



**HAL**  
open science

# A Non-asymptotic Approach to Best-Arm Identification for Gaussian Bandits

Antoine Barrier, Aurélien Garivier, Tomáš Kocák

► **To cite this version:**

Antoine Barrier, Aurélien Garivier, Tomáš Kocák. A Non-asymptotic Approach to Best-Arm Identification for Gaussian Bandits. 2021. hal-03236583v1

**HAL Id: hal-03236583**

**<https://hal.science/hal-03236583v1>**

Preprint submitted on 26 May 2021 (v1), last revised 4 Mar 2022 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A Non-asymptotic Approach to Best-Arm Identification for Gaussian Bandits

Antoine Barrier<sup>1, 2</sup>, Aurélien Garivier<sup>1</sup>, and Tomáš Kocák<sup>1</sup>

<sup>1</sup>UMPA, CNRS, Inria, ÉNS Lyon, Lyon (France)  
<sup>2</sup>LMO, CNRS, Université Paris-Saclay, Orsay (France)

May 26, 2021

## Abstract

We propose a new strategy for best-arm identification with fixed confidence of Gaussian variables with bounded means and unit variance. This strategy called EXPLORATION-BIASED SAMPLING is not only asymptotically optimal: we also prove non-asymptotic bounds occurring with high probability. To the best of our knowledge, this is the first strategy with such guarantees. But the main advantage over other algorithms like TRACK-AND-STOP is an improved behavior regarding exploration: EXPLORATION-BIASED SAMPLING is slightly biased in favor of exploration in a subtle but natural way that makes it more stable and interpretable. These improvements are allowed by a new analysis of the sample complexity optimization problem, which yields a faster numerical resolution scheme and several quantitative regularity results that we believe of high independent interest.

**Keywords:** Best arm identification · Fixed confidence · Multi-armed bandits · Sequential Learning

## 1 Introduction

As detailed for example in [Lattimore and Szepesvári, 2020], many modern systems of automatic decisions (from recommender systems to clinical trials, through auto-ML and parameter tuning) require to find the best among a set of options, using noisy observations obtained by successive calls to a random mechanism. The simplest formal model for such situations is the *standard Gaussian multi-armed bandit*, a collection of  $K \geq 2$  independent Gaussian distributions called *arms* of unknown means  $\boldsymbol{\mu} = (\mu_a)_{a \in [K]} \in \mathbb{R}^K$  and variances all equal to 1. They are sampled sequentially and independently: at every discrete time step  $t \in \mathbb{N}^+$ , an agent chooses an arm  $A_t \in [K]$  based on past information, and observes an independent draw  $Y_t$  from distribution  $\mathcal{N}(\mu_{A_t}, 1)$ .

Among the set  $\mathcal{G}$  of all standard Gaussian multi-armed bandits with means in the interval  $[0, 1]$ , we focus in this work on the subset  $\mathcal{G}^*$  that have exactly one arm  $a^*(\boldsymbol{\mu}) \in [K]$  with the highest mean  $\mu_{a^*(\boldsymbol{\mu})} = \mu_* = \max_{a \in [K]} \mu_a$ : we address the problem of optimally sampling the arms in order to identify  $a^*(\boldsymbol{\mu})$  as quickly as possible. We consider the sequential statistics framework often called *fixed confidence setting* (see [Even-Dar et al., 2006, Kalyanakrishnan et al., 2012]): in this framework,  $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$  denotes the sigma-field generated by the observations up to time  $t$  and a strategy consists of a sampling rule  $(A_t)_{t \geq 1}$  where each  $A_t$  is  $\mathcal{F}_{t-1}$ -measurable, a stopping rule  $\tau$  with respect to  $\mathcal{F}_t$ , and an  $\mathcal{F}_\tau$ -measurable decision rule  $\hat{a}_\tau$ . Given a risk parameter  $\delta \in (0, 1)$ , a strategy is called  $\delta$ -correct if, whatever the parameter  $\boldsymbol{\mu} = (\mu_a)_{a \in [K]} \in \mathcal{G}^*$ , it holds that  $\mathbb{P}(\hat{a}_\tau \notin \operatorname{argmax}_{a \in [K]} \mu_a) \leq \delta$ . The goal is to find a  $\delta$ -correct strategy that minimizes the expected number of observations  $\mathbb{E}[\tau_\delta]$  needed to identify the best arm.

The sample complexity of  $\delta$ -correct strategies cannot be arbitrarily good: it has been proved in [Garivier and Kaufmann, 2016] that they essentially obey the lower bound  $\mathbb{E}[\tau_\delta] \geq T(\boldsymbol{\mu}) \log 1/\delta$  for any

$\boldsymbol{\mu} \in \mathcal{G}^*$ , where the *characteristic time*  $T(\boldsymbol{\mu})$  is the solution of the following optimization problem

$$T(\boldsymbol{\mu})^{-1} = \sup_{\mathbf{v} \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \sum_{a \in [K]} v_a \frac{(\mu_a - \lambda_a)^2}{2}, \quad (1)$$

where  $\Sigma_K = \{\mathbf{v} \in [0, 1]^K : v_1 + \dots + v_K = 1\}$  is the simplex of dimension  $K - 1$  and  $\text{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \mathcal{G}^* : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$  the set of Gaussian bandit models with an optimal arm different from  $a^*(\boldsymbol{\mu})$ . Moreover, this bound is tight: the authors introduced TRACK-AND-STOP, a strategy for which they proved that  $\limsup_{\delta \rightarrow 0} \mathbb{E}[\tau_\delta] / \log(1/\delta) = T(\boldsymbol{\mu})$  (see also [Russo, 2016]). Later, [Ménard, 2019] proved that it is not necessary to solve the optimization problem in every time step. Instead, they perform a single gradient step in every round which enables them to prove a similar result while reducing the computational complexity of their algorithm.

The information-theoretic analysis of [Garivier and Kaufmann, 2016] also highlights the nature of the optimal sampling strategy: whatever the value of the risk  $\delta$ , one should sample the arms with frequencies proportional to  $\mathbf{v} = \mathbf{w}(\boldsymbol{\mu})$ , the (unique and well-defined) maximizer in the right-hand side of Equation (1). In fact, TRACK-AND-STOP merely boils down to estimating these optimal frequencies of draws, from past observations, and following them. The only exception is that some exploration is forced in order to avoid vicious circles where an arm might be first under-estimated and then consequently under-sampled.

The TRACK-AND-STOP algorithm is not only a theoretical contribution, it also proved to be numerically efficient, far exceeding its competitors in a wide variety of settings. It was improved in different directions [Degenne and Koolen, 2019, Degenne et al., 2019, Shang et al., 2019], and also provides a simple template for extensions, for bandit problems with structure [Kocák and Garivier, 2020], as long as the optimization problem (1) can be solved. Yet, TRACK-AND-STOP suffers from certain shortcomings. First, a close look into the proofs shows that the theoretical guarantees proved so far are really asymptotic in nature. Second, the forced exploration in TRACK-AND-STOP appears very arbitrary, with a rate of  $\sqrt{t}$  that has no other justification than lying somewhere between constant and linear functions. Third, the sampling strategy of TRACK-AND-STOP appears to be pretty unstable, especially at the beginning: the target frequencies can vary significantly as the estimated means fluctuate before stabilizing around their expectations. Fourth, TRACK-AND-STOP does not present the intuitively desirable behavior to sample uniformly in the beginning, until sufficient information has been gathered for significant differences between the arms to emerge. This is in contrast with strategies like Racing [Kaufmann and Kalyanakrishnan, 2013], which are sub-optimal but intuitively appealing. Altogether, these issues lead for example to unpredictable and irregular conduct at the beginning of multiple A/B testing cases with many arms very close to optimal.

The present paper addresses all these issues and proposes a new algorithm that permits solving all of them together. The exploration is conducted very differently, in a statistically natural way that softens the fluctuations of empirical means and avoids arbitrary parameters. It results in a stabilized sampling strategy, that is much easier to follow and understand. We propose for this strategy a non-asymptotic analysis with finite risk bounds. These results have required developing a careful analysis of the quantitative regularity of the solution to the optimization problem (1). As a by-product, we obtain an accelerated algorithm for its numerical resolution, which permits a significant speed-up for the TRACK-AND-STOP or the Gradient Ascent algorithms in the Gaussian case. Actually, the algorithms discussed here are not restricted to Gaussian arms but apply equally to sub-Gaussian arms with a known upper bound on the variances (in these settings, the sample complexity bounds proved in this paper apply but are not necessarily optimal).

The paper is organized as follows. We present in Section 2 our new EXPLORATION-BIASED SAMPLING strategy and present its main properties and guarantees. We then turn in Section 3 to the analysis of the optimization problem (1) and to the resulting new algorithm for its numerical resolution. Section 4 is dedicated to the proof of the non-asymptotic sample complexity bound of the EXPLORATION-BIASED SAMPLING strategy (other proofs are gathered in the appendix). Lastly, we illustrate its performance and behavior by numerical experiments in Section 5, and propose concluding remarks in Section 6.

## 2 The Exploration-Biased Sampling strategy

In this section, we introduce our new strategy called EXPLORATION-BIASED SAMPLING. We describe the algorithm and state its guarantees.

For  $\boldsymbol{\mu} \in \mathcal{G}$ , we denote by  $\boldsymbol{\Delta}(\boldsymbol{\mu}) \in [0, 1]^K$  its *gap vector* defined by  $\Delta_a(\boldsymbol{\mu}) = \mu_* - \mu_a$  for  $a \in [K]$  and by  $a^*(\boldsymbol{\mu}) = \{a \in [K] : \mu_a = \mu_*\} = \{a \in [K] : \Delta_a(\boldsymbol{\mu}) = 0\}$  its set of optimal arms. When  $\boldsymbol{\mu} \in \mathcal{G}^*$ ,  $a^*(\boldsymbol{\mu})$  has one element that we also denote by  $a^*(\boldsymbol{\mu})$  and recall that the *optimal weight vector*  $\boldsymbol{w}(\boldsymbol{\mu})$  is the unique maximizer of optimization problem (1). Otherwise, when  $\boldsymbol{\mu} \in \mathcal{G} \setminus \mathcal{G}^*$  has at least two optimal arms, we define

$$\boldsymbol{w}(\boldsymbol{\mu}) = \frac{1}{\text{card}(a^*(\boldsymbol{\mu}))} (\mathbb{1}_{1 \in a^*(\boldsymbol{\mu})}, \dots, \mathbb{1}_{K \in a^*(\boldsymbol{\mu})})^T.$$

Since these quantities play a special role in the sequel, we set  $w_{\min}(\boldsymbol{\mu}) = \min_{a \in [K]} w_a(\boldsymbol{\mu})$ ,  $\Delta_{\min}(\boldsymbol{\mu}) = \min_{a \in [K]: \Delta_a(\boldsymbol{\mu}) > 0} \Delta_a(\boldsymbol{\mu})$  (which is not defined when  $a^*(\boldsymbol{\mu}) = [K]$ ) and  $\Delta_{\max}(\boldsymbol{\mu}) = \max_{a \in [K]} \Delta_a(\boldsymbol{\mu})$ .

Given a sampling strategy, let  $N_a(t) = \sum_{1 \leq s \leq t} \mathbb{1}\{A_s = a\}$  be the (random) number of draws of arm  $a \in [K]$  up to time  $t \in \mathbb{N}$ , and  $\boldsymbol{N}(t) = (N_a(t))_{a \in [K]}$ . For every arm  $a \in [K]$  and time  $t \geq 1$  such that  $N_a(t) \geq 1$ , the maximum likelihood estimate of  $\mu_a$  at time  $t$  is  $\hat{\mu}_a(t) = N_a(t)^{-1} \sum_{s=1}^t Y_s \mathbb{1}_{A_s=a}$ .

In the rest of this section, we fix  $\boldsymbol{\mu} \in \mathcal{G}$ .

### 2.1 Conservative tracking

The main idea of the algorithm is to design a sampling policy of arms that naturally encourages exploration without forcing it like TRACK-AND-STOP does. To do so, the objective is to “wrap” the optimal weight vector  $\boldsymbol{w}(\boldsymbol{\mu})$  “from above”, by ensuring that we never under-estimate its minimal value. Indeed, even an arm with low mean needs to be sampled sufficiently often until one is very confident that it is suboptimal. The idea is to construct a confidence region  $\mathcal{CR}_{\boldsymbol{\mu}} \subset [0, 1]^d$  for  $\boldsymbol{\mu}$  on which one can efficiently find a bandit  $\tilde{\boldsymbol{\mu}} \in \mathcal{CR}_{\boldsymbol{\mu}}$  maximizing the minimal weight  $w_{\min}$ :

$$\tilde{\boldsymbol{\mu}} \in \operatorname{argmax}_{\boldsymbol{\nu} \in \mathcal{CR}_{\boldsymbol{\mu}}} w_{\min}(\boldsymbol{\nu}). \quad (2)$$

As long as  $\boldsymbol{\mu}$  belongs to the confidence region  $\mathcal{CR}_{\boldsymbol{\mu}}$ , choosing the target weights  $\boldsymbol{w}(\tilde{\boldsymbol{\mu}})$  guarantees that every arm is explored sufficiently. The exploration bias decreases with the number of observations, as the confidence region  $\mathcal{CR}_{\boldsymbol{\mu}}$  shrinks to  $\{\boldsymbol{\mu}\}$ , and in the end every arm is sampled with a frequency proportional to  $w_a(\boldsymbol{\mu})$ .

This approach to exploration requires two ingredients:

- the exploration-biased bandit  $\tilde{\boldsymbol{\mu}}$  needs to be efficiently computable. It turns out to be the case if the confidence region is a product of confidence intervals on each arm (a mild requirement since the arms are independent). We propose Algorithm 1, an efficient procedure for computing  $\tilde{\boldsymbol{\mu}}$ . Intuitively, maximizing  $w_{\min}$  over  $\mathcal{CR}(\boldsymbol{\mu})$  requires to increase and equalize all the positive gaps as much as possible. The associated bandit will indeed be the one for which it is harder to identify the second best arm and thus it will require to sample the worst arms more frequently. This gives a candidate bandit for each potential best arm, and our Algorithm compares those candidates. Figure 1 illustrates on an example the principle of Algorithm 1, whose correctness is proved in Proposition 1. Note that it requires an efficient procedure for computing the optimal weight vector  $\boldsymbol{w}(\boldsymbol{\nu})$  of any bandit  $\boldsymbol{\nu} \in \mathcal{G}^*$ : such a procedure (see Algorithm 3) is presented in Section 3.
- the regularity of the mapping  $\boldsymbol{\nu} \mapsto \boldsymbol{w}(\boldsymbol{\nu})$  needs to be explicitly known. Indeed, the confidence region will decrease with the number of observations, and  $\tilde{\boldsymbol{\mu}}$  will come close to  $\boldsymbol{\mu}$ . The continuity proved in [Garivier and Kaufmann, 2016] for the asymptotic optimality of TRACK-AND-STOP is not sufficient: the first quantitative bounds are given below in Section 3.4.

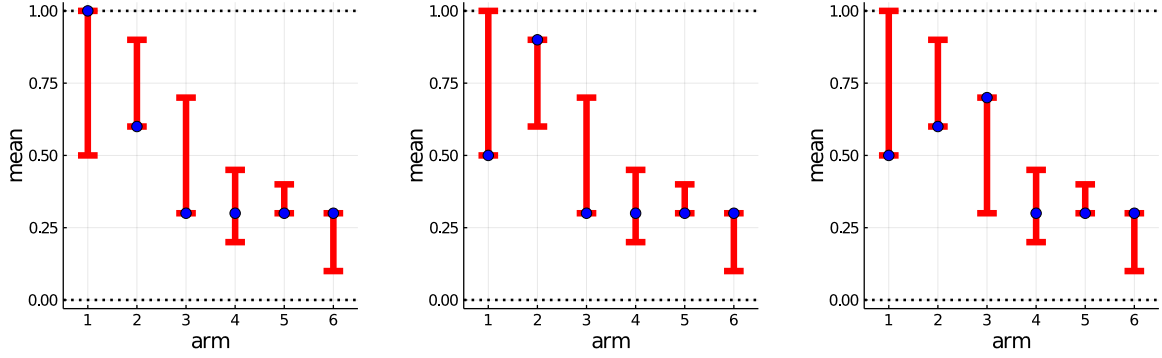


Figure 1: List of the bandits  $(\tilde{\mu}^{\text{test}(a)})_{a \in \text{PotentialBestArms}}$  tried by Algorithm 1 for the example confidence region in red with  $\text{PotentialBestArms} = \{1, 2, 3\}$ . From left to right:  $\tilde{\mu}^{\text{test}(1)}$ ,  $\tilde{\mu}^{\text{test}(2)}$  and  $\tilde{\mu}^{\text{test}(3)}$

One can remark that as long as the confidence intervals have a non-empty intersection, which means the observations do not permit to exclude that any of them is optimal, the exploration-biased weights returned by Algorithm 1 are uniform and the arms are sampled in a round-robin way (as in a Racing or Successive Elimination algorithm like [Even-Dar et al., 2006]).

---

**Algorithm 1: EXPLORATION-BIASED WEIGHTS**


---

**Input:** confidence region  $\mathcal{CR} = \prod_{a \in [K]} [\underline{\mu}_a, \bar{\mu}_a]$

**Output:** exploration-biased bandit  $\tilde{\mu} \in \mathcal{CR}$   
exploration-biased optimal weight vector  $\mathbf{w} = \mathbf{w}(\tilde{\mu})$

```

maxLB  $\leftarrow$   $\max_{a \in [K]} \underline{\mu}_a$ 
minUB  $\leftarrow$   $\min_{a \in [K]} \bar{\mu}_a$ 
if  $\text{minUB} \geq \text{maxLB}$  then
     $\tilde{\mu} \leftarrow (\text{minUB}, \dots, \text{minUB})$ 
     $\mathbf{w} \leftarrow (1/K, \dots, 1/K)$ 
else
    PotentialBestArms  $\leftarrow$   $\{a \in [K] : \bar{\mu}_a > \text{maxLB}\}$ 
     $\mathbf{w} \leftarrow (0, \dots, 0)$ 
    for  $a \in \text{PotentialBestArms}$  do
         $\tilde{\mu}_a^{\text{test}(a)} \leftarrow \bar{\mu}_a$ 
        for  $b \neq a$  do
             $\tilde{\mu}_b^{\text{test}(a)} \leftarrow \max(\underline{\mu}_b, \text{minUB})$ 
         $\mathbf{w}^{\text{test}(a)} \leftarrow \text{OPTIMAL WEIGHTS}(\tilde{\mu}^{\text{test}(a)})$ 
        if  $w_{\text{min}}^{\text{test}(a)} > w_{\text{min}}$  then
             $\mathbf{w} \leftarrow \mathbf{w}^{\text{test}(a)}$ 
             $\tilde{\mu} \leftarrow \tilde{\mu}^{\text{test}(a)}$ 

```

---

**Proposition 1.** Let  $\mathcal{CR} \subset [0, 1]^K$  be some region that can be decomposed as a product of intervals and  $(\tilde{\mu}, \mathbf{w}) \leftarrow \text{EXPLORATION-BIASED WEIGHTS}(\mathcal{CR})$ . Then  $\mathbf{w} = \mathbf{w}(\tilde{\mu})$  and:

$$w_{\text{min}} = \max_{\nu \in \mathcal{CR}} w_{\text{min}}(\nu).$$

The proof of Proposition 1 is given in Appendix B.4 and relies on the results of Section 3.3.

## 2.2 The strategy

We are now able to introduce our strategy called EXPLORATION-BIASED SAMPLING. Given a risk  $\delta \in (0, 1)$  and a threshold function  $\beta(t, \delta)$ , we compute at each time confidence intervals for each  $\mu_a$  that will ensure  $\boldsymbol{\mu}$  to belong to each associated confidence region with probability at least  $1 - \gamma$ , where  $\gamma \in (0, 1)$  is a fixed parameter. We can then ensure enough exploration by biasing the optimal weights  $\boldsymbol{w}(\boldsymbol{\mu})$  using Algorithm 1. As explained in [Garivier and Kaufmann, 2016], one can either follow the exploration-biased weights directly (D-tracking) or their cumulative sums (C-tracking). For the simplicity of the proofs, we use C-tracking in the analysis, but we ran the experiments with both options (D-tracking appears to perform slightly better).

---

**Algorithm 2:** EXPLORATION-BIASED SAMPLING

---

**Input:** confidence level  $\delta$   
threshold function  $\beta(t, \delta)$   
confidence parameter  $\gamma$

**Output:** stopping time  $\tau_\delta$   
estimated best arm  $\hat{a}_{\tau_\delta}$

Observe each arm once  
**for**  $s = 0$  **to**  $K - 1$  **do**  
|  $\tilde{\boldsymbol{w}}(s) \leftarrow (1/K, \dots, 1/K)$   
 $t \leftarrow K$   
**while**  $Z(t) \leq \beta(t, \delta)$  **do**  
| **for**  $a \in [K]$  **do**  
| |  $\underline{\mu}_a(t), \bar{\mu}_a(t) \leftarrow \hat{\mu}_a(t) - C_{\gamma/K}(N_a(t)), \hat{\mu}_a(t) + C_{\gamma/K}(N_a(t))$   
|  $\mathcal{CR}_\mu(t) \leftarrow \prod_{a \in [K]} [\underline{\mu}_a(t), \bar{\mu}_a(t)]$   
|  $(\tilde{\boldsymbol{\mu}}(t), \tilde{\boldsymbol{w}}(t)) \leftarrow \text{EXPLORATION-BIASED WEIGHTS}(\mathcal{CR}_\mu(t))$   
| Choose  $A_{t+1} \in \operatorname{argmin}_{a \in [K]} N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s)$   
| Observe  $Y_{A_{t+1}}$   
| Increase  $t$  by 1  
 $\tau_\delta \leftarrow t$   
 $\hat{a}_{\tau_\delta} \leftarrow \operatorname{argmax}_{a \in [K]} \hat{\mu}_a(t)$

---

**Confidence regions** Confidence regions are designed to satisfy two requirements: Algorithm 1 requires a product of confidence intervals, and we need that  $\boldsymbol{\mu}$  belongs to them at all times. For  $\gamma \in (0, 1)$ , we define for  $t \in \llbracket K, \tau_\delta \rrbracket$

$$\mathcal{CR}_\mu(t) = \prod_{a \in [K]} [\hat{\mu}_a(t) \pm \ell_a(t)], \quad (3)$$

where  $\ell_a(t) = C_{\gamma/K}(N_a(t))$  is the half-width of the confidence interval for  $\mu_a$  and  $C_\gamma(s) = 2\sqrt{\frac{\log(4s/\gamma)}{s}}$ . The following Lemma, proved in Appendix A, states a time-uniform confidence guarantee for  $\boldsymbol{\mu}$ .

**Lemma 2.** *For any  $\boldsymbol{\mu} \in \mathcal{G}$  and  $\gamma \in ]0, 1[$ , we have*

$$\mathbb{P}_\mu \left( \exists t \in \llbracket K, \tau_\delta \rrbracket : \boldsymbol{\mu} \notin \mathcal{CR}_\mu(t) \right) \leq \gamma.$$

It happens that the choice of confidence regions given by Equation (3) leads to a minimal exploration rate for each arm of order  $\sqrt{t}$ . What is surprising is that this is exactly the arbitrary rate that TRACK-AND-STOP is using for forced exploration, which appears here naturally.

**Lemma 3.** *For any choice of parameters and  $\boldsymbol{\mu} \in \mathcal{G}$ , EXPLORATION-BIASED SAMPLING satisfies*

$$\forall t \in [0, \tau_\delta], \forall a \in [K], \quad N_a(t) \geq \frac{2}{K} \sqrt{t} - K.$$

The proof of this Lemma can be found in Appendix D.1.

**Stopping rule** Following [Garivier and Kaufmann, 2016], our stopping rule relies on a statistic  $Z(t)$  defined as

$$Z(t) = \max_{a \in [K]} \min_{b \neq a} Z_{a,b}(t),$$

where  $Z_{a,b}(t)$  is the Generalized Likelihood Ratio statistic (see [Chernoff, 1959]), equal in the Gaussian case to

$$Z_{a,b}(t) = \frac{1}{2} \frac{N_a(t)N_b(t)}{N_a(t) + N_b(t)} (\hat{\mu}_a(t) - \hat{\mu}_b(t))^2 \operatorname{sgn}(\hat{\mu}_a(t) - \hat{\mu}_b(t)),$$

with  $\operatorname{sgn}(u) = \mathbf{1}\{u > 0\} - \mathbf{1}\{u < 0\}$  the sign of the real number  $u$ .

### 2.3 Theoretical results

**A  $\delta$ -correct strategy** The  $\delta$ -correctness of EXPLORATION-BIASED SAMPLING, which relies on the same stopping rule as TRACK-AND-STOP, is a simple consequence of [Garivier and Kaufmann, 2016, Proposition 12].

**Proposition 4.** *For any  $\delta, \gamma \in (0, 1)$  and  $\alpha > 1$ , there exists a constant  $R = R(K, \alpha)$  such that EXPLORATION-BIASED SAMPLING with parameters  $\delta, \gamma$  and threshold*

$$\beta(t, \delta) = \log \left( \frac{Rt^\alpha}{\delta} \right) \tag{4}$$

is  $\delta$ -correct.

Our main result is to obtain high probability bounds for  $\tau_\delta$  in finite horizon for the EXPLORATION-BIASED SAMPLING algorithm, which is summarized in the following theorem. Note that such a result cannot be proved for the TRACK-AND-STOP algorithm.

**Theorem 5** (Non-asymptotic bound). *Fix  $\gamma \in (0, 1)$ ,  $\alpha \in [1, 2]$ ,  $\eta \in (0, 1]$  and let  $\boldsymbol{\mu} \in \mathcal{G}^*$ . There exists an event  $\mathcal{E}$  of probability at least  $1 - \gamma$  and  $\delta_0 = \delta_0(\boldsymbol{\mu}, K, \gamma, \eta, \alpha) > 0$  such that for any  $0 < \delta \leq \delta_0$ , algorithm EXPLORATION-BIASED SAMPLING with the threshold of Equation (4) satisfies*

$$\forall t > (1 + \eta)T \log(1/\delta), \quad \mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta > t \cap \mathcal{E}) \leq 2Kt \exp \left( - \frac{tw_{\min}}{4T^2} \frac{1}{\log^{\frac{2}{3}}(1/\delta)} \right) \tag{5}$$

and

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta \mathbf{1}_{\mathcal{E}}] \leq (1 + \eta)T \log(1/\delta) + \frac{2^7 KT^4}{w_{\min}^2} \exp \left( - \frac{w_{\min}}{4T} \log^{\frac{1}{3}}(1/\delta) \right) \log^2(1/\delta) \tag{6}$$

where  $T = T(\boldsymbol{\mu})$  and  $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{\mu})$ .

Note that the proof of Theorem 5 provides an explicit expression for  $\delta_0$ . Also note that the second term tends to 0, and is hence neglectible with respect to the first term, when  $\delta$  decreases to 0. We additionally prove that, from an asymptotic point of view, the EXPLORATION-BIASED SAMPLING algorithm presents the same guarantees as TRACK-AND-STOP:

**Theorem 6** (Almost sure asymptotic bound). *Fix  $\gamma \in (0, 1)$ ,  $\alpha \in [1, e/2]$ . For any  $\boldsymbol{\mu} \in \mathcal{G}^*$ , Algorithm EXPLORATION-BIASED SAMPLING with the threshold of Equation (4) satisfies*

$$\limsup_{\delta \rightarrow 0} \frac{\tau_\delta}{\log(1/\delta)} \leq \alpha T(\boldsymbol{\mu}) \quad \mathbb{P}_{\boldsymbol{\mu}\text{-a.s.}}$$

**Theorem 7** (Asymptotic optimality in expectation). *Fix  $\gamma \in (0, 1)$ ,  $\alpha \in (1, e/2]$  and let  $\boldsymbol{\mu} \in \mathcal{G}^*$ . Algorithm EXPLORATION-BIASED SAMPLING with the threshold of Equation (4) satisfies*

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T(\boldsymbol{\mu}).$$

Section 4 will be devoted to the proof of Theorem 5 while the proof of Theorems 6 and 7 can be found in Appendix D. It is worth mentioning that the guarantees of EXPLORATION-BIASED SAMPLING presented in this Section hold true not only for Gaussian arms, but more generally for 1-subGaussian arms with means in  $[0, 1]$ , as in those proofs the Gaussian assumption is only be used in exponential deviation bounds.

### 3 About the sample complexity optimization problem

We now introduce a new method for solving the sample complexity optimization problem (1). It comes with a new analysis that yields various bounds for the bandits characteristic constants together with monotonicity and regularity results. Detailed discussions and proofs are deferred to Appendix B.

In this section and its associated Appendix B, letters  $a, b, c$  always refer to arm indices, that is elements of  $[K]$ . In subindices for sums and infima, we sometimes omit to explicitly mention  $[K]$  for simplicity: for example, given a fixed arm  $b$ ,  $\sum_{a \neq b}$  denotes the sum over arms  $a \in [K] \setminus \{b\}$ .

For any bandit  $\boldsymbol{\mu} \in \mathcal{G}$  and  $\mathbf{v} \in \Sigma_K$ , we define:

$$g(\boldsymbol{\mu}, \mathbf{v}) = \inf_{\lambda \in \text{Alt}(\boldsymbol{\mu})} \sum_{a \in [K]} v_a \frac{(\mu_a - \lambda_a)^2}{2} = \frac{1}{2} \min_{a \neq a^*} \frac{v_{a^*} v_a}{v_{a^*} + v_a} \Delta_a(\boldsymbol{\mu})^2. \quad (7)$$

The easy proof of the second equality can be found in Appendix B.1. Function  $g$  is twice useful, as the solution to the inner optimization problem (1), and for the expression of the statistic  $Z(t)$ :

$$T(\boldsymbol{\mu})^{-1} = g(\boldsymbol{\mu}, \mathbf{w}(\boldsymbol{\mu})), \quad (8)$$

$$\text{and} \quad Z(t) = t g\left(\hat{\boldsymbol{\mu}}(t), \frac{\mathbf{N}(t)}{t}\right) \quad (9)$$

with the convention  $T(\boldsymbol{\mu}) = +\infty$  when  $\boldsymbol{\mu} \in \mathcal{G} \setminus \mathcal{G}^*$ .

Let in this section  $\boldsymbol{\mu} \in \mathcal{G}^*$  be a fixed bandit parameter. For the simplicity of the presentation, we denote  $a^* = a^*(\boldsymbol{\mu})$ ,  $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\mu})$ ,  $\mathbf{w} = \mathbf{w}(\boldsymbol{\mu})$  and  $T = T(\boldsymbol{\mu})$ .

#### 3.1 Solving the optimization problem

We define

$$\phi_{\boldsymbol{\mu}} : r \in \left(\frac{1}{\Delta_{\min}^2}, +\infty\right) \mapsto \sum_{a \neq a^*} \frac{1}{(r \Delta_a^2 - 1)^2} - 1. \quad (10)$$

**Lemma 8.**  *$\phi_{\boldsymbol{\mu}}$  is convex and strictly decreasing on  $(1/\Delta_{\min}^2, +\infty)$ , and thus has a unique root.*

The following proposition shows that solving  $\phi_{\boldsymbol{\mu}}(r) = 0$  directly gives a solution to Problem (1).



**Proposition 9.** *Let  $r = r(\boldsymbol{\mu})$  be the solution of  $\phi_{\boldsymbol{\mu}}(r) = 0$ . Then*

$$w_{a^*} = \frac{1}{1 + \sum_{a \neq a^*} \frac{1}{r\Delta_a^2 - 1}}, \quad (11)$$

$$\forall a \neq a^*, \quad w_a = \frac{w_{a^*}}{r\Delta_a^2 - 1}, \quad (12)$$

$$\text{and} \quad T = 2 \frac{r}{w_{a^*}}. \quad (13)$$

Besides,

$$w_{a^*} = \sqrt{\sum_{a \neq a^*} w_a^2}. \quad (14)$$

Recall that in the case of  $K = 2$  arms,  $\mathbf{w}(\boldsymbol{\mu}) = (0.5, 0.5)$  for any  $\boldsymbol{\mu} \in \mathcal{G}^*$ . Besides, the monotonicity of the optimal weights with respect to the gaps follows from Equation (12).

**Corollary 10.** *Assume that  $K \geq 3$ . Then*

$$\forall a, b \in [K], \quad \mu_a > \mu_b \implies w_a > w_b.$$

Equation (12) also implies that

$$\forall a, b \neq a^*, \quad \frac{w_a}{w_b} = \frac{\Delta_b^2 - 1/r}{\Delta_a^2 - 1/r}$$

Intuitively, it requires about  $\Delta_a^2$  samplings of arms  $a^*$  and  $a$  before being able to distinguish them, so that one could expect  $\frac{w_a}{w_b}$  to be  $\frac{\Delta_b^2}{\Delta_a^2}$ . This would be the case if the comparisons between arms were independent. In our problem, sampling the best arm benefits the comparison with all arms, so that it is worth sampling the optimal arm a little more than any single comparison would require, and hence each sub-optimal arm a little less. As a result, the ratio  $\frac{w_a}{w_b}$  is closer to 1, and the factor can be seen as a ‘‘discount’’ on each squared gap for sharing the comparisons. We now derive other important consequences of Proposition 9.

### 3.2 Bounds and computation of the problem characteristics

By Proposition 9, it suffices to compute  $r$  to obtain the values of both  $T$  and  $\mathbf{w}$ . As  $\phi_{\boldsymbol{\mu}}$  is a strictly convex and strictly decreasing function, Newton’s iterates initialized with a value  $r_0 < r$  converge to  $r$  from below at quadratic speed. The procedure is summarized in Algorithm 3 of Appendix B.3. The number of correct digits roughly doubles at every step, which implies that a few iterations are sufficient to guarantee machine precision. The cost of the algorithm can hence be considered proportional to that of evaluating  $\phi_{\boldsymbol{\mu}}(r)$ , which is linear in the number of arms.

It remains to show that it is possible to find  $r_0 < r$ , and possibly close to  $r$ . The next proposition offers such a lower bound as simple functions of the gaps. This also yields tight bounds on the optimal weight vector  $\mathbf{w}$  and the characteristic time  $T$ .

**Proposition 11.** *Denoting by  $\overline{\Delta^2} = \frac{1}{K-1} \sum_{a \neq a^*} \Delta_a^2$  the average squared gap,*

$$\max \left( \frac{2}{\Delta_{\min}^2}, \frac{1 + \sqrt{K-1}}{\overline{\Delta^2}} \right) \leq r \leq \frac{1 + \sqrt{K-1}}{\Delta_{\min}^2}, \quad (15)$$

$$\frac{1}{1 + \sqrt{K-1}} \leq w_{\max} \leq \frac{1}{2}, \quad (16)$$

$$\text{and} \quad \max \left( \frac{8}{\Delta_{\min}^2}, 4 \frac{1 + \sqrt{K-1}}{\overline{\Delta^2}} \right) \leq T \leq 2 \frac{(1 + \sqrt{K-1})^2}{\Delta_{\min}^2}. \quad (17)$$

All of these inequalities can be reached for certain parameters  $\boldsymbol{\mu}$ , as discussed in Appendix B.2 after the proof of Proposition 11.

### 3.3 About the monotonicity of the min-max problem

We now show the monotonicity results of the mappings  $\boldsymbol{\nu} \mapsto T(\boldsymbol{\nu})$  and  $\boldsymbol{\nu} \mapsto \mathbf{w}(\boldsymbol{\nu})$  when moving one or several arms. When  $K = 2$ , the optimization problem is simple and leads to  $\mathbf{w}(\boldsymbol{\mu}) = (0.5, 0.5)$  and  $T(\boldsymbol{\mu}) = 8\Delta_2^2$ , so that we consider in the remaining of this section that  $K \geq 3$ .

Let  $\boldsymbol{\mu}' \in \mathcal{G}^*$  be another bandit problem sharing the same unique optimal arm  $a^*$  as  $\boldsymbol{\mu}$  and define  $T'$ ,  $\mathbf{w}'$ ,  $\Delta'$ ,  $r'$  the characteristic time, optimal weight vector, gaps and root of  $\phi_{\boldsymbol{\mu}'}$  relative to the parameter  $\boldsymbol{\mu}'$ . The three following lemmas are proved in Appendix B.4.

**Lemma 12.** *Assume that  $\Delta'_b > \Delta_b$  for a fixed  $b \neq a^*$  while  $\Delta'_a = \Delta_a$  for all  $a \neq b$ . Then*

1.  $w'_b < w_b$ ,
2.  $w'_a > w_a$  for any  $a \notin \{a^*, b\}$ ,
3.  $T' < T$ .

**Lemma 13.** *Assume that  $\Delta'_a = \Delta_a + d$  for every  $a \neq a^*$  and some positive constant  $d$ . Then  $w'_{\min} \geq w_{\min}$ . The inequality is strict whenever  $\Delta_a \neq \Delta_b$  for some  $a, b \neq a^*$ .*

**Lemma 14.** *Let  $B = \operatorname{argmin}_{a \in [K]} \mu_a$  (resp.  $B' = \operatorname{argmin}_{a \in [K]} \mu'_a$ ) be the set of the worst arms of  $\boldsymbol{\mu}$  (resp.  $\boldsymbol{\mu}'$ ) and assume that  $B \subset B'$  and  $\Delta'_{\max} < \Delta_{\max}$ , while  $\Delta'_a = \Delta_a$  for all  $a \notin B'$ . Then  $w'_{\min} \geq w_{\min}$ .*

Note that those Lemmas are the key ingredients of the proof of Proposition 1 when  $K \geq 3$  (when  $K = 2$  the proposition is trivial).

### 3.4 Regularity of $\mathbf{w}$ , $T$ and $g$

Lastly, we show explicit bounds on the regularity of  $\boldsymbol{\nu} \mapsto \mathbf{w}(\boldsymbol{\nu})$  and  $\boldsymbol{\nu} \mapsto T(\boldsymbol{\nu})$ . We keep the notations of the last section.

**Theorem 15.** *Assume that  $(1 - \varepsilon)\Delta_a^2 \leq \Delta_a'^2 \leq (1 + \varepsilon)\Delta_a^2$  for all  $a \neq a^*$  and some  $\varepsilon \in [0, 1/7]$ . Then*

$$(1 - 3\varepsilon)T \leq T' \leq (1 + 6\varepsilon)T ,$$

$$\forall a \in [K], \quad (1 - 10\varepsilon)w_a \leq w'_a \leq (1 + 10\varepsilon)w_a .$$

Independently, we show the following property of function  $g$ .

**Proposition 16.** *Let  $\mathbf{v} \in \Sigma_K$ . Then:*

$$g(\boldsymbol{\mu}', \mathbf{v}) \geq \frac{(1 - \eta)^2}{1 + \eta} (g(\boldsymbol{\mu}, \mathbf{w}(\boldsymbol{\mu})) - \varepsilon/2)$$

where  $\varepsilon = \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_{\infty}$  and  $\eta = \max_{a \in [K]} \frac{|w_a(\boldsymbol{\mu}) - v_a|}{w_a(\boldsymbol{\mu})}$ .

These results will prove to be essential to prove Theorem 5.

## 4 Proof of the main result

The aim of this section is to prove Theorem 5. Let  $\gamma \in (0, 1)$  and  $\boldsymbol{\mu} \in \mathcal{G}^*$ . We assume, without loss of generality, that  $a^*(\boldsymbol{\mu}) = 1$ . We also write for simplicity  $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\mu})$ ,  $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{\mu})$  and  $T = T(\boldsymbol{\mu})$ .

Let  $\mathcal{E}$  denotes an event such that  $\boldsymbol{\mu}$  belongs to all confidence regions:

$$\mathcal{E} = \bigcap_{t=K}^{\tau_\delta} (\boldsymbol{\mu} \in \mathcal{CR}_\boldsymbol{\mu}(t))$$

and recall that the confidence regions defined by Equation (3) are chosen so as to ensure that  $\mathbb{P}_\boldsymbol{\mu}(\mathcal{E}) \geq 1 - \gamma$  (see Lemma 2). Furthermore, when  $\mathcal{E}$  occurs, EXPLORATION-BIASED SAMPLING has been designed so that arms are observed with some minimal linear rate, specified by Lemma 17 and proved in Appendix C.1.

**Lemma 17.** *On  $\mathcal{E}$  one has:*

$$\forall t \in \mathbb{N}^*, \quad \min_{a \in [K]} N_a(t) \geq tw_{\min} - K .$$

This inequality directly implies the following lower bound:

$$\forall t \geq \frac{2K}{w_{\min}} \implies \min_{a \in [K]} N_a(t) \geq \frac{tw_{\min}}{2} . \quad (18)$$

The proof is organized in 3 steps:

1. We first show that, on event  $\mathcal{E}$ , the optimal vector  $\boldsymbol{w}$  and the sampling frequency vector  $\boldsymbol{N}(t)/t$  are very close for any  $t \geq T_1$ , where  $T_1$  is a (problem-dependent) constant. To do so, we will make use of the regularity results of Section 3.4 and the fact that the confidence regions shrink with time.
2. Then, we control the event  $(\tau_\delta > t) \cap \mathcal{E}$  for  $t > T \log(1/\delta)$  by another event for which we can easily bound the probability using Hoeffding's inequality. This inclusion relies once again on the regularity results of Section 3.4 and on conditions on  $\delta$ , in particular we will require to have  $T \log(1/\delta) \geq T_1$  with  $T_1$  obtained at Step 1.
3. Finally, we derive the two bounds of the theorem from Hoeffding's inequality and elementary calculations.

The proof uses some technical lemmas introduced and shown in Appendix C.

### Step 1: controlling the difference between vectors $\boldsymbol{w}$ and $\boldsymbol{N}(t)/t$

**In this Step we assume that event  $\mathcal{E}$  occurs.**

Let  $t \geq \frac{2K}{w_{\min}}$ . Equation (18) implies that

$$\forall a \in [K], \quad \ell_a(t) = 2\sqrt{\frac{\log(4N_a(t)K/\gamma)}{N_a(t)}} \leq \sqrt{8\frac{\log(4tK/\gamma)}{tw_{\min}}} =: L(t) .$$

$L(t)$  is an arm-independent bound on the half-length of the confidence interval of each  $\mu_a$ . In other words,  $\|\tilde{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}\|_\infty \leq L(t)$  as we are on event  $\mathcal{E}$ . Note that  $L(t)$  is deterministic and goes to 0 as  $t$  goes to  $+\infty$ . This control of  $\|\tilde{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}\|_\infty$  together with Theorem 15 allows to control the difference between  $\boldsymbol{w}$  and  $\tilde{\boldsymbol{w}}(t)$  for  $t$  large enough, as the following Lemma claims.

**Lemma 18.** *Let*

$$T_0 = \max \left( \frac{224^2}{\Delta_{\min}^2 w_{\min}} \log \left( \frac{2 \times 224^2 eK}{\Delta_{\min}^2 w_{\min} \gamma} \right), \frac{2K}{w_{\min}} \right). \quad (19)$$

*Then for every  $t \geq T_0$ , one has, introducing  $\varepsilon_t = \frac{80L(t)}{\Delta_{\min}}$ :*

$$\forall a \in [K], \quad w_a(1 - \varepsilon_t) \leq \tilde{w}_a(t) \leq w_a(1 + \varepsilon_t). \quad (20)$$

*Proof.* Let  $t \geq \frac{2K}{w_{\min}}$  and assume that  $t$  is such that  $4L(t) < \Delta_{\min}$ . On event  $\mathcal{E}$ , one has  $\boldsymbol{\mu} \in \mathcal{CR}_{\boldsymbol{\mu}}(t) = \prod_{a \in [K]} [\underline{\mu}_a(t), \bar{\mu}_a(t)]$ , hence for any  $a \neq 1$ :

$$\underline{\mu}_1(t) - \bar{\mu}_a(t) \geq \mu_1 - 2L(t) - (\mu_a + 2L(t)) \geq \Delta_a - 4L(t) > 0$$

so that the confidence interval for  $\mu_1$  is strictly above all other confidence intervals. Hence  $\tilde{\boldsymbol{\mu}}(t)$  has a unique optimal arm which is arm 1.

For each arm  $a \neq 1$ , define  $\tilde{\Delta}_a(t) = \Delta_a(\tilde{\boldsymbol{\mu}}(t)) = \tilde{\mu}_1(t) - \tilde{\mu}_a(t)$ . Then

$$\begin{aligned} \tilde{\Delta}_a(t)^2 &\leq (\Delta_a + 2L(t))^2 = \Delta_a^2 \left( 1 + \frac{4L(t)}{\Delta_a} + \frac{4L(t)^2}{\Delta_a^2} \right) \leq \Delta_a^2 \left( 1 + \frac{8L(t)}{\Delta_{\min}} \right) \\ \text{and } \tilde{\Delta}_a(t)^2 &\geq (\Delta_a - 2L(t))^2 = \Delta_a^2 \left( 1 - \frac{4L(t)}{\Delta_a} + \frac{4L(t)^2}{\Delta_a^2} \right) \geq \Delta_a^2 \left( 1 - \frac{8L(t)}{\Delta_{\min}} \right). \end{aligned}$$

If  $t$  is such that  $\frac{8L(t)}{\Delta_{\min}} \leq 1/7$  (this condition is stronger than  $4L(t) < \Delta_{\min}$ ), we can apply Theorem 15 which gives

$$\forall a \in [K], \quad w_a(1 - \varepsilon_t) \leq \tilde{w}_a(t) \leq w_a(1 + \varepsilon_t).$$

It remains to understand when the condition  $\frac{8L(t)}{\Delta_{\min}} \leq 1/7$  holds. We have:

$$\frac{8L(t)}{\Delta_{\min}} \leq 1/7 \iff \frac{\log(4tK/\gamma)}{t} \leq \frac{\Delta_{\min}^2 w_{\min}}{(7 \times 8)^2 \times 8} = \frac{\Delta_{\min}^2 w_{\min}}{2 \times 112^2}$$

and this inequality is satisfied, by Lemma 27, for

$$t \geq \frac{224^2}{\Delta_{\min}^2 w_{\min}} \log \left( \frac{2 \times 224^2 eK}{\Delta_{\min}^2 w_{\min} \gamma} \right).$$

Combining with the initial condition  $t \geq \frac{2K}{w_{\min}}$  leads to the definition of  $T_0$ .  $\square$

As each  $N_a(t)/t$  is nearly the Cesaro sum of the  $(\tilde{w}_a(s))_{0 \leq s \leq t-1}$  (see Lemma 26), and as  $\varepsilon_t \rightarrow_{t \rightarrow +\infty} 0$ , we are able to control the difference between  $\mathbf{w}$  and  $\mathbf{N}(t)/t$  after a deterministic time  $T_1$ .

**Lemma 19.** *Fix  $\eta \in (0, 1)$  and let*

$$T_1 = \frac{\max(640^2, 8K)}{\eta^2 \Delta_{\min}^2 w_{\min}^2} \log \left( \frac{2 \times 640^2 eK}{\eta^2 \Delta_{\min}^2 w_{\min} \gamma} \right). \quad (21)$$

*Then for any  $t \geq T_1$  one has:*

$$\forall a \in [K], \quad w_a(1 - \eta) \leq \frac{N_a(t)}{t} \leq w_a(1 + \eta). \quad (22)$$

*Proof.* Let  $T_0$  be defined by Equation (19). Let  $t > T_0$  and  $a \in [K]$ . Equation (20) of Lemma 18 gives:

$$\left| \sum_{s=0}^{t-1} \tilde{w}_a(s) - tw_a \right| \leq \sum_{s=0}^{T_0-1} |\tilde{w}_a(s) - w_a| + \sum_{s=T_0}^{t-1} |\tilde{w}_a(s) - w_a| \leq T_0 + w_a \sum_{s=T_0}^{t-1} \varepsilon_s.$$

By definition of  $\varepsilon_t$  one has:

$$\sum_{s=T_0}^{t-1} \varepsilon_s = \frac{80\sqrt{8}}{\Delta_{\min}\sqrt{w_{\min}}} \sum_{s=T_0}^{t-1} \sqrt{\frac{\log(4sK/\gamma)}{s}} \leq \frac{80\sqrt{8}\sqrt{\log(4tK/\gamma)}}{\Delta_{\min}\sqrt{w_{\min}}} \sum_{s=T_0}^{t-1} \frac{1}{\sqrt{s}} \leq \frac{80\sqrt{8}\sqrt{t\log(4tK/\gamma)}}{\Delta_{\min}\sqrt{w_{\min}}}$$

so that we have, using Lemma 26:

$$\begin{aligned} \left| \frac{N_a(t)}{t} - w_a \right| &\leq \frac{1}{t} \left[ \left| N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s) \right| + \left| \sum_{s=0}^{t-1} \tilde{w}_a(s) - w_a \right| \right] \\ &\leq \frac{K + T_0}{t} + w_a \frac{80\sqrt{8}\sqrt{\log(4tK/\gamma)}}{\Delta_{\min}\sqrt{w_{\min}t}} \\ &\leq w_a \left( \frac{K + T_0}{tw_{\min}} + \frac{80\sqrt{8}\sqrt{\log(4tK/\gamma)}}{\Delta_{\min}\sqrt{w_{\min}t}} \right). \end{aligned}$$

Thus the conclusion of the Lemma holds when:

$$\max \left( \frac{K + T_0}{tw_{\min}}, \frac{80\sqrt{8}\sqrt{\log(4tK/\gamma)}}{\Delta_{\min}\sqrt{w_{\min}t}} \right) \leq \frac{\eta}{2}$$

and this inequality is satisfied, using Lemma 27, when:

$$t \geq \max \left( \frac{2}{\eta} \frac{K + T_0}{w_{\min}}, \frac{640^2}{\eta^2 \Delta_{\min}^2 w_{\min}} \log \left( \frac{2 \times 640^2 eK}{\eta^2 \Delta_{\min}^2 w_{\min} \gamma} \right) \right).$$

The definition of  $T_0$  implies  $K + T_0 \leq \frac{4 \max(112^2, K)}{\Delta_{\min}^2 w_{\min}} \log \left( \frac{2 \times 224^2 eK}{\Delta_{\min}^2 w_{\min} \gamma} \right)$ , hence the inequality still holds for

$$t \geq \max \left( \frac{8 \max(112^2, K)}{\eta \Delta_{\min}^2 w_{\min}^2} \log \left( \frac{2 \times 224^2 eK}{\Delta_{\min}^2 w_{\min} \gamma} \right), \frac{640^2}{\eta^2 \Delta_{\min}^2 w_{\min}} \log \left( \frac{2 \times 640^2 eK}{\eta^2 \Delta_{\min}^2 w_{\min} \gamma} \right) \right)$$

and  $T_1$  is greater than this lower bound.  $\square$

## Step 2: a useful inclusion of events

We want to control the event  $(\tau_\delta > t) \cap \mathcal{E}$  for  $t > T \log(1/\delta)$ . For  $\delta$  small enough, we have the following inclusion of events.

**Lemma 20.** Fix  $\eta \in (0, 0.15]$  and let  $\delta$  be such that

$$T \log(1/\delta) \geq T_1 \tag{C1}$$

where  $T_1$  is defined by Equation (21) and

$$\log(1/\delta) > \frac{4}{\eta} \log \left( \frac{8eTR^{1/2}}{\eta} \right). \tag{C2}$$

Then for any  $C \in (0, 1]$ :

$$\forall t \geq (1 + C) \frac{(1 + \eta)^2}{(1 - \eta)^2} T \log(1/\delta), \quad (\tau_\delta > t) \cap \mathcal{E} \subseteq \left( \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t)\|_\infty \geq \frac{C}{T} \right) \cap \mathcal{E}.$$

**Remark 21.** Latter, we will use this Lemma with  $C = \frac{1}{\log^{\frac{1}{3}}(1/\delta)}$ .

*Proof.* Assume in the following that  $T \log(1/\delta) \geq T_1$  and let  $t \geq T \log(1/\delta)$ . By definition of  $T_1$  and Lemma 19, one has

$$\max_{a \in [K]} \left| \frac{w_a - N_a(t)/t}{w_a} \right| \leq \eta. \quad (23)$$

Then using Proposition 16 and Equation (8):

$$\begin{aligned} (\tau_\delta > t) \cap \mathcal{E} &\subseteq \left( Z(t) = tg(\hat{\boldsymbol{\mu}}(t), \mathbf{N}(t)/t) \leq \beta(t, \delta) \right) \cap \mathcal{E} \\ &\subseteq \left( t \frac{(1-\eta)^2}{1+\eta} \left( g(\boldsymbol{\mu}, \mathbf{w}) - \frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t)\|_\infty}{2} \right) \leq \beta(t, \delta) \right) \cap \mathcal{E} \\ &\subseteq \left( \frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t)\|_\infty}{2} \geq \frac{1}{T} - \frac{1+\eta}{(1-\eta)^2} \frac{\beta(t, \delta)}{t} \right) \cap \mathcal{E}. \end{aligned}$$

Consider now

$$f(t) = \frac{1+\eta}{(1-\eta)^2} \frac{\beta(t, \delta)}{t} = \frac{1+\eta}{(1-\eta)^2} \frac{\log\left(\frac{Rt^\alpha}{\delta}\right)}{t}.$$

As  $\alpha \leq 2$ , one can check that  $f$  is decreasing on  $(4, +\infty)$ . Let us show that

$$\forall C \in (0, 1], \quad f\left((1+C) \frac{(1+\eta)^2}{(1-\eta)^2} T \log(1/\delta)\right) \leq \frac{1}{(1+C)T}. \quad (24)$$

Fix  $C \in (0, 1]$ . As  $\alpha \leq 2$  and as  $\eta \leq 0.15$  is such that  $\frac{(1+\eta)^2}{(1-\eta)^2} \leq 2$ , we have:

$$\begin{aligned} f\left((1+C) \frac{(1+\eta)^2}{(1-\eta)^2} T \log(1/\delta)\right) &\leq \frac{1+\eta}{(1-\eta)^2} \frac{\log\left(\frac{R(4T \log(1/\delta))^2}{\delta}\right)}{(1+C) \frac{(1+\eta)^2}{(1-\eta)^2} T \log(1/\delta)} \\ &\leq \frac{1}{(1+C)T} \frac{1}{1+\eta} \left( 1 + 2 \frac{\log(4R^{1/2} T \log(1/\delta))}{\log(1/\delta)} \right). \end{aligned}$$

hence Inequality (24) is satisfied if

$$\log(4R^{1/2} T \log(1/\delta)) \leq \frac{\eta}{2} \log(1/\delta)$$

which is the case, by Lemma 27, when:

$$\log(1/\delta) > \frac{4}{\eta} \log\left(\frac{8eTR^{1/2}}{\eta}\right).$$

Finally when Inequality (24) holds we have for  $t \geq (1+C) \frac{(1+\eta)^2}{(1-\eta)^2} T \log(1/\delta)$ :

$$\begin{aligned} (\tau_\delta > t) \cap \mathcal{E} &\subseteq \left( \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t)\|_\infty \geq \frac{2}{T} - \frac{2}{(1+C)T} \right) \cap \mathcal{E} \\ &\subseteq \left( \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t)\|_\infty \geq \frac{C}{T} \right) \cap \mathcal{E} \end{aligned}$$

where we use  $C \leq 1$  in the last inclusion.  $\square$

### Step 3: bounding $\mathbb{P}_\mu(\tau_\delta > t \cap \mathcal{E})$ and $\mathbb{E}_\mu[\tau_\delta \mathbf{1}_\mathcal{E}]$ .

Fix  $\eta \in (0, 1]$  and assume in the following that conditions (C1) and (C2) of Lemma 20 are satisfied with  $\eta' = \eta/7 \leq 0.15$ .

We set  $\zeta = \frac{(1+\eta')^2}{(1-\eta')^2}$ . Let  $C \in (0, 1]$ ,  $t > (1+C)\zeta T \log(1/\delta)$  and define

$$\mathcal{E}_t = \left( \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t)\|_\infty \geq \frac{C}{T} \right) \cap \mathcal{E}.$$

Lemmas 20 and 28 – a consequence of Hoeffding’s inequality – (note that Condition (C1) ensures that  $t \geq \frac{2K}{w_{\min}}$ ) give the bound:

$$\mathbb{P}_{\boldsymbol{\mu}}(\tau_{\delta} > t \cap \mathcal{E}) \leq \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_t) \leq 2Kt \exp\left(-\frac{tw_{\min}}{4T^2}C^2\right). \quad (25)$$

By taking  $C = \frac{1}{\log^{\frac{1}{3}}(1/\delta)}$ , we obtained so far that

$$\forall t > \left(1 + \frac{1}{\log^{\frac{1}{3}}(1/\delta)}\right)\zeta T \log(1/\delta), \quad \mathbb{P}_{\boldsymbol{\mu}}(\tau_{\delta} > t \cap \mathcal{E}) \leq 2Kt \exp\left(-\frac{tw_{\min}}{4T^2} \frac{1}{\log^{\frac{2}{3}}(1/\delta)}\right)$$

giving Bound (5) as long as  $(1 + \frac{1}{\log^{\frac{1}{3}}(1/\delta)})\zeta \leq 1 + \eta$ . Note that  $\zeta \leq 1 + 6\eta'$  as  $\eta' \leq 0.15$  so that when

$$\frac{1}{\log^{\frac{1}{3}}(1/\delta)} \leq \frac{\eta'}{2} \iff \log(1/\delta) \geq \frac{8 \times 7^3}{\eta^3} \quad (C3)$$

the condition holds as

$$\left(1 + \frac{1}{\log^{\frac{1}{3}}(1/\delta)}\right)\zeta \leq \left(1 + \frac{\eta'}{2}\right)(1 + 6\eta') \leq 1 + 6.6\eta' \leq 1 + \eta.$$

It remains to focus on the bound of  $\mathbb{E}_{\boldsymbol{\mu}}[\tau_{\delta}\mathbb{1}_{\mathcal{E}}]$ . Using Equation (25) we have:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\mu}}[\tau_{\delta}\mathbb{1}_{\mathcal{E}}] &= \sum_{t=0}^{\lfloor (1+C)\zeta T \log(1/\delta) \rfloor} \mathbb{P}_{\boldsymbol{\mu}}(\tau_{\delta} > t \cap \mathcal{E}) + \sum_{t > (1+C)\zeta T \log(1/\delta)} \mathbb{P}_{\boldsymbol{\mu}}(\tau_{\delta} > t \cap \mathcal{E}) \\ &\leq (1+C)\zeta T \log(1/\delta) + 1 + 2K \sum_{t > (1+C)\zeta T \log(1/\delta)} t \exp\left(-\frac{tw_{\min}}{4T^2}C^2\right). \end{aligned}$$

Define

$$S(C) = \sum_{t > C\zeta T \log(1/\delta)} t \exp\left(-\frac{tw_{\min}}{4T^2}C^2\right).$$

With some technical calculations, one can obtain that:

**Lemma 22.** *One has*

$$S(C) \leq \frac{32T^4}{w_{\min}^2} \exp\left(-\frac{w_{\min}}{4T}C^2 \log(1/\delta)\right) \left(\frac{\log(1/\delta)}{C^2} + \frac{1}{C^4}\right).$$

Once again, taking  $C = \frac{1}{\log^{\frac{1}{3}}(1/\delta)}$  leads to

$$\begin{aligned} S(C) &\leq \frac{32T^4}{w_{\min}^2} \exp\left(-\frac{w_{\min}}{4T} \log^{\frac{1}{3}}(1/\delta)\right) \left(\log^{\frac{5}{3}}(1/\delta) + \log^{\frac{4}{3}}(1/\delta)\right) \\ &\leq \frac{64T^4}{w_{\min}^2} \exp\left(-\frac{w_{\min}}{4T} \log^{\frac{1}{3}}(1/\delta)\right) \log^2(1/\delta) \end{aligned}$$

thus

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_{\delta}\mathbb{1}_{\mathcal{E}}] \leq \zeta \left(1 + \frac{1}{\log^{\frac{1}{3}}(1/\delta)}\right) T \log(1/\delta) + 1 + \frac{2^7 K T^4}{w_{\min}^2} \exp\left(-\frac{w_{\min}}{4T} \log^{\frac{1}{3}}(1/\delta)\right) \log^2(1/\delta).$$

Under Condition (C3) we get

$$\zeta \left(1 + \frac{1}{\log^{\frac{1}{3}}(1/\delta)}\right) T \log(1/\delta) + 1 \leq (1 + 6.6\eta') T \log(1/\delta) + 1 \leq (1 + \eta) T \log(1/\delta)$$

and obtain the Bound (6) claimed in the theorem. Combining conditions (C1), (C2) and (C3) together, one can define  $\delta_0$  satisfying:

$$\log(1/\delta_0) \geq \frac{7^3 \times \max(2 \times 160^2, K)}{\eta^3 \Delta_{\min}^2 w_{\min}^2} \log \left( \frac{7^2 \times 2 \times 640^2 e K R^{1/2}}{\eta^2 \Delta_{\min}^2 w_{\min} \gamma} \right),$$

with some simplifications allowed by Equation (17) of Proposition 11.

## 5 Numerical experiments

In this section, we discuss the behavior and performance of our strategy EXPLORATION-BIASED SAMPLING, for practical values of confidence  $\delta$ . We will especially propose a comparison with TRACK-AND-STOP and CHERNOFF-RACING. Let us first recall the three strategies and introduce the parameters used in our simulations.

**TRACK-AND-STOP** The strategy ([Garivier and Kaufmann, 2016]) tracks the optimal weights  $\mathbf{w}(\boldsymbol{\mu})$  by estimating it by  $\mathbf{w}(\hat{\boldsymbol{\mu}}(t))$ . Using the current weight vector could lead to an under-sampling of some arms suffering from bad initial observations, so that a forced exploration is required to tackle this issue: the strategy ensures that each  $N_a(t)$  grows at least in  $\sqrt{t}$ . The decision process is the same as the one presented for EXPLORATION-BIASED SAMPLING.

**CHERNOFF-RACING** With the racing algorithm, arms are eliminated one by one, the strategy is divided into rounds in which all currently active arms are sampled once. At the end of the round, a decision is made to keep or eliminate the current worst arm. Several decision rules are possible, we will use the Chernoff rule presented in [Garivier and Kaufmann, 2016], which eliminates arm  $b$  at the end of round  $r$  if

$$Z_{\hat{a}_r, b}(t) = \frac{r}{4} (\hat{\mu}_{\hat{a}_r}(t) - \hat{\mu}_b(t))^2 > \beta(t, \delta)$$

where  $\hat{a}_r$  (resp.  $t$ ) is the best arm (resp. the time) at the end of round  $r$ .

**EXPLORATION-BIASED SAMPLING** We ran our experiments with confidence lengths  $C_\gamma(s) = \sqrt{\frac{\log(s/\gamma)}{s}}$ .

With all strategies, we used the same threshold

$$\beta(t, \delta) = \log((\log(t) + 1)/\delta)$$

which seems numerically sufficient for  $\delta$ -correctness, at least for EXPLORATION-BIASED SAMPLING and TRACK-AND-STOP. We now discuss the numerical pros and cons of EXPLORATION-BIASED SAMPLING comparing to TRACK-AND-STOP and CHERNOFF-RACING.

**Improving the stability of Track-and-Stop** In Section 1, we highlighted that TRACK-AND-STOP had several weaknesses, especially a forced exploration parameter and a non-interpretable and unstable sampling strategy during the first rounds, at which the strategy already tracks the optimal weights according to the current (poor) estimates. On Figures 2 and 3 we can see how EXPLORATION-BIASED SAMPLING fixes those behaviours. During the first rounds, as for a racing algorithm, a uniform sampling is observed as the confidence intervals on all arms are not separated. Then the best arms are sampled more and more often, but still in a more cautious way than TRACK-AND-STOP. We see how stable the sampling strategies are comparing to TRACK-AND-STOP during the first rounds: on Figure 3 the targeted weights of EXPLORATION-BIASED SAMPLING are stable and separate from each other cautiously (note that the three last arms still have the same weight at time 1200) whereas for TRACK-AND-STOP, we observe an important variation of the targeted weights with time. As a matter of facts, there is a clear discontinuity each time the estimated best arm changes, as we can see with the red and green arms. We also remark that TRACK-AND-STOP uses the forced exploration at regular rounds (giving the yellow and blue peaks). With EXPLORATION-BIASED SAMPLING this force exploration is unnecessary as a natural exploration is always performed (see Lemma 3).



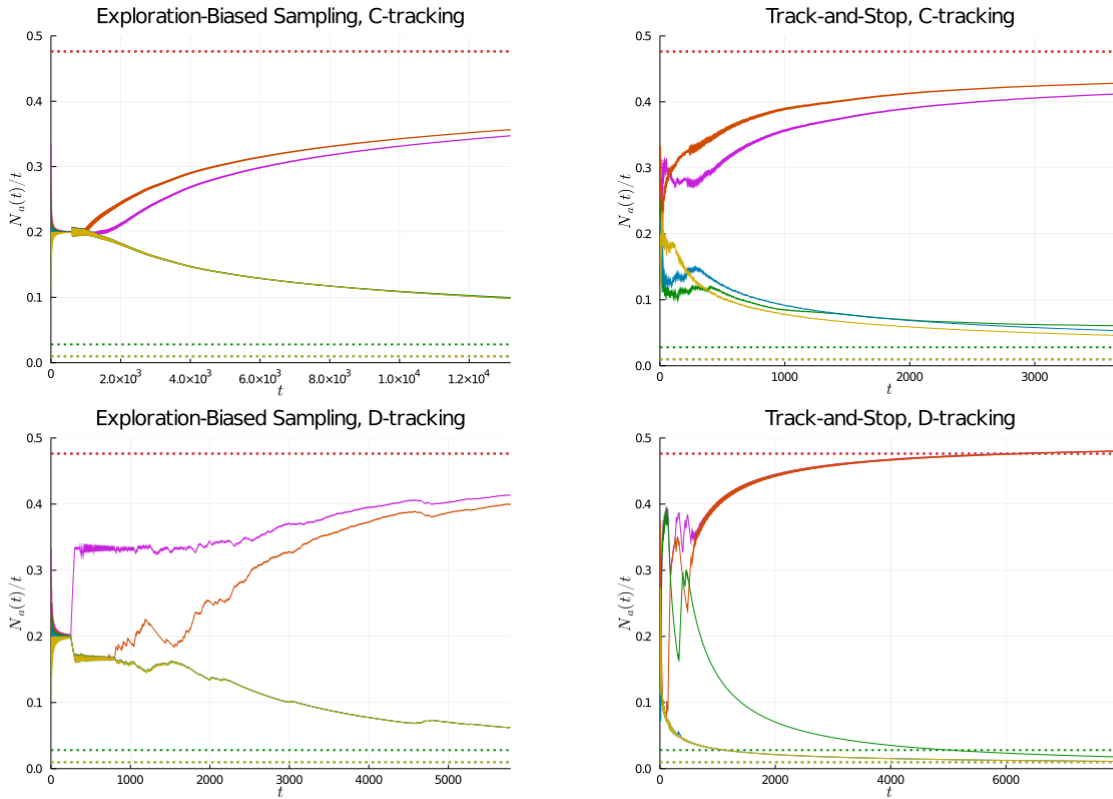


Figure 2: Evolution of the sampling frequencies  $N(t)/t$  when running EXPLORATION-BIASED SAMPLING or TRACK-AND-STOP, with  $\boldsymbol{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$ ,  $\delta = 0.01$  and  $\gamma = 0.2$ . The values of  $\boldsymbol{w}(\boldsymbol{\mu}) = (0.477, 0.476, 0.028, 0.010, 0.010)$  are dotted.

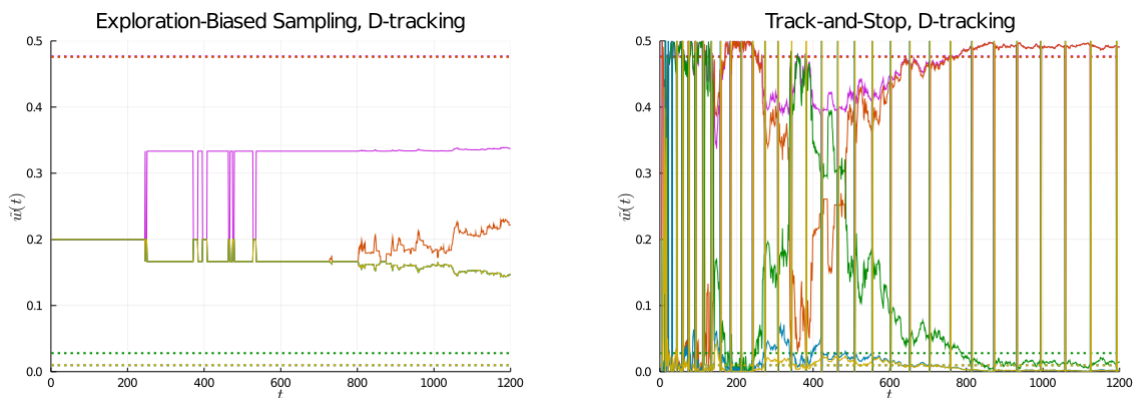


Figure 3: Evolution of the exploration-biased optimal vectors  $\tilde{\boldsymbol{w}}(t)$  during the first 1200 rounds when running EXPLORATION-BIASED SAMPLING or TRACK-AND-STOP, with  $\boldsymbol{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$ ,  $\delta = 0.01$  and  $\gamma = 0.2$ .

**Comparisons of the three strategies** The cost of the exploration-biased weights of our strategy is that it takes a little longer for the proportions of draws of EXPLORATION-BIASED SAMPLING to converge to the optimal weights. This results in a slightly larger stopping time than TRACK-AND-STOP. This is what can be observed on Table 1, where we present the performances of EXPLORATION-BIASED SAMPLING, TRACK-AND-STOP and CHERNOFF-RACING with two scenarios and a set of parameters. EXPLORATION-BIASED SAMPLING globally performs correctly but we see that TRACK-AND-STOP and CHERNOFF-RACING are always a little more efficient. Note that when increasing  $\gamma$ , the confidence intervals reduces so that the targeted weights are closer to  $\mathbf{w}$ , which improves the performance of the algorithm. Also note that the cautiousness of the strategy during the first rounds is compensated after long times, so that when  $\delta$  is very small the relative performance of TRACK-AND-STOP and EXPLORATION-BIASED SAMPLING get closer. Of course, EXPLORATION-BIASED SAMPLING overperforms CHERNOFF-RACING in the long run when the optimal weights are far from the sampling proportions of CHERNOFF-RACING (e.g. when  $w_1 \gg w_2$ ).

Concerning CHERNOFF-RACING, the strategy shows great performance with both  $\mu^{(1)}$  and  $\mu^{(2)}$ . As this strategy samples the two last arms of the race equally often, it can be optimal only when  $\mathbf{w}(\mu)$  has its two highest components of similar value which happens when the two best arms are well separated from the others (for instance with bandit  $\mu^{(1)}$ ). The case of  $\mu^{(2)}$  is a bit different: the problem being easy, any strategy performs well. However, note that there is no theoretical guarantees for CHERNOFF-RACING, and that the strategy leads to a few more misidentifications in our experiments that might be linked to the stopping rule we chose here (which, for fairness reasons, was taken identical to that of the other algorithms).

Finally, note that D-tracking shows better performance than C-tracking, either for EXPLORATION-BIASED SAMPLING and TRACK-AND-STOP. D-tracking indeed benefits directly of the current estimate of  $\mu$ , while the impact is diluted in time with C-tracking.

Bandit	$\delta$	$\gamma$	$T \text{kl}(\delta, 1 - \delta)$	EBS C	TaS C	EBS D	TaS D	Racing
$\mu^{(1)}$	0.1	0.05	1476	4727	3597	4191	3477	3124
$\mu^{(1)}$	0.01	0.05	3782	7363	5664	6330	5584	5419
$\mu^{(1)}$	0.01	0.2	3782	7090	5664	6136	5584	5419
$\mu^{(1)}$	$10^{-5}$	0.2	9669	13801	12181	12376	11439	11557
$\mu^{(2)}$	0.1	0.05	135	476	367	470	322	405
$\mu^{(2)}$	0.01	0.05	347	708	588	699	485	542

Table 1: Empirical expected number of draws  $\mathbb{E}_\mu[\tau_\delta]$ , averaged over  $N = 1000$  experiments:  $\mu^{(1)} = (0.9, 0.8, 0.6, 0.4, 0.4)$ ,  $\mathbf{w}(\mu^{(1)}) = (0.477, 0.476, 0.028, 0.010, 0.010)$ ;  $\mu^{(2)} = (0.9, 0.5, 0.45, 0.4)$ ,  $\mathbf{w}(\mu^{(2)}) = (0.375, 0.286, 0.195, 0.144)$

Lastly, we numerically investigated the behavior of  $\tau_\delta$  for rather large values of the risk  $\delta$  that are not covered by the theory. On Figure 4, we can see the empirical distribution and tail of  $\tau_\delta$  for fixed parameters of EXPLORATION-BIASED SAMPLING and a bandit parameter. The histogram shows that it is quite rare to require less than  $T \log(1/\delta)$  samples before making a decision, but the right tail of the histogram shows a rather large spread after this value. Decreasing  $\delta$  leads to a better concentration. Concerning the tail of  $\tau_\delta$ , Bound (5) proves an exponential bound of the type  $\exp(-Ct)$ . Empirically, Figure 4 reveals a Gaussian behaviour in  $\exp(-Ct^2)$ .

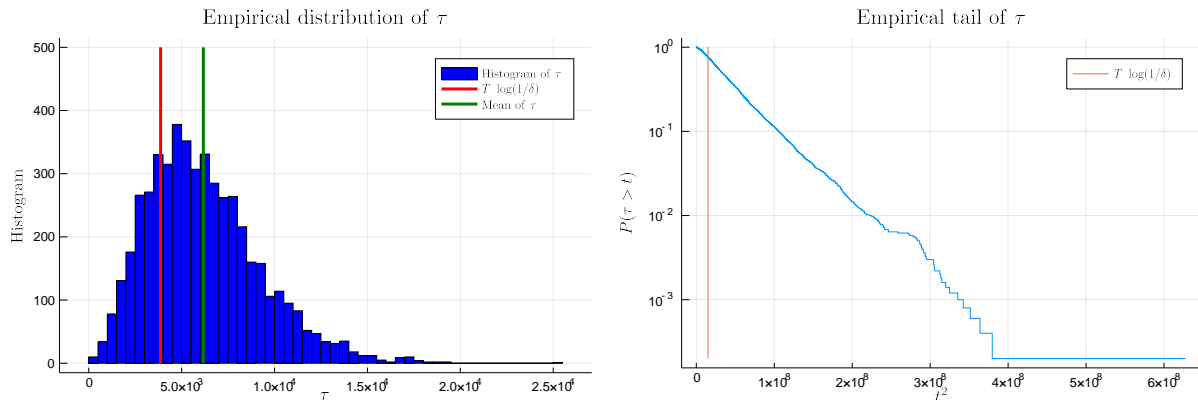


Figure 4: Empirical distribution and tail of  $\tau_\delta$ , when running 5000 experiments of EXPLORATION-BIASED SAMPLING with parameters  $\delta = 0.01$  and  $\gamma = 0.2$ , with bandit  $\boldsymbol{\mu}^{(1)} = (0.9, 0.8, 0.6, 0.4, 0.4)$ .

## 6 Conclusion

We introduced EXPLORATION-BIASED SAMPLING, a new strategy for the problem of Best Arm Identification with fixed confidence. In addition to asymptotic optimal results, we proved non-asymptotic bounds for this strategy in the case of (sub-)Gaussian bandits. Those finite risk bounds were made possible by a new analysis of the sample complexity optimization problem, and by the design of our strategy which tackles the shortcomings of TRACK-AND-STOP: the procedure ensures exploration in an unforced way and stabilizes the sampling strategy by observing uniformly before having a high certainty that one arm is better than another.

In future work, we will consider a generalization of this work to non-Gaussian models, which requires to adapt our approach of the sample-complexity optimization problem (a technically challenging task for which the simple and clean arguments developed here are likely to be replaced by much more involved derivations). In addition, it will be necessary to correctly modify the confidence intervals on the arm means in a way that ensures exploration. Another direction of improvement will be to investigate if similar analysis and strategies are possible for the problem of  $\varepsilon$ -best arm identification.

## Acknowledgements

The authors acknowledge the support of the Project IDEXLYON of the University of Lyon, in the framework of the Programme Investissements d’Avenir (Grant number ANR-16-IDEX-0005), and Chaire SeqALO (Grant number ANR-20-CHIA-0020-01).

## References

- [Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities : a non asymptotic theory of independence*. Oxford University Press.
- [Chernoff, 1959] Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.
- [Degenne and Koolen, 2019] Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

- [Degenne et al., 2019] Degenne, R., Koolen, W. M., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- [Even-Dar et al., 2006] Even-Dar, E., Mannor, S., and Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105.
- [Garivier et al., 2019] Garivier, A., Hadiji, H., Ménard, P., and Stoltz, G. (2019). KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints.
- [Garivier and Kaufmann, 2016] Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence.
- [Kalyanakrishnan et al., 2012] Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). Pac subset selection in stochastic multi-armed bandits. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 1.
- [Kaufmann and Kalyanakrishnan, 2013] Kaufmann, E. and Kalyanakrishnan, S. (2013). Information complexity in bandit subset selection. In *Proceeding of the 26th Conference On Learning Theory*.
- [Kocák and Garivier, 2020] Kocák, T. and Garivier, A. (2020). Best arm identification in spectral bandits. *International Joint Conference on Artificial Intelligence*.
- [Lattimore and Szepesvári, 2020] Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.
- [Ménard, 2019] Ménard, P. (2019). Gradient ascent for active exploration in bandit problems.
- [Russo, 2016] Russo, D. (2016). Simple bayesian algorithms for best arm identification. In *Proceedings of the 29th Conference On Learning Theory*.
- [Shang et al., 2019] Shang, X., de Heide, R., Kaufmann, E., Ménard, P., and Valko, M. (2019). Fixed-confidence guarantees for bayesian best-arm identification. *CoRR*, abs/1910.10945.

## Appendix

### Outline

The appendix is organized as follows:

- A. Proof of the time-uniform confidence regions guarantees for  $\boldsymbol{\mu}$  (Lemma 2)
- B. Proofs of the results on the sample complexity for Gaussian arms (Section 3)
- C. Technical results associated to the proof of Theorem 5 (complements to Section 4)
- D. Asymptotic analysis of EXPLORATION-BIASED SAMPLING (Theorems 6 and 7)

Without loss of generality (see [Garivier et al., 2019]), we assume that for any  $a \in [K]$ ,  $(X_{a,n})_{n \geq 1}$  is a sequence of random variables independent and identically distributed with distribution  $\mathcal{N}(\mu_a, 1)$ , we denote  $\hat{\mu}_{a,n} = \frac{1}{n} \sum_{p=1}^n X_{a,p}$  for all  $n \geq 1$  and assume that

$$\forall t \geq K, \quad \hat{\mu}_a(t) = \hat{\mu}_{a, N_a(t)}. \quad (26)$$

## A Proof of Lemma 2

By union bound we only have to show that for any  $\gamma \in (0, 1)$  and  $a \in [K]$ :

$$\mathbb{P}_{\boldsymbol{\mu}} \left( \exists t \geq K : |\hat{\mu}_a(t) - \mu_a| \geq C_\gamma(N_a(t)) \right) \leq \gamma.$$

Fix  $\gamma \in (0, 1)$  and  $a \in [K]$ . Note that as all arms are observed once at the beginning (see Algorithm 2), we have  $N_a(K) = 1$ . Thus using Equation (26), we have:

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu}} \left( \exists t \geq K : |\hat{\mu}_a(t) - \mu_a| \geq C_\gamma(N_a(t)) \right) &= \mathbb{P}_{\boldsymbol{\mu}} \left( \exists t \geq K : |\hat{\mu}_{a, N_a(t)} - \mu_a| \geq C_\gamma(N_a(t)) \right) \\ &= \mathbb{P}_{\boldsymbol{\mu}} \left( \exists n \in \mathbb{N}^* : |\hat{\mu}_{a,n} - \mu_a| \geq C_\gamma(n) \right). \end{aligned}$$

Then we use a peeling trick (see for instance [Boucheron et al., 2013]):

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu}} \left( \exists n \in \mathbb{N}^* : |\hat{\mu}_{a,n} - \mu_a| \geq C_\gamma(n) \right) &\leq \sum_{k \geq 0} \mathbb{P} \left( \exists n \in [2^k, 2^{k+1}] : \left| \frac{1}{n} \sum_{p=1}^n (X_{a,p} - \mu_a) \right| \geq C_\gamma(n) \right) \\ &= \sum_{k \geq 0} \mathbb{P} \left( \exists n \in [2^k, 2^{k+1}] : \left| \sum_{p=1}^n X_{a,p} - \mu_a \right| \geq n C_\gamma(n) \right) \\ &\stackrel{(a)}{\leq} \sum_{k \geq 0} \mathbb{P} \left( \exists n \in [0, 2^{k+1}] : \left| \sum_{p=1}^n X_{a,p} - \mu_a \right| \geq 2^k C_\gamma(2^k) \right) \\ &\stackrel{(b)}{\leq} 2 \sum_{k \geq 0} \exp \left( - \frac{(2^k C_\gamma(2^k))^2}{2 \times 2^{k+1}} \right) \\ &= 2 \sum_{k \geq 0} \exp \left( - \log(2^{k+2}/\gamma) \right) = 2\gamma \sum_{k \geq 0} \frac{1}{2^{k+2}} = \gamma. \end{aligned}$$

(a) is obtained using the fact that  $n \mapsto n C_\gamma(n)$  is non-decreasing and (b) is a well-known inequality for the sum of sub-Gaussian variables, see for instance [Lattimore and Szepesvári, 2020, Theorem 9.2].

## B Proofs of results presented in Section 3

In this appendix, we start with the proof of Proposition 9. Then we prove the consequences developed in Section 3.

For the sake of simplicity, we assume that  $a^* = 1$  in this appendix, except in the last section where there is no uniqueness assumption on the best arm of the bandits.

### B.1 Solving the optimization problem

*Proof of Equation (7).* Let  $\mathbf{v} \in \Sigma_K$ . One has:

$$\begin{aligned} g(\boldsymbol{\mu}, \mathbf{v}) &= \inf_{\lambda \in \text{Alt}(\boldsymbol{\mu})} \sum_{a \in [K]} v_a \frac{(\mu_a - \lambda_a)^2}{2} \\ &= \frac{1}{2} \min_{a \neq 1} \inf_{\lambda_1 < \lambda_a} v_1 (\mu_1 - \lambda_1)^2 + v_a (\mu_a - \lambda_a)^2 \\ &= \frac{1}{2} \min_{a \neq 1} \inf_{\mu_1 \leq \lambda \leq \mu_a} v_1 (\mu_1 - \lambda)^2 + v_a (\mu_a - \lambda)^2 \\ &= \frac{1}{2} \min_{a \neq 1} \frac{v_1 v_a}{v_1 + v_a} (\mu_1 - \mu_a)^2 \end{aligned}$$

since the minimum is reached at  $\lambda = \frac{v_1 \mu_1 + v_a \mu_a}{v_1 + v_a}$ .  $\square$

*Proof of Proposition 9.* Let us define, for some  $v_1 \in [0, 1]$ :

$$C(v_1) = \max_{v_{2:K} : \mathbf{v} \in \Sigma_K} \min_{a \neq 1} \frac{v_1 v_a}{v_1 + v_a} \Delta_a^2 \quad (27)$$

so that

$$T^{-1} = \max_{\mathbf{v} \in \Sigma_K} g(\boldsymbol{\mu}, \mathbf{v}) = \frac{1}{2} \max_{v_1 \in [0,1]} C(v_1). \quad (28)$$

Fix  $v_1 \in [0, 1]$ . The maximum in Equation (27) is reached for  $v_{2:K}$  such that all the  $(\frac{v_1 v_a}{v_1 + v_a} \Delta_a^2)_{a \neq 1}$  are equal, which happens when the  $(v_a)_{a \neq 1}$  equalize those costs:  $C$  is such that

$$\forall a \neq 1, \quad C = \frac{v_1 v_a}{v_1 + v_a} \Delta_a^2$$

and hence:

$$\forall a \neq 1, \quad v_a = \frac{v_1 C}{v_1 \Delta_a^2 - C}. \quad (29)$$

The fact that  $\mathbf{v} \in \Sigma_K$  yields:

$$\Phi(v_1, C) := v_1 + \sum_{a \neq 1} \frac{v_1 C}{v_1 \Delta_a^2 - C} - 1 = 0. \quad (30)$$

By the implicit function theorem, there exists a mapping  $C(v_1)$  such that  $\Phi(v_1, C(v_1)) = 0$  and

$$\begin{aligned} C'(v_1) &= - \frac{\frac{\partial \Phi}{\partial v_1}(v_1, C(v_1))}{\frac{\partial \Phi}{\partial C}(v_1, C(v_1))} = - \frac{1 + \sum_{a \neq 1} \frac{C(v_1)(v_1 \Delta_a^2 - C(v_1)) - v_1 C(v_1) \Delta_a^2}{(v_1 \Delta_a^2 - C(v_1))^2}}{v_1^2 \sum_{a \neq 1} \frac{\Delta_a^2}{(v_1 \Delta_a^2 - C(v_1))^2}} \\ &= - \frac{1 - \sum_{a \neq 1} \frac{1}{(v_1 \Delta_a^2 / C(v_1) - 1)^2}}{v_1^2 \sum_{a \neq 1} \frac{\Delta_a^2}{(v_1 \Delta_a^2 - C(v_1))^2}}. \end{aligned}$$

Hence  $C(v_1)$  is a smooth non-negative function with a continuous derivative. By Equation (28), it vanishes when  $v_1 \rightarrow 0$  and  $v_1 \rightarrow 1$ , and hence its maximum is reached at a point  $w_1$  where  $C'(w_1) = 0$ . Define  $r = w_1/C(w_1)$  by the relation

$$C'(w_1) = 0 \iff 1 - \sum_{a \neq 1} \frac{1}{\left(\frac{w_1}{C(w_1)} \Delta_a^2 - 1\right)^2} = 0$$

$r$  is the unique solution of  $\phi_{\boldsymbol{\mu}}(r) = 0$ .

Equations (11), (12) and (13) can be respectively derived from (30), (29) and (28). It remains to obtain Equation (14) by combining Equation (12) and the characterization  $\phi_{\boldsymbol{\mu}}(r) = 0$ :

$$\sum_{a \neq 1} w_a^2 = w_1^2 \sum_{a \neq 1} \frac{1}{(r \Delta_a^2 - 1)^2} = w_1^2 (\phi_{\boldsymbol{\mu}}(r) + 1) = w_1^2 .$$

□

*Proof of Corollary 10.* When  $a, b$  are not optimal, the result is a direct consequence of Equation (12) of Proposition 9. It remains to show that  $w_1 > \max_{a \neq 1} w_a$ . Define  $q_a = \frac{1}{r \Delta_a^2 - 1}$  for  $a \in [K]$ . As  $\phi_{\boldsymbol{\mu}}(r) = 0$ , the  $(q_a^2)_{a \neq 1}$  are positive and sum to 1, hence for any  $a \neq 1$ ,  $q_a < 1$  (strict inequality occurs as  $K \geq 3$ ) and thus by Equation (12) of Proposition 9:

$$w_a = w_1 q_a < w_1 .$$

□

## B.2 Proof of Proposition 11

We will use that  $q_a = \frac{1}{r \Delta_a^2 - 1} \leq 1$  for any  $a \neq 1$ , as shown in the proof of Corollary 10 (large inequality occurs when  $K = 2$ ). Let us begin with Equation (16). As we assume  $a^* = 1$ ,  $w_{\max} = w_1$  by Corollary 10. Using Equation (11) of Proposition 9 one has:

- on the one hand

$$\begin{aligned} w_1 &= \left(1 + \sum_{a \neq 1} \frac{1}{r \Delta_a^2 - 1}\right)^{-1} && \text{by Equation (11) of Proposition 9} \\ &\leq \left(1 + \sum_{a \neq 1} \frac{1}{(r \Delta_a^2 - 1)^2}\right)^{-1} && \text{as } q_a \leq 1 \\ &= \frac{1}{2} && \text{as } \phi_{\boldsymbol{\mu}}(r) = 0 \end{aligned}$$

giving the upper bound ;

- on the other hand, by the Cauchy-Schwarz inequality:

$$w_1 \geq \left(1 + \sqrt{(K-1) \sum_{a \neq 1} \frac{1}{(r \Delta_a^2 - 1)^2}}\right)^{-1} = \frac{1}{1 + \sqrt{K-1}} .$$

We now prove Inequalities (15).

- Since  $q_a \leq 1$  or equivalently  $r \Delta_a^2 \geq 2$  for every  $a \neq 1$ ,

$$r \geq \frac{2}{\Delta_{\min}^2} .$$

- Since  $\overline{\Delta^2} = \frac{1}{K-1} \sum_{a \neq 1} \Delta_a^2$ , by convexity of  $x \mapsto \frac{1}{(rx-1)^2}$ :

$$\frac{1}{K-1} \sum_{a \neq 1} \frac{1}{\left(\frac{1+\sqrt{K-1}}{\Delta^2} \Delta_a^2 - 1\right)^2} \geq \frac{1}{\left(\frac{1+\sqrt{K-1}}{\Delta^2} \overline{\Delta^2} - 1\right)^2} = \frac{1}{K-1}$$

and hence  $\phi_{\boldsymbol{\mu}}\left(\frac{1+\sqrt{K-1}}{\Delta^2}\right) \geq 0$ , which by decreasing of  $\phi_{\boldsymbol{\mu}}$  (Lemma 8) gives  $r \geq \frac{1+\sqrt{K-1}}{\Delta^2}$ .

- One can also check that

$$\phi_{\boldsymbol{\mu}}\left(\frac{1+\sqrt{K-1}}{\Delta_{\min}^2}\right) = \sum_{a \neq 1} \frac{1}{\left(\frac{1+\sqrt{K-1}}{\Delta_{\min}^2} \Delta_a^2 - 1\right)^2} - 1 \leq 0$$

so that  $r \leq \frac{1+\sqrt{K-1}}{\Delta_{\min}^2}$ .

Finally, combining the obtained inequalities with Equation (13) yields Equation (17).

To conclude this section, we discuss about the tightness of the proved inequalities.

- First note that when  $K = 2$ , lower and upper bounds match in Inequalities (15), (16) and (17). In that case the problem is easy as we always have  $\boldsymbol{w} = (0.5, 0.5)$ .
- In fact, equalities  $r = 2/\Delta_{\min}^2$ ,  $w_1 = 1/2$  and  $T = 8/\Delta_{\min}^2$  occur if and only if  $K = 2$ . This is because the  $(q_a)_{a \neq 1}$  are positive and sum to 1 (thus  $q_2 = 1$  only when  $K = 2$ ). The presence of other arms thus increases  $r$  and  $T$  while decreases  $w_1$ .
- If there is at least 3 arms, then the remaining equalities  $w_1 = (1 + \sqrt{K-1})^{-1}$ ,  $r = (1 + \sqrt{K-1})/\overline{\Delta^2}$ ,  $r = (1 + \sqrt{K-1})/\Delta_{\min}^2$  and  $T = 2(1 + \sqrt{K-1})^2/\Delta_{\min}^2$  are reached if and only if  $\Delta_{\min} = \Delta_{\max}$ , or in other words  $\Delta_2 = \dots = \Delta_K$ . Indeed, the condition can be obtained by studying the equality cases in the proof above, using the equality case of the Cauchy-Schwarz inequality for  $w_1$ , the strict convexity of  $x \mapsto \frac{1}{(rx-1)^2}$  and the decreasing of  $\phi_{\boldsymbol{\mu}}$  for  $r$  and finally the link  $T = 2r/w_1$  for  $T$ . Note that in that case,  $T$  grows linearly with  $K$ .

### B.3 Computing $r$

At the sight of Proposition 9, it suffices to compute  $r$  to obtain the values of both the optimal weight vector and the sample complexity.

The function  $\phi_{\boldsymbol{\mu}}$  is convex and strictly decreasing on  $(1/\Delta_{\min}^2, +\infty)$  (Lemma 8). Hence, when initialized with a value  $r_0 < r$ , the iterates of a Newton procedure remain smaller than  $r$ . The lower bound of Inequalities (15) of Proposition 11 permits such an initialization. The convergence is quadratic (the number of correct digits roughly doubles at every step), which implies that a few iterations are sufficient to guarantee machine precision. The cost of the algorithm can hence be considered proportional to that of evaluating  $\phi_{\boldsymbol{\mu}}(r)$ , which is linear in the number of arms. See Algorithm 3 for details.



---

**Algorithm 3: OPTIMAL WEIGHTS**


---

**Input:** bandit  $\mu \in \mathcal{G}^*$  with best arm 1  
tolerance parameter  $\text{tol}$  (typically  $10^{-10}$ )

**Output:** optimal weight vector  $w$   
characteristic time  $T$

**for**  $a = 2$  **to**  $K$  **do**  
|  $\Delta_a \leftarrow \mu_1 - \mu_a$   
 $\phi_\mu(r) \leftarrow \sum_{a \neq 1} \frac{1}{(r\Delta_a^2 - 1)^2} - 1$   
 $\phi'_\mu(r) \leftarrow -2 \sum_{a \neq 1} \frac{\Delta_a^2}{(r\Delta_a^2 - 1)^3}$   
 $r \leftarrow \max\left(\frac{2}{\Delta_{\min}^2}, \frac{1 + \sqrt{K-1}}{\Delta^2}\right)$   
**while**  $|\phi_\mu(r)| \geq \text{tol}$  **do**  
|  $r \leftarrow r - \frac{\phi_\mu(r)}{\phi'_\mu(r)}$   
 $w_1 \leftarrow \frac{1}{1 + \sum_{a \neq 1} \frac{1}{r\Delta_a^2 - 1}}$   
**for**  $a = 2$  **to**  $K$  **do**  
|  $w_a \leftarrow \frac{w_1}{r\Delta_a^2 - 1}$   
 $T \leftarrow 2 \frac{r}{w_1}$

---

## B.4 On the monotonicity of the min-max problem

We prove Lemmas 12, 13 and 14. We then use those Lemmas to prove Proposition 1. Recall that we assume  $K \geq 3$  in this section.

*Proof of Lemma 12.*

1. Since

$$\sum_{a \neq 1} \frac{1}{(r\Delta_a'^2 - 1)^2} < \sum_{a \neq 1} \frac{1}{(r\Delta_a^2 - 1)^2} = 1,$$

it holds that  $r' < r$ . It implies that for  $a \neq \{1, b\}$  one has:

$$\frac{1}{r'\Delta_a'^2 - 1} > \frac{1}{r\Delta_a^2 - 1}.$$

As  $K \geq 3$ , such an arm  $a$  exists and hence as  $\phi_\mu(r) = 0 = \phi_{\mu'}(r')$ :

$$\frac{1}{r'\Delta_b'^2 - 1} < \frac{1}{r\Delta_b^2 - 1}$$

or equivalently  $r'\Delta_b'^2 - 1 > r\Delta_b^2 - 1$ .

Combining those inequalities with Equation (12) of Proposition 9, we have for all  $a \notin \{1, b\}$ :

$$\frac{w'_a}{w'_b} = \frac{r'\Delta_b'^2 - 1}{r'\Delta_a'^2 - 1} > \frac{r\Delta_b^2 - 1}{r\Delta_a^2 - 1} = \frac{w_a}{w_b}.$$

Besides,  $w'_1/w'_b = r'\Delta_b'^2 - 1 > r\Delta_b^2 - 1 = w_1/w_b$ . Hence,

$$\frac{1 - w'_b}{w'_b} = \sum_{a \neq b} \frac{w'_a}{w'_b} > \sum_{a \neq b} \frac{w_a}{w_b} = \frac{1 - w_b}{w_b}$$

and thus  $w'_b < w_b$ .

2. For any  $\boldsymbol{\nu} \in \mathcal{G}^*$  with best arm 1, one can see  $\mathbf{w}(\boldsymbol{\nu})$  or its components as a function of  $\boldsymbol{\Delta}^2(\boldsymbol{\nu})$ . Fix  $a \notin \{1, b\}$  and define  $F_a(\boldsymbol{\Delta}^2(\boldsymbol{\nu}))$  as

$$F_a(\boldsymbol{\Delta}^2(\boldsymbol{\nu})) = \frac{1}{w_a(\boldsymbol{\nu})} = \frac{r(\boldsymbol{\nu})\Delta_a(\boldsymbol{\nu})^2 - 1}{w_1(\boldsymbol{\nu})} = (r(\boldsymbol{\nu})\Delta_a^2 - 1) + \sum_{c \neq 1} \frac{r(\boldsymbol{\nu})\Delta_a^2 - 1}{r(\boldsymbol{\nu})\Delta_c^2 - 1}$$

where the right-inequalities are derived from Equations (11) and (12) of Proposition 9. Recall that  $r(\boldsymbol{\nu})$  also depends uniquely on the gaps, as the unique solution of  $\phi_{\boldsymbol{\nu}} = 0$ . In the following calculations we write  $r$  for  $r(\boldsymbol{\nu})$  but the dependency with respect to the gaps is crucial.

Fix  $d_1 = 0$  and  $d_a = \Delta_a^2$  for  $c \neq \{1, b\}$ . We want to see the change of  $F_a$  with respect to  $d_b = \Delta_b^2$ . We can take the partial derivative:

$$\begin{aligned} \frac{\partial F_a}{\partial d_b} &= \frac{\partial r}{\partial d_b} d_a + \sum_{c \neq 1} \left[ \frac{\frac{\partial r}{\partial d_b} d_a}{rd_c - 1} - \frac{rd_a - 1}{(rd_c - 1)^2} \left( \frac{\partial r}{\partial d_b} d_c \right) \right] - \frac{rd_a - 1}{(rd_b - 1)^2} r \\ &= \frac{\partial r}{\partial d_b} d_a \left( 1 + \sum_{c \neq 1} \frac{1}{rd_c - 1} - \frac{rd_c}{(rd_c - 1)^2} \right) + \frac{\partial r}{\partial d_b} \sum_{c \neq 1} \frac{d_c}{(rd_c - 1)^2} - \frac{rd_a - 1}{(rd_b - 1)^2} r \\ &= \frac{\partial r}{\partial d_b} d_a \sum_{c \neq 1} \underbrace{\frac{1 + (rd_c - 1) - rd_c}{(rd_c - 1)^2}}_{=0} + \frac{\partial r}{\partial d_b} \sum_{c \neq 1} \frac{d_c}{(rd_c - 1)^2} - \frac{rd_a - 1}{(rd_b - 1)^2} r \\ &= \frac{\partial r}{\partial d_b} \sum_{c \neq 1} \frac{d_c}{(rd_c - 1)^2} - \frac{rd_a - 1}{(rd_b - 1)^2} r \end{aligned}$$

(to obtain the third equality, we used that  $\sum_{c \neq 1} \frac{1}{(rd_c - 1)^2} = 1$  by definition of  $r$ ).

It remains to see that  $\frac{\partial r}{\partial d_b}$  is nonpositive, that is that  $r$  is nondecreasing when  $\Delta_b$  increases. In fact, we already remark that by showing that  $r' < r$  in the first part of the proof of Lemma 12. Note that one can also use the implicit function theorem to obtain

$$\frac{\partial r}{\partial d_b} = - \frac{r(rd_b - 1)^{-3}}{\sum_{c \neq 1} d_c (rd_c - 1)^{-3}} < 0.$$

Hence  $\frac{\partial F_a}{\partial d_b} < 0$ , so that as  $\Delta_b' > \Delta_b$ :

$$\frac{1}{w_a} = F_a(\boldsymbol{\Delta}^2) > F_a(\boldsymbol{\Delta}'^2) = \frac{1}{w'_a} \quad \text{giving} \quad w'_a > w_a.$$

3. Using Equations (8) and (7):

$$T'^{-1} = \frac{1}{2} \min_{a \neq 1} \frac{w'_1 w'_a}{w'_1 + w'_a} \Delta_a'^2 \geq \frac{1}{2} \min_{a \neq 1} \frac{w'_1 w'_a}{w'_1 + w'_a} \Delta_a^2 > \frac{1}{2} \min_{a \neq 1} \frac{w_1 w_a}{w_1 + w_a} \Delta_a^2 = T^{-1},$$

the first inequality comes from the assumption on  $\boldsymbol{\mu}$  and  $\boldsymbol{\mu}'$ , and the second is a consequence of the uniqueness of the optimal weight vector  $\mathbf{w}$  and the fact that  $\mathbf{w} \neq \mathbf{w}'$ , as previously obtained.

□

Before proving Lemmas 13 and 14, we show the following result.

**Lemma 23.** *Assume that there exists  $\kappa > 0$  such that  $\Delta'_a = \kappa\Delta_a$  for any  $a \neq 1$ . Then  $\mathbf{w}' = \mathbf{w}$ .*

*Proof of Lemma 23.* As  $r$  is the unique solution of  $\phi_{\boldsymbol{\mu}}(r) = 0$ , one has:

$$0 = \phi_{\boldsymbol{\mu}}(r) = \sum_{a \neq 1} \frac{1}{(r\Delta_a^2 - 1)^2} - 1 = \sum_{a \neq 1} \frac{1}{\left(\frac{r}{\kappa^2}(\kappa\Delta_a)^2 - 1\right)^2} - 1 = \sum_{a \neq 1} \frac{1}{\left(\frac{r}{\kappa^2}\Delta_a'^2 - 1\right)^2} - 1 = \phi_{\boldsymbol{\mu}'}\left(\frac{r}{\kappa^2}\right)$$

and thus  $r' = r/\kappa^2$ .

This implies  $r\Delta_a^2 = r'\Delta_a'^2$  for any  $a \neq 1$ , hence  $\mathbf{w}' = \mathbf{w}$  by Equations (11) and (12) of Proposition 9.  $\square$

*Proof of Lemma 13.* Let us rescale the gaps of  $\boldsymbol{\mu}'$  to obtain the same maximal gap, by multiplying by constant  $\kappa = \frac{\Delta_{\max}}{\Delta_{\max} + d}$ . Denoting by  $\boldsymbol{\mu}''$  the obtained bandit, with  $\boldsymbol{\Delta}'' = \boldsymbol{\Delta}(\boldsymbol{\mu}'') = \kappa\boldsymbol{\Delta}'$  and  $\mathbf{w}'' = \mathbf{w}(\boldsymbol{\mu}'')$ , we have  $\mathbf{w}'' = \mathbf{w}'$  by Lemma 23. Let  $a$  be (one of) the worst arm of  $\boldsymbol{\mu}$ , such that  $\Delta_a = \Delta_{\max}$ . Then

$$\Delta''_{\max} = \Delta''_a = \kappa\Delta'_a = \frac{\Delta_a}{\Delta_a + d}(\Delta_a + d) = \Delta_a = \Delta_{\max}$$

and for any  $b \neq 1$ , one has  $\Delta_b \leq \Delta_a$  so that the nondecreasing of  $x \mapsto \frac{x}{x+d}$  leads to:

$$\Delta''_b = \kappa\Delta'_b = \frac{\Delta_a}{\Delta_a + d}(\Delta_b + d) \geq \frac{\Delta_b}{\Delta_b + d}(\Delta_b + d) = \Delta_b$$

Now we can use Lemma 12 to every arm  $b \notin \{1, a\}$  to go from  $\boldsymbol{\mu}$  to  $\boldsymbol{\mu}''$ , and by Point 2 we know that those transformations can only increase  $w_a$ , so that by Corollary 10

$$w'_{\min} = w'_a = w''_a \geq w_a = w_{\min}.$$

If in addition there exists an arm  $b$  for which  $\Delta_b < \Delta_a$ , then strict inequality  $\Delta_b < \Delta''_b$  occurs in the above inequality and hence Lemma 12 gives a strict increasing of  $w_{\min}$ .  $\square$

*Proof of Lemma 14.* Using scaling argument from Lemma 23, like in the proof of Lemma 13, we can scale  $\boldsymbol{\mu}'$  to keep gap between arm 1 and arms of  $B$  unchanged. That would increase the gaps of all the other arms which in consequence, using Point 2 of Lemma 12, would mean that corresponding  $w_{\min}$  increases.  $\square$

Finally we can prove that Algorithm 1 correctly computes the optimistic bandit.

*Proof of Proposition 1.* We stick to the notation of Algorithm 1, and first observe that  $\mathbf{w} = \mathbf{w}(\tilde{\boldsymbol{\mu}})$ . When  $\min\text{UB} \geq \max\text{LB}$  the algorithm returns a constant bandit and  $\mathbf{w} = (1/K, \dots, 1/K)$  which is its optimal weight vector by convention. As all weight vectors belong to  $\Sigma_K$ , the result is clear.

Now assume that  $\min\text{UB} < \max\text{LB}$  and fix  $\boldsymbol{\nu} \in \mathcal{CR}$ . If  $\boldsymbol{\nu}$  as several optimal arms, then  $w_{\min}(\boldsymbol{\nu}) = 0$  so that  $w_{\min}(\boldsymbol{\nu}) \leq w_{\min}(\tilde{\boldsymbol{\mu}})$ . Assume now that  $\boldsymbol{\nu}$  has a unique optimal arm denoted by  $a$ . Note that  $a \in \text{PotentialBestArms}$ , so that we will show that  $w_{\min}(\boldsymbol{\nu}) \leq w_{\min}(\tilde{\boldsymbol{\mu}}^{\text{test}(a)})$  by transforming  $\boldsymbol{\nu}$  to  $\tilde{\boldsymbol{\mu}}^{\text{test}(a)}$  with changes that will only increase the quantity of interest  $w_{\min}$ . Remark that the value of  $w_{\min}$  is the vector value associated to any of the worst arms of a bandit due to Corollary 10. The procedure, illustrated in Figure 5, is the following:

1. Transform  $\boldsymbol{\nu}$  into  $\boldsymbol{\nu}^{(1)}$  by increasing arm  $a$  so that  $\nu_a^{(1)} = \bar{\mu}_a$ . Using Lemma 13, one has  $w_{\min}(\boldsymbol{\nu}^{(1)}) \geq w_{\min}(\boldsymbol{\nu})$ .

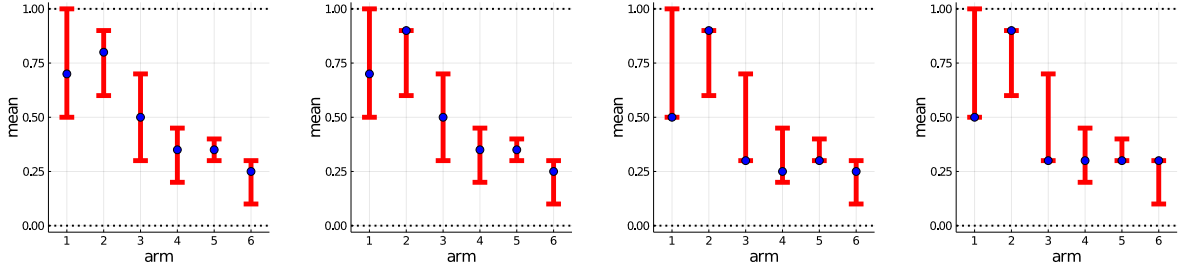


Figure 5: Transformations in the proof of Proposition 1, for some instance bandit  $\nu$ . From left to right:  $\nu$ ,  $\nu^{(1)}$ ,  $\nu^{(2)}$ ,  $\nu^{(3)} = \tilde{\mu}^{\text{test}(2)}$

2. Transform  $\nu^{(1)}$  into  $\nu^{(2)}$  by decreasing, for each arm  $b \neq a$ ,  $\mu_b$  to  $\max(\underline{\mu}_b, \nu_{\min})$ . By several applications of Lemma 12, one has  $w_{\min}(\nu^{(2)}) \geq w_{\min}(\nu^{(1)})$  (remark that imposing to stay above  $\nu_{\min}$  ensures that the associated worst arm stays one of the worst arms at each modification).
3. Transform  $\nu^{(2)}$  into  $\nu^{(3)}$  by increasing all the worst arms to minUB. By Lemma 14, one has  $w_{\min}(\nu^{(3)}) \geq w_{\min}(\nu^{(2)})$ .

We now have  $\nu^{(3)} = \tilde{\mu}^{\text{test}(a)}$  so that  $w_{\min}(\nu) \leq w_{\min}(\tilde{\mu}^{\text{test}(a)})$ . We thus showed that

$$\max_{\nu \in \mathcal{C}\mathcal{R}} w_{\min}(\nu) = \max_{a \in \text{PotentialBestArms}} w_{\min}(\tilde{\mu}^{\text{test}(a)}) = w(\tilde{\mu}),$$

where the last inequality comes from the procedure defining  $\tilde{\mu}$ . □

## B.5 Proof of Theorem 15

We have that

$$\phi_{\mu'}\left(\frac{r}{1+\varepsilon}\right) = \sum_{a \neq 1} \frac{1}{\left(\frac{r}{1+\varepsilon} \Delta'_a{}^2 - 1\right)^2} - 1 \geq \sum_{a \neq 1} \frac{1}{\left(\frac{r}{1+\varepsilon} \Delta_a^2 (1+\varepsilon) - 1\right)^2} - 1 = \phi_{\mu}(r) = 0$$

and

$$\phi_{\mu'}\left(\frac{r}{1-\varepsilon}\right) = \sum_{a \neq 1} \frac{1}{\left(\frac{r}{1-\varepsilon} \Delta'_a{}^2 - 1\right)^2} - 1 \leq \sum_{a \neq 1} \frac{1}{\left(\frac{r}{1-\varepsilon} \Delta_a^2 (1-\varepsilon) - 1\right)^2} - 1 = \phi_{\mu}(r) = 0$$

hence by monotonicity of  $\phi_{\mu'}$  and definition of  $r'$ :

$$\frac{r}{1+\varepsilon} \leq r' \leq \frac{r}{1-\varepsilon}.$$

Consequently, for every  $a \neq 1$ ,  $r' \Delta'_a{}^2 \leq (1+\eta)r \Delta_a^2$  for  $1+\eta = (1+\varepsilon)/(1-\varepsilon)$ , and

$$\frac{1}{r' \Delta'_a{}^2 - 1} \geq \frac{1}{(r \Delta_a^2 - 1) \left(1 + \frac{\eta r \Delta_a^2}{r \Delta_a^2 - 1}\right)} \geq \frac{1}{r \Delta_a^2 - 1} \left(1 - \frac{\eta r \Delta_a^2}{r \Delta_a^2 - 1}\right) = \frac{1}{r \Delta_a^2 - 1} - \eta \frac{1}{r \Delta_a^2 - 1} - \eta \frac{1}{(r \Delta_a^2 - 1)^2}$$

so that

$$\begin{aligned}
 (w'_1)^{-1} &= 1 + \sum_{a \neq 1} \frac{1}{r' \Delta'_a{}^2 - 1} \\
 &\geq 1 + (1 - \eta) \sum_{a \neq 1} \frac{1}{r \Delta_a^2 - 1} - \eta \underbrace{\sum_{a \neq 1} \frac{1}{(r \Delta_a^2 - 1)^2}}_{=1} \\
 &= (1 - \eta) w_1^{-1} = \frac{1 - 3\varepsilon}{1 - \varepsilon} w_1^{-1} \geq (1 - 3\varepsilon) w_1^{-1}.
 \end{aligned}$$

Furthermore,  $r \Delta_a^2 \geq 2$  (see the lower bound in Inequalities (15) of Proposition 11), hence  $\frac{r \Delta_a^2}{r \Delta_a^2 - 1} \leq 2$  by decreasing of  $x \mapsto \frac{x}{x-1}$  on  $(2, +\infty)$ . Thus, for every  $\eta \leq 1/4$ ,  $u = \eta \frac{r \Delta_a^2}{r \Delta_a^2 - 1} \leq 1/2$  and  $\frac{1}{1-u} \leq 1 + 2u$ . One has  $r' \Delta'_a{}^2 \geq (1 - \eta) r \Delta_a^2$  for  $1 - \eta = (1 - \varepsilon)/(1 + \varepsilon)$ , and one checks that  $\eta \leq 1/4$  for  $\varepsilon \leq 1/7$ , hence

$$\frac{1}{r' \Delta'_a{}^2 - 1} \leq \frac{1}{(r \Delta_a^2 - 1) \left(1 - \frac{\eta r \Delta_a^2}{r \Delta_a^2 - 1}\right)} \leq \frac{1}{r \Delta_a^2 - 1} \left(1 + 2 \frac{\eta r \Delta_a^2}{r \Delta_a^2 - 1}\right) = \frac{1}{r \Delta_a^2 - 1} + 2\eta \frac{1}{r \Delta_a^2 - 1} + 2\eta \frac{1}{(r \Delta_a^2 - 1)^2}$$

Consequently,

$$\begin{aligned}
 (w'_1)^{-1} &= 1 + \sum_{a \neq 1} \frac{1}{r' \Delta'_a{}^2 - 1} \\
 &\leq 1 + (1 + 2\eta) \sum_{a \neq 1} \frac{1}{r \Delta_a^2 - 1} + 2\eta \underbrace{\sum_{a \neq 1} \frac{1}{(r \Delta_a^2 - 1)^2}}_{=1} \\
 &= (1 + 2\eta) w_1^{-1} = \frac{1 + 5\varepsilon}{1 + \varepsilon} w_1^{-1} \leq (1 + 5\varepsilon) w_1^{-1}.
 \end{aligned}$$

To summarize, for  $\varepsilon \leq 1/7$ , by Equation (13) of Proposition 9, on the one hand:

$$T' = 2r' w'_1{}^{-1} \geq 2 \times \frac{r}{1 + \varepsilon} \times \frac{1 - 3\varepsilon}{1 - \varepsilon} w_1^{-1} = \frac{1 - 3\varepsilon}{1 + \varepsilon^2} \times T \geq (1 - 3\varepsilon) T$$

and on the other hand

$$T' = 2r' (w'_1)^{-1} \leq 2 \times \frac{r}{1 - \varepsilon} \times \frac{1 + 5\varepsilon}{1 + \varepsilon} w_1^{-1} = \frac{1 + 5\varepsilon}{1 - \varepsilon^2} \times T \leq (1 + 6\varepsilon) T$$

as  $1 + 5\varepsilon \leq (1 + 6\varepsilon)(1 - \varepsilon^2)$ .

We also have

$$(1 - 5\varepsilon) w_1 \leq \frac{w_1}{1 + 5\varepsilon} \leq w'_1 \leq \frac{w_1}{1 - 3\varepsilon} \leq (1 + 6\varepsilon) w_1$$

which yields by Equation (12) of Proposition 9, for any  $a \neq 1$ :

$$\begin{aligned}
 (1 - 10\varepsilon) w_a &\leq \frac{w_1 / (1 + 5\varepsilon)}{(r \Delta_a^2 - 1) \left(1 + \frac{2\varepsilon}{1 + \varepsilon}\right)} \leq w'_a = \frac{w'_1}{r' \Delta'_a{}^2 - 1} \\
 &\leq \frac{w_1 / (1 - 3\varepsilon)}{(r \Delta_a^2 - 1) \left(1 - 2 \frac{2\varepsilon}{1 + \varepsilon}\right)} = \frac{1 + \varepsilon}{(1 - 3\varepsilon)^2} w_a \leq (1 + 10\varepsilon) w_a.
 \end{aligned}$$

## B.6 Proof of Proposition 16

We will prove Proposition 16 by combining two Lemmas. Note that in this subsection  $\boldsymbol{\mu}$  and  $\boldsymbol{\mu}'$  are general bandits, with possibly more than one best arm.

**Lemma 24.** *Let  $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{G}$  and  $\mathbf{v} \in \Sigma_K$  be any optimal vector. Then:*

$$g(\boldsymbol{\mu}', \mathbf{v}) \geq g(\boldsymbol{\mu}, \mathbf{v}) - \varepsilon/2$$

where  $\varepsilon = \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_\infty$ .

*Proof.*

- Assume first that  $\boldsymbol{\mu}$  and  $\boldsymbol{\mu}'$  have a common best arm. Without loss of generality we assume that this arm is 1. Then:

$$\begin{aligned} g(\boldsymbol{\mu}', \mathbf{v}) - g(\boldsymbol{\mu}, \mathbf{v}) &= \frac{1}{2} \min_{a \neq 1} \frac{v_1 v_a}{v_1 + v_a} \Delta_a'^2 - \frac{1}{2} \min_{b \neq 1} \frac{v_1 v_b}{v_1 + v_b} \Delta_b^2 && \text{by Equation (7)} \\ &= \frac{1}{2} \min_{a \neq 1} \max_{b \neq 1} \frac{v_1 v_a}{v_1 + v_a} \Delta_a'^2 - \frac{v_1 v_b}{v_1 + v_b} \Delta_b^2 \\ &\geq \frac{1}{2} \min_{a \neq 1} \frac{v_1 v_a}{v_1 + v_a} (\Delta_a'^2 - \Delta_a^2) && \text{taking } b = a. \end{aligned}$$

Then for any  $a \neq 1$ , one has:

$$|\Delta_a - \Delta_a'| = |(\mu_1 - \mu'_1) - (\mu_a - \mu'_a)| \leq |\mu_1 - \mu'_1| + |\mu_a - \mu'_a| \leq 2\varepsilon$$

from which we obtain, using that the gaps are in  $[0, 1]$  in  $\mathcal{G}$

$$\left| \Delta_a^2 - \Delta_a'^2 \right| = |\Delta_a - \Delta_a'| (\Delta_a + \Delta_a') \leq 4\varepsilon.$$

As  $\mathbf{v}$  is an optimal vector, we have  $0 \leq v_a \leq v_1 \leq \frac{1}{2}$  using Equation (16), so that:

$$\frac{v_1 v_a}{v_1 + v_a} \leq \frac{1}{2} \frac{v_a}{v_1 + v_a} \leq \frac{1}{2} \frac{v_a}{2v_a} = \frac{1}{4}$$

hence

$$\frac{v_1 v_a}{v_1 + v_a} (\Delta_a'^2 - \Delta_a^2) \geq -\varepsilon.$$

- In case  $\boldsymbol{\mu}$  and  $\boldsymbol{\mu}'$  do not share a best arm, define the family of bandits  $(\boldsymbol{\mu}^{(t)})_{t \in [0,1]}$  by

$$\forall t \in [0, 1], \forall a \in [K], \quad \mu_a^{(t)} = (1-t)\mu_a + t\mu'_a.$$

One can check that

- $\boldsymbol{\mu} = \boldsymbol{\mu}^{(0)}$ ,
- $\boldsymbol{\mu}' = \boldsymbol{\mu}^{(1)}$ ,
- $\|\boldsymbol{\mu}^{(t_1)} - \boldsymbol{\mu}^{(t_2)}\|_\infty \leq |t_1 - t_2| \varepsilon$  for every  $t_1, t_2 \in [0, 1]$ .

Select the subdivision  $0 = t_0 < t_1 < \dots < t_N = 1$  of times at which the optimal arms of  $\boldsymbol{\mu}^{(t)}$  are modified. Note that  $N \geq 2$  as  $\boldsymbol{\mu}$  and  $\boldsymbol{\mu}'$  do not have a common best arm. Note that by continuity:

- for any  $n \in \llbracket 1, N-1 \rrbracket$ ,  $\boldsymbol{\mu}^{(t_n)}$  has at least two best arms so that  $g(\boldsymbol{\mu}^{(t_n)}, \mathbf{v}) = 0$ ,
- $\boldsymbol{\mu}^{(1)}$  and  $\boldsymbol{\mu}$  have a common best arm,

–  $\boldsymbol{\mu}^{(N-1)}$  and  $\boldsymbol{\mu}'$  have a common best arm.

Thus

$$\begin{aligned} g(\boldsymbol{\mu}', \mathbf{v}) - g(\boldsymbol{\mu}, \mathbf{v}) &= g(\boldsymbol{\mu}', \mathbf{v}) - g(\boldsymbol{\mu}^{(1)}, \mathbf{v}) + g(\boldsymbol{\mu}^{(N-1)}, \mathbf{v}) - g(\boldsymbol{\mu}, \mathbf{v}) \\ &\geq -\frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}^{(1)}\|_\infty + \|\boldsymbol{\mu}^{(N-1)} - \boldsymbol{\mu}'\|_\infty}{2} \\ &\geq -\frac{(t_1 + (1 - t_{N-1}))\varepsilon}{2} \geq -\frac{\varepsilon}{2}. \end{aligned}$$

□

**Lemma 25.** *Let  $\boldsymbol{\mu}' \in \mathcal{G}$  be a Gaussian bandit and  $\mathbf{u}, \mathbf{v} \in \Sigma_K$  be such that*

$$\max_{a \in [K]} \frac{|u_a - v_a|}{u_a} \leq \eta$$

for a fixed  $0 \leq \eta \leq 1$ . Then:

$$g(\boldsymbol{\mu}', \mathbf{v}) \geq \frac{(1 - \eta)^2}{1 + \eta} g(\boldsymbol{\mu}', \mathbf{u}).$$

*Proof.* Without loss of generality, assume that arm 1 is one of the best arms of  $\boldsymbol{\mu}'$ . Note that the condition of the lemma can be rewritten as

$$\forall a \in [K], \quad (1 - \eta)u_a \leq v_a \leq (1 + \eta)u_a.$$

Then for every  $a \neq 1$ :

$$\frac{v_1 v_a}{v_1 + v_a} \geq \frac{(1 - \eta)^2 u_1 u_a}{(1 + \eta)(u_1 + u_a)}.$$

Thus:

$$g(\boldsymbol{\mu}', \mathbf{v}) = \min_{a \neq 1} \frac{v_1 v_a}{v_1 + v_a} \Delta_a'^2 \geq \frac{(1 - \eta)^2}{1 + \eta} \min_{a \neq 1} \frac{u_1 u_a}{u_1 + u_a} \Delta_a'^2 = \frac{(1 - \eta)^2}{1 + \eta} g(\boldsymbol{\mu}', \mathbf{u}).$$

□

*Proof of Proposition 16.* The result follows directly by applying Lemmas 25 and 24 with  $\mathbf{u} = \mathbf{w}(\boldsymbol{\mu})$ :

$$g(\boldsymbol{\mu}', \mathbf{v}) \geq \frac{(1 - \eta)^2}{1 + \eta} g(\boldsymbol{\mu}', \mathbf{w}(\boldsymbol{\mu})) \geq \frac{(1 - \eta)^2}{1 + \eta} (g(\boldsymbol{\mu}, \mathbf{w}(\boldsymbol{\mu})) - \varepsilon/2).$$

□

## C Technical details for the proof of Section 4

### C.1 Proof of Lemma 17

We will use the following deterministic Lemma:

**Lemma 26.** *One has:*

$$\forall t > 0, \quad \max_{1 \leq a \leq K} \left| N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s) \right| \leq K - 1.$$

*Proof.* Apply [Garivier and Kaufmann, 2016, Lemma 15] with  $p(s) = \tilde{w}(s)$ . □

The claim is true for  $t \in \llbracket 0, K \rrbracket$  as Equation (16) of Proposition 11 gives

$$w_{\min}K - K \leq \frac{K}{2} - K \leq 0.$$

Otherwise, fix  $t \in \llbracket K + 1, \tau_\delta \rrbracket$  and  $a \in [K]$ . For any  $s \in \llbracket 0, K - 1 \rrbracket$ , one has  $\tilde{w}_a(s) = \frac{1}{K}$  by convention (as all arms are drawn once during the  $K$  first rounds, the only request is  $\sum_{s=0}^{K-1} \tilde{w}_a(s) = 1$ ), and thus  $\tilde{w}_a(s) \geq w_{\min}$  ( $\mathbf{w} \in \Sigma_K$  implies  $w_{\min} \leq \frac{1}{K}$ ). For any  $s \in \llbracket K, \tau_\delta - 1 \rrbracket$ , one has by Proposition 1 :

$$\tilde{w}_a(s) \geq \tilde{w}_{\min}(s) = \max_{\boldsymbol{\nu} \in \mathcal{CR}_\mu(s)} w_{\min}(\boldsymbol{\nu}) \geq w_{\min}$$

as  $\boldsymbol{\mu} \in \mathcal{CR}_\mu(s)$  on event  $\mathcal{E}$ . Hence by Lemma 26

$$N_a(t) \geq \sum_{s=0}^{t-1} \tilde{w}_a(s) - (K - 1) \geq tw_{\min} - (K - 1) \geq tw_{\min} - K.$$

## C.2 A technical lemma

**Lemma 27.** *For any  $c_1, c_2 > 0$ ,*

$$x = \frac{2}{c_1} \log \left( \frac{c_2 e}{c_1} \right)$$

*is such that  $c_1 x \geq \log(c_2 x)$ .*

This is a direct consequence of [Garivier and Kaufmann, 2016, Lemma 18].

## C.3 Deviation bound

We prove the following simple consequence of Hoeffding's inequality.

**Lemma 28.** *For any  $t \geq \frac{2K}{w_{\min}}$  and  $x > 0$ , one has*

$$\mathbb{P} \left( \max_{a \in [K]} |\hat{\mu}_a(t) - \mu_a| > x \cap \mathcal{E} \right) \leq 2Kt \exp \left( - \frac{tw_{\min}}{4} x^2 \right).$$

*Proof.* Fix  $t \geq \frac{2K}{w_{\min}}$  and  $x > 0$ . For any  $a \in [K]$ , one has with  $T = \frac{tw_{\min}}{2}$ :

$$\begin{aligned} \mathbb{P} \left( |\hat{\mu}_a(t) - \mu_a| > x \cap \mathcal{E} \right) &= \sum_{s=T}^t \mathbb{P} \left( |\hat{\mu}_a(t) - \mu_a| > x \cap \mathcal{E} \cap N_a(t) = s \right) && \text{by Equation (18)} \\ &\leq \sum_{s=T}^t \mathbb{P} \left( |\hat{\mu}_{a,s} - \mu_a| > x \right) && \text{by Equation (26)} \\ &\leq \sum_{s=T}^t 2 \exp \left( - \frac{s}{2} x^2 \right) && \text{by Hoeffding's inequality} \\ &\leq 2t \exp \left( - \frac{T}{2} x^2 \right) \end{aligned}$$

giving the desired bound by union bound.  $\square$



## C.4 Proof of Lemma 22

We have

$$S(C) = \sum_{t > (1+C)\zeta T \log(1/\delta)} t \exp\left(-\frac{tw_{\min} C^2}{4T^2}\right) = \sum_{t > B} f(t)$$

where  $f : t \mapsto t \exp(-At)$ ,  $A = \frac{w_{\min}}{4T^2} C^2$  and  $B = (1+C)\zeta T \log(1/\delta)$ .  $f$  is increasing until  $1/A$  and then decreasing. Let  $n_0 = \lfloor \frac{1}{A} \rfloor$ .

- If  $B > n_0$  then  $f$  is decreasing on  $[B, +\infty[$  and one has:

$$S(C) \leq \int_B^{+\infty} f(t) dt.$$

- Otherwise, one has:

$$\begin{aligned} S(C) &= \sum_{t=\lceil B \rceil}^{n_0-1} f(t) + f(n_0) + f(n_0+1) + \sum_{t > n_0+1} f(t) \\ &\leq \sum_{t=\lceil B \rceil}^{n_0-1} \int_t^{t+1} f(t) dt + f(n_0) + f(n_0+1) + \sum_{t > n_0+1} \int_{t-1}^t f(t) dt \\ &\leq \int_{\lceil B \rceil}^{+\infty} f(t) dt + f(n_0) + f(n_0+1) \end{aligned}$$

where in the second inequality, we use the increasing of  $f$  on  $[B, n_0]$  and its decreasing on  $[n_0+1, +\infty]$ . The result will be true if

$$f(n_0) + f(n_0+1) \leq \int_B^{+\infty} f(t) dt.$$

One has:

$$\begin{aligned} f(n_0) + f(n_0+1) &= \left\lfloor \frac{1}{A} \right\rfloor e^{-A \lfloor \frac{1}{A} \rfloor} + \left\lceil \frac{1}{A} \right\rceil e^{-A \lceil \frac{1}{A} \rceil} \\ &\leq \left( \left\lfloor \frac{1}{A} \right\rfloor + \left\lceil \frac{1}{A} \right\rceil \right) e^{-A \lfloor \frac{1}{A} \rfloor} \\ &\leq \left( \left\lfloor \frac{1}{A} \right\rfloor \frac{1}{A} + \frac{1}{A^2} \right) e^{-A \lfloor \frac{1}{A} \rfloor} \quad \text{as } A < \frac{1}{2} \\ &= \int_{\lfloor \frac{1}{A} \rfloor}^{+\infty} f(t) dt \leq \int_B^{+\infty} f(t) dt \quad \text{as } B \leq \left\lfloor \frac{1}{A} \right\rfloor = n_0. \end{aligned}$$

where in the last inequality, we used the simple calculation

$$\int_Y^{+\infty} t \exp(-tX) dt = \exp(-YX) \left( \frac{Y}{X} + \frac{1}{X^2} \right)$$

for  $X, Y > 0$ .

In both cases we have:

$$S(C) \leq 2 \int_{(1+C)\zeta T \log(1/\delta)}^{\infty} t \exp\left(-\frac{tw_{\min} C^2}{4T^2}\right) dt$$

and using the same calculation as before

$$S(C) \leq 2 \exp\left(-\frac{\zeta w_{\min}}{4T} (1+C) C^2 \log(1/\delta)\right) \left( \frac{4(1+C)\zeta T^3 \log(1/\delta)}{w_{\min} C^2} + \frac{16T^4}{w_{\min}^2 C^4} \right).$$

Bounding  $C \in (0, 1]$  and  $\zeta \in [1, 2]$  (remind that  $\zeta \leq 1 + 6\eta'$ ):

$$\begin{aligned} S(C) &\leq 2 \exp\left(-\frac{w_{\min}}{4T} C^2 \log(1/\delta)\right) \left(\frac{16T^3 \log(1/\delta)}{w_{\min} C^2} + \frac{16T^4}{w_{\min}^2 C^4}\right) \\ &\leq \frac{32T^4}{w_{\min}^2} \exp\left(-\frac{w_{\min}}{4T} C^2 \log(1/\delta)\right) \left(\frac{\log(1/\delta)}{C^2} + \frac{1}{C^4}\right). \end{aligned}$$

## D Proof of asymptotic results

### D.1 Proof of Lemma 3

We will need the following Lemma concerning the optimistic bandit computed by Algorithm 1.

**Lemma 29.** *Let  $\mathcal{CR} = \prod_{a \in [K]} [\underline{\mu}_a, \bar{\mu}_a]$  be a confidence region such that  $\underline{\mu}_a < \bar{\mu}_a$  for  $a \in [K]$  and  $\max_{a \in [K]} \underline{\mu}_a = \max LB > \min UB = \min_{a \in [K]} \bar{\mu}_a$ , and  $(\tilde{\boldsymbol{\mu}}, \boldsymbol{\nu}) \leftarrow \text{OPTIMISTICWEIGHTS}(\mathcal{CR})$ . Then*

$$\Delta_{\min}(\tilde{\boldsymbol{\mu}}) \geq \min_{a \in [K]} \bar{\mu}_a - \underline{\mu}_a$$

*Proof.* We proceed by contradiction: let us assume that  $\tilde{\boldsymbol{\mu}}$  is such that

$$\Delta_{\min}(\tilde{\boldsymbol{\mu}}) < \min_{a \in [K]} \bar{\mu}_a - \underline{\mu}_a$$

By the two hypothesis and the Algorithm's procedure, it is clear that  $\tilde{\boldsymbol{\mu}}$  has a unique best arm. Without loss of generality let us arrange the arms so that  $\tilde{\mu}_1 > \tilde{\mu}_2 \geq \tilde{\mu}_3 \geq \dots \geq \tilde{\mu}_K$ . Note that  $\Delta_{\min}(\tilde{\boldsymbol{\mu}}) = \tilde{\mu}_1 - \tilde{\mu}_2$ .

As 1 is the best arm, once again the Algorithm's procedure ensures that  $\tilde{\mu}_1 = \bar{\mu}_1$ . In addition, our assumption implies  $\Delta_{\min}(\tilde{\boldsymbol{\mu}}) < \bar{\mu}_1 - \underline{\mu}_1$ , giving  $\tilde{\mu}_2 > \underline{\mu}_1$ . Recall that  $\tilde{\mu}_2 = \max(\underline{\mu}_2, \min UB)$ , so that we split our analysis to the two possible cases:

- if  $\tilde{\mu}_2 = \underline{\mu}_2$ , then we cannot have  $\bar{\mu}_2 \leq \bar{\mu}_1 = \tilde{\mu}_1$  otherwise  $\Delta_{\min}(\tilde{\boldsymbol{\mu}}) > \bar{\mu}_2 - \underline{\mu}_2$ , which is impossible. Then  $\bar{\mu}_2 > \bar{\mu}_1$ . By defining  $\boldsymbol{\nu} = (\tilde{\mu}_2, \bar{\mu}_2, \tilde{\mu}_3, \dots, \tilde{\mu}_K)$ , one has  $\boldsymbol{\nu} \in \mathcal{CR}$  and  $w_{\min}(\boldsymbol{\nu}) > w_{\min}(\tilde{\boldsymbol{\mu}})$  by Lemma 13. Thus  $\tilde{\boldsymbol{\mu}}$  cannot maximize  $w_{\min}$  over  $\mathcal{CR}$  which is in contradiction with Proposition 1.
- if  $\tilde{\mu}_2 = \min UB$ , then  $\tilde{\mu}_2 = \tilde{\mu}_3 = \dots = \tilde{\mu}_K$  and thus all confidence intervals share a common point equal to  $\tilde{\mu}_2$  (recall that  $\tilde{\mu}_2 \in [\underline{\mu}_1, \bar{\mu}_1]$ ), which is a contradiction with  $\max LB > \min UB$ .

□

We can now prove Lemma 3. Let  $t \in \llbracket 0, \tau_\delta - 1 \rrbracket$ . We want to lower bound  $\tilde{w}_{\min}(t)$ .

- If at time  $t$  one has  $\tilde{\boldsymbol{w}}(t) = (1/K, \dots, 1/K)$  then  $\tilde{w}_{\min}(t) = \frac{1}{K}$ .
- Otherwise, by construction of Algorithms 1 and 2 we know that  $t \geq K$  and the confidence region  $\mathcal{CR}(t)$  is such that at least two confidence intervals are separated. In that case, the optimistic bandit

$\tilde{\boldsymbol{\mu}}(t)$  has a unique optimal arm, so that we have:

$$\begin{aligned}
 \tilde{w}_{\min}(t) &= \frac{\tilde{w}_{\max}(t)}{r\tilde{\Delta}_{\max}(t) - 1} && \text{by Equation (12) of Proposition 9} \\
 &\geq \frac{1}{\sqrt{K-1} + 1} \times \frac{1}{\frac{\sqrt{K-1}+1}{\tilde{\Delta}_{\min}(t)}\tilde{\Delta}_{\max}(t) - 1} && \text{by Inequalities (15) and (16)} \\
 &\geq \frac{\tilde{\Delta}_{\min}(t)}{(\sqrt{K-1} + 1)^2} && \text{as } \tilde{\Delta}_{\max}(t) \leq 1 \\
 &\geq \frac{\tilde{\Delta}_{\min}(t)}{2K}.
 \end{aligned}$$

One can use Lemma 29 and note that as  $t \geq K$ , all arms have already been pulled at least once, hence

$$\tilde{\Delta}_{\min}^{(t)} \geq \min_{a \in [K]} 2\ell_a(t) \geq 4 \min_{a \in [K]} \sqrt{\frac{\log(4N_a(t)K/\gamma)}{N_a(t)}} \geq 4\sqrt{\frac{\log(4K/\gamma)}{t}} \geq 4\sqrt{\frac{\log 8}{t}} \geq \frac{4}{\sqrt{t}}.$$

Putting everything together one can obtain

$$\tilde{w}_{\min}(t) \geq \frac{2}{K} \frac{1}{\sqrt{t}}.$$

In both cases we obtained:

$$\tilde{w}_{\min}(t) \geq \min\left(\frac{2}{K} \frac{1}{\sqrt{t}}, \frac{1}{K}\right) \geq \frac{1}{K} \frac{1}{\sqrt{t}}$$

hence for any  $a \in [K]$  and  $t \in \mathbb{N}$ , we have using Lemma 26:

$$N_a(t) \geq \sum_{s=0}^{t-1} \tilde{w}_a(s) - (K-1) \geq \sum_{s=2}^{t-1} \tilde{w}_{\min}(s) - K \geq \frac{1}{K} \sum_{s=2}^{t-1} \frac{1}{\sqrt{s}} - K \geq \frac{1}{K} \int_1^t \frac{1}{\sqrt{s}} ds - K \geq \frac{2}{K} \sqrt{t} - K.$$

## D.2 Proof of Theorem 6

The result was obtained by [Garivier and Kaufmann, 2016, Proposition 13]. The adaptation to EXPLORATION-BIASED SAMPLING is straightforward, as soon as we prove the following result

**Proposition 30.** *For any choice of parameters and  $\boldsymbol{\mu} \in \mathcal{G}^*$ , the sampling rule of EXPLORATION-BIASED SAMPLING satisfies:*

$$\lim_{t \rightarrow +\infty} \hat{\boldsymbol{\mu}}(t) = \boldsymbol{\mu} \quad \mathbb{P}_{\boldsymbol{\mu}}\text{-a.s.} \quad \text{and} \quad \lim_{t \rightarrow +\infty} \frac{N(t)}{t} = w(\boldsymbol{\mu}) \quad \mathbb{P}_{\boldsymbol{\mu}}\text{-a.s.} .$$

*Proof.* Lemma 3 implies that  $N_a(t) \rightarrow_{t \rightarrow +\infty} +\infty$  for all  $a \in [K]$ , so that the law of large number gives

$$\lim_{t \rightarrow +\infty} \hat{\boldsymbol{\mu}}(t) = \boldsymbol{\mu} \quad \mathbb{P}_{\boldsymbol{\mu}}\text{-a.s.} .$$

Remark that for  $a \in [K]$  one has

$$|\tilde{\boldsymbol{\mu}}_a(t) - \hat{\boldsymbol{\mu}}_a(t)| \leq C_{\gamma/K}(N_a(t)) = 2\sqrt{\frac{\log(4N_a(t)K/\gamma)}{N_a(t)}} \xrightarrow{t \rightarrow +\infty} 0$$

so that we also have

$$\lim_{t \rightarrow +\infty} \tilde{\boldsymbol{\mu}}(t) = \boldsymbol{\mu} \quad \mathbb{P}_{\boldsymbol{\mu}}\text{-a.s.}$$

and thus by continuity of function  $\mathbf{w}$  in  $\boldsymbol{\mu}$  (as  $\boldsymbol{\mu}$  has a unique optimal arm):

$$\lim_{t \rightarrow +\infty} \tilde{\mathbf{w}}(t) = \mathbf{w}(\boldsymbol{\mu}) \quad \mathbb{P}_{\boldsymbol{\mu}\text{-a.s.}}$$

Now for all  $t \in \mathbb{N}^*$  and  $a \in [K]$  we have:

$$\begin{aligned} \left| \frac{N_a(t)}{t} - w_a(\boldsymbol{\mu}) \right| &\leq \frac{1}{t} \left| N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s) \right| + \left| \frac{1}{t} \sum_{s=0}^{t-1} (\tilde{w}_a(s) - w_a(\boldsymbol{\mu})) \right| \\ &\leq \frac{K-1}{t} + \left| \frac{1}{t} \sum_{s=0}^{t-1} (\tilde{w}_a(s) - w_a(\boldsymbol{\mu})) \right| && \text{by Lemma 26} \\ &\xrightarrow{t \rightarrow +\infty} 0 \end{aligned}$$

(using the Cesaro Lemma for the second term).  $\square$

### D.3 Proof of Theorem 7

Once again this is a direct adaptation of [Garivier and Kaufmann, 2016, Theorem 14]. Indeed, we can follow the proof as long as the two Lemmas are satisfied.

Let us recall the notations following notations. We assume that 1 is the best arm of  $\boldsymbol{\mu}$ . Fix  $\varepsilon > 0$ . By continuity of  $\mathbf{w}$  in  $\boldsymbol{\mu}$ , let  $\xi \leq \Delta_{\min}(\boldsymbol{\mu})/4$  such that

$$\max_{\boldsymbol{\mu}' \in \mathcal{I}_\varepsilon} \|\mathbf{w}(\boldsymbol{\mu}') - \mathbf{w}(\boldsymbol{\mu})\|_\infty \leq \varepsilon \quad \text{where} \quad \mathcal{I}_\varepsilon = \prod_{a \in [K]} [\mu_a \pm \xi].$$

Let  $T \in \mathbb{N}$  and define  $h(T) = T^{1/4}$  and the event

$$\mathcal{E}_T = \bigcap_{t=h(T)}^T (\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}_\varepsilon).$$

**Lemma 31.** *There exists two positive constants  $B, C$  (that depend on  $\boldsymbol{\mu}$  and  $\varepsilon$ ) such that*

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_T^c) \leq BT \exp(-CT^{1/8}).$$

*Proof.* We have by union bound

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_T^c) \leq \sum_{t=h(T)}^T \sum_{a \in [K]} \mathbb{P}_{\boldsymbol{\mu}}(|\hat{\mu}_a(t) - \mu_a| > \xi).$$

Then

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu}}(|\hat{\mu}_a(t) - \mu_a| > \xi) &= \sum_{s=\frac{2}{K}\sqrt{t}-K}^t \mathbb{P}_{\boldsymbol{\mu}}(|\hat{\mu}_a(t) - \mu_a| > \xi \cap N_a(t) = s) && \text{by Lemma 3} \\ &\leq \sum_{s=\frac{2}{K}\sqrt{t}-K}^t \mathbb{P}(|\hat{\mu}_{a,s} - \mu_a| > \xi) && \text{by Equation (26)} \\ &\leq 2 \sum_{s=\frac{2}{K}\sqrt{t}-K}^t \exp\left(-s \frac{\xi^2}{2}\right) && \text{by Hoeffding's inequality} \\ &\leq 2 \frac{\exp(-(\frac{2}{K}\sqrt{t}-K)\xi^2/2)}{1 - \exp(-\xi^2/2)}. \end{aligned}$$

With

$$B = 2K \frac{\exp(K\xi^2/2)}{1 - \exp(-\xi^2/2)} \quad \text{and} \quad C = \frac{\xi^2}{K},$$

one has

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_T^c) \leq \sum_{t=h(T)}^T B \exp(-\sqrt{t}C) \leq BT \exp(-\sqrt{h(T)}C) \leq BT \exp(-CT^{1/8}).$$

□

**Lemma 32.** *There exists a constant  $T_\varepsilon$  such that for  $T \geq T_\varepsilon$ , it holds that on  $\mathcal{E}_T$*

$$\forall t \geq \sqrt{T}, \quad \max_{a \in [K]} \left| \frac{N_a(t)}{t} - w_a(\boldsymbol{\mu}) \right| \leq 3\varepsilon.$$

*Proof.* For any  $t \geq \sqrt{T} = h(T)^2$  and  $a \in [K]$  we have:

$$\begin{aligned} \left| \frac{N_a(t)}{t} - w_a(\boldsymbol{\mu}) \right| &\leq \frac{1}{t} \left| N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s) \right| + \left| \frac{1}{t} \sum_{s=0}^{t-1} (\tilde{w}_a(s) - w_a(\boldsymbol{\mu})) \right| \\ &\leq \frac{K-1}{t} + \frac{h(T)}{t} + \left| \frac{1}{t} \sum_{s=h(T)}^{t-1} (\tilde{w}_a(s) - w_a(\boldsymbol{\mu})) \right| && \text{by Lemma 26} \\ &\leq \frac{K-1}{T^{1/2}} + \frac{1}{T^{1/4}} + \varepsilon && \text{by definition of } \mathcal{E}_T \\ &\leq \frac{K}{T^{1/4}} + \varepsilon \leq 3\varepsilon \end{aligned}$$

whenever  $T \geq (K/2\varepsilon)^4 = T_\varepsilon$ .

□