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

Antoine Barrier, Aurélien Garivier, Tomáš Kocák

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

Antoine Barrier[1, 2], Aurélien Garivier[1], and Tomáš Kocák[3]

[1]ENS de Lyon, UMPA UMR 5669, 46 allée d'Italie, 69364 Lyon Cedex 07, France
[2]Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
[3]Institute for Mathematics, University of Potsdam, Germany

March 4, 2022

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: it is to the best of our knowledge the first strategy with non-asymptotic bounds that asymptotically matches the sample complexity. But the main advantage over other algorithms like Track-and-Stop is an improved behavior regarding exploration: Exploration-Biased Sampling is biased towards 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

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 (see e.g. Lattimore and Szepesvári, 2020). 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 $\boldsymbol{\mu} = (\mu_a)_{1\leq a\leq K} \in \mathbb{R}^K$ 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 $A_t \in [K] = \{1, \dots, K\}$ 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 $\mathcal{G}^*$ of bandits $\boldsymbol{\mu} \in \mathcal{G}$ that have exactly one arm $a^*(\boldsymbol{\mu}) \in [K]$ with the highest mean, that is $\mu_* = \mu_{a^*(\boldsymbol{\mu})} > \max_{a\in[K]\setminus\{a^*(\boldsymbol{\mu})\}} \mu_a$, and 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; Kalyanakrishnan1 et al., 2012): by defining $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$ the sigma-field generated by the observations up to time t, 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 τ with respect to $(\mathcal{F}_t)_{t\geq 0}$, and a $\mathcal{F}_\tau$-measurable decision rule $\hat{a}_\tau$. Given a risk parameter $\delta \in (0, 1)$, a strategy is called δ-*correct* if, whatever the parameter $\boldsymbol{\mu} \in \mathcal{G}^*$, it holds that $\mathbb{P}_{\boldsymbol{\mu}}(\tau < +\infty, \hat{a}_\tau \neq a^*(\boldsymbol{\mu})) \leq \delta$. The goal is to find a δ-correct strategy that minimizes the expected number of observations $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ needed to identify $a^*(\boldsymbol{\mu})$.

The sample complexity of δ-correct strategies cannot be arbitrarily good: it has been proved by Garivier and Kaufmann (2016) that they essentially obey the lower bound $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T(\boldsymbol{\mu}) \log(1/\delta)$ for any $\boldsymbol{\mu} \in \mathcal{G}^*$, where the *characteristic time* $T(\boldsymbol{\mu})$ is the solution of the following optimization problem

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

where $\Sigma_K = \{\boldsymbol{v} \in [0,1]^K : v_1 + \cdots + v_K = 1\}$ and $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \mathcal{G}^* : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$ is the set of 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 $\limsup_{\delta \to 0} \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] / \log(1/\delta) = T(\boldsymbol{\mu})$ (see also Russo, 2016).

The information-theoretic analysis of Garivier and Kaufmann (2016) also highlights the nature of the optimal sampling strategy: whatever the value of the risk δ, one should sample the arms with frequencies proportional to $\boldsymbol{v} = \boldsymbol{w}(\boldsymbol{\mu})$, the (unique and well-defined) maximizer in the right-hand side of Equation (1). Indeed, the Track-and-Stop algorithm works as follows: at every time step t, an estimate $\hat{\boldsymbol{\mu}}(t)$ of the mean parameter $\boldsymbol{\mu}$ is computed thanks to the available observations. The optimal frequencies relative to this estimate are computed, and used to determine which action is to be selected next: we pick the action that lays the most behind its estimated optimal frequency, unless one action was severely undersampled (in which case its exploration is forced). A formal description of the strategy is recalled in Appendix A (see Algorithm 3). Some improvements were proposed: for example, 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 (see also Tirinzoni et al., 2020).

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., 2020), 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 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 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.

Contributions The present paper addresses the issues of Track-and-Stop and proposes a new algorithm that solves all of them. We focus on Gaussian bandits with known and equal variances. 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, allowing a significant speed-up for the Track-and-Stop or the Gradient Ascent algorithms in the Gaussian case. Actually, the algorithms discussed here 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).

While the proven optimality of Track-and-Stop is purely asymptotic, a different approach is followed in (Karnin et al., 2013; Jamieson et al., 2014; Chen et al., 2017) for moderate values of δ. The proposed strategies are sub-optimal by a multiplicative constant, but are proved to satisfy explicit non-asymptotic bounds. More recently, Degenne et al. (2019) obtained a general non-asymptotic bound, a remarkable but hardly comparable result in particular settings. In this contribution, we try to make a link between both approaches by introducing a strategy with a non-asymptotic bound that asymptotically matches the sample complexity.

The paper is organized as follows. We present in Section 2 our new strategy with 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. Lastly, we illustrate the performance and behavior of our strategy by numerical experiments in Section 4, and propose concluding remarks in Section 5.

2 The Exploration-Biased Sampling strategy

In this section, we introduce our new strategy called Exploration-Biased Sampling. Instead of Track-and-Stop's greedy choice of actions based on a plug-in estimate of $\boldsymbol{\mu}$, it relies on a specific estimator that is biased toward uniform exploration.

For $\boldsymbol{\mu} \in \mathcal{G}$, let $\boldsymbol{\Delta}(\boldsymbol{\mu}) = (\mu_* - \mu_a)_{a\in[K]} \in [0,1]^K$ be its *gap vector* and $a^*(\boldsymbol{\mu}) = \{a \in [K] : \Delta_a(\boldsymbol{\mu}) = 0\}$ 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 we 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 $\boldsymbol{w}(\boldsymbol{\mu}) = \frac{1}{\mathrm{card}(a^*(\boldsymbol{\mu}))}(\mathbb{1}_{1\in a^*(\boldsymbol{\mu})}, \dots, \mathbb{1}_{K\in a^*(\boldsymbol{\mu})})^T$. 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 $N_a(t) = \sum_{s\in[t]} \mathbb{1}_{A_s=a}$ be the random number of draws of arm $a \in [K]$ up to time $t \in \mathbb{N}^*$, and if $N_a(t) \geq 1$, let $\hat{\mu}_a(t) = N_a(t)^{-1} \sum_{s\in[t]} Y_s \mathbb{1}_{A_s=a}$ be the maximum likelihood estimate of μ_a at time t. We use the vector notations $\boldsymbol{N}(t) = (N_a(t))_{a\in[K]}$ and $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_a(t))_{a\in[K]}$.

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 $\mathcal{CR}_{\boldsymbol{\mu}} \subset [0,1]^K$ for $\boldsymbol{\mu}$ on which one can efficiently find a bandit $\tilde{\boldsymbol{\mu}} \in \mathcal{CR}_{\boldsymbol{\mu}}$ maximizing the minimal weight $w_{\min}$:

$$\tilde{\boldsymbol{\mu}} \in \mathrm{argmax}_{\boldsymbol{\nu}\in\mathcal{CR}_{\boldsymbol{\mu}}} w_{\min}(\boldsymbol{\nu}) \,. \tag{2}$$

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, as $w_{\min}(\tilde{\boldsymbol{\mu}}) \geq w_{\min}(\boldsymbol{\mu})$. The exploration bias decreases with the number of observations, as $\mathcal{CR}_{\boldsymbol{\mu}}$ shrinks to $\{\boldsymbol{\mu}\}$, and in the end arms are sampled with frequencies close to the optimal weight vector $\boldsymbol{w}(\boldsymbol{\mu})$.

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,

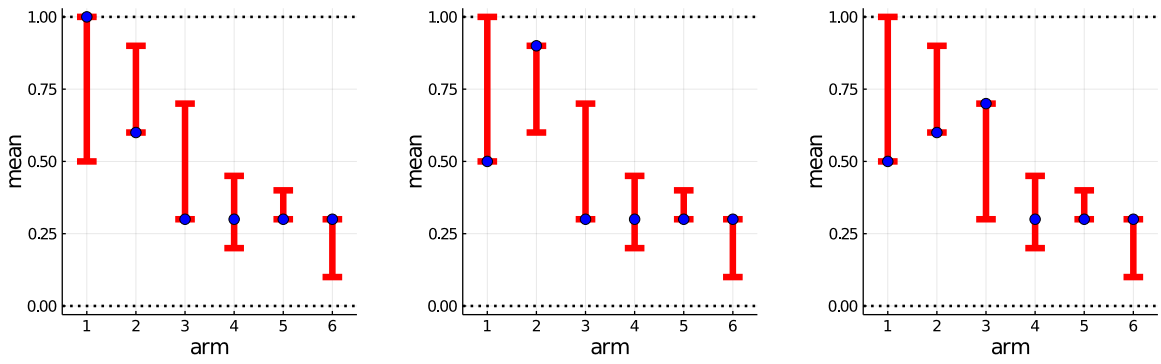

Figure 1: List of bandits $(\tilde{\boldsymbol{\mu}}^{\text{test}(a)})_{a\in\text{PotentialBest}}$ tried by Algorithm 1 for the example confidence region in red with PotentialBest $= \{1, 2, 3\}$. From left to right: $\tilde{\boldsymbol{\mu}}^{\text{test}(1)}$, $\tilde{\boldsymbol{\mu}}^{\text{test}(2)}$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

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. The algorithm requires Optimal Weights (Algorithm 4 of Appendix C.3), an efficient procedure for solving optimization problem (1) (see also Section 3.2).

- 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 by 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.

Algorithm 1: Exploration-Biased Weights

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

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

maxLB $\leftarrow \max_{a\in[K]} \underline{\mu}_a$; minUB $\leftarrow \min_{a\in[K]} \overline{\mu}_a$
if *minUB* $\geq$ *maxLB* **then**
  $\tilde{\boldsymbol{\mu}} \leftarrow (\text{minUB}, \ldots, \text{minUB})$; $\boldsymbol{w} \leftarrow (\frac{1}{K}, \ldots, \frac{1}{K})$
else
  PotentialBest $\leftarrow \{a \in [K] \ : \ \overline{\mu}_a > \text{maxLB}\}$
  $\boldsymbol{w} \leftarrow (0, \ldots, 0)$
  for $a \in$ *PotentialBest* **do**
    $\tilde{\mu}_a^{\text{test}(a)} \leftarrow \overline{\mu}_a$
    for $b \in [K] \setminus \{a\}$ **do**
      $\tilde{\mu}_b^{\text{test}(a)} \leftarrow \max(\underline{\mu}_b, \text{minUB})$
    $\boldsymbol{w}^{\text{test}(a)} \leftarrow$ Optimal Weights$(\tilde{\boldsymbol{\mu}}^{\text{test}(a)})$
    if $\min_{b\in[K]} w_b^{test(a)} > \min_{b\in[K]} w_b$ **then**
      $\boldsymbol{w} \leftarrow \boldsymbol{w}^{\text{test}(a)}$; $\tilde{\boldsymbol{\mu}} \leftarrow \tilde{\boldsymbol{\mu}}^{\text{test}(a)}$

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 in (Even-Dar et al., 2006)).

Proposition 1. *Let* $\mathcal{CR} = \prod_{a\in[K]}[\underline{\mu}_a, \overline{\mu}_a] \subset [0,1]^K$ *and* $(\tilde{\boldsymbol{\mu}}, \boldsymbol{w}) \leftarrow$ Exploration-Biased Weights$(\mathcal{CR})$. *Then* $\boldsymbol{w} = \boldsymbol{w}(\tilde{\boldsymbol{\mu}})$ *and* $\tilde{\boldsymbol{\mu}}$ *satisfies Equation* (2).

The proof of Proposition 1 is given in Appendix C.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 μ_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.

Confidence regions Confidence regions are designed to satisfy two requirements. First we need products of confidence intervals in order to use Algorithm 1, and then we will require a time-uniform confidence guarantee as a key ingredient for the non-asymptotic analysis of Exploration-Biased Sampling. For $\gamma \in (0,1)$, we define for $t \in [\![K, \tau_\delta]\!]$

$$\mathcal{CR}_{\boldsymbol{\mu}}(t) = \prod_{a\in[K]}\left[\hat{\mu}_a(t) \pm C_{\gamma/K}(N_a(t))\right] , \tag{3}$$

where $C_\gamma(s) = 2\sqrt{\frac{\log(4s/\gamma)}{s}}$. The following Lemma, proved in Appendix B, states a time-uniform γ-confidence guarantee for $\boldsymbol{\mu}$.

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

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in [\![K, \tau_\delta]\!] : \boldsymbol{\mu} \notin \mathcal{CR}_{\boldsymbol{\mu}}(t)\right) \le \gamma .$$

Stopping rule Following Garivier and Kaufmann (2016), our stopping rule relies on the statistic

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

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

$$Z_{a,b}(t) = \frac{1}{2}\frac{N_a(t)N_b(t)}{N_a(t)+N_b(t)}(\hat{\mu}_a(t) - \hat{\mu}_b(t))\left|\hat{\mu}_a(t) - \hat{\mu}_b(t)\right| .$$

The Exploration-Biased Sampling strategy is summarized in Algorithm 2. 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, as D-tracking appears to perform slightly better (replace $\sum_{s\in[t]} \tilde{w}_a(s)$ by $t\tilde{w}_a(t)$ in the description of Algorithm 2 for D-tracking).

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 used by Track-and-Stop for forced exploration, which appears here naturally.

Lemma 3. *For any choice of parameters and* $\boldsymbol{\mu} \in \mathcal{G}$, Exploration-Biased Sampling *satisfies*

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

Algorithm 2: Exploration-Biased Sampling

Input: confidence level δ
threshold function $\beta(t,\delta)$
confidence parameter γ
Output: stopping time τ_δ
estimated best arm $\hat{a}_{\tau_\delta}$

Observe each arm once ; $t \leftarrow K$
for $s = 0$ **to** $K-1$ **do**
  $\tilde{\boldsymbol{w}}(s) \leftarrow (1/K, \dots, 1/K)$
while $Z(t) \le \beta(t,\delta)$ **do**
  $\mathcal{CR}_{\boldsymbol{\mu}}(t) \leftarrow \prod_{a\in[K]}[\hat{\mu}_a(t) \pm C_{\gamma/K}(N_a(t))]$
  $(\tilde{\boldsymbol{\mu}}(t), \tilde{\boldsymbol{w}}(t)) \leftarrow$ Exploration-Biased Weights$(\mathcal{CR}_{\boldsymbol{\mu}}(t))$
  Choose $A_{t+1} \in \operatorname{argmin}_{a\in[K]} N_a(t) - \sum_{s\in[t]} \tilde{w}_a(s)$
  Observe $Y_{A_{t+1}}$ and increase t by 1
$\tau_\delta \leftarrow t$; $\hat{a}_{\tau_\delta} \leftarrow \operatorname{argmax}_{a\in[K]} \hat{\mu}_a(t)$

The proof of this lemma can be found in Appendix F.1.

The practical advantages of Exploration-Biased Sampling over Track-and-Stop are discussed in Section 4. On the theoretical level, we now show that (contrary to Track-and-Stop) this exploration strategy is adequate for obtaining non-asymptotic bounds.

2.3 Theoretical Results

A δ-correct strategy The δ-correctness of our strategy, 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 δ, γ and threshold*

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

is δ-correct.

Our main result is to obtain high probability bounds for τ_δ in finite horizon for Exploration-Biased Sampling, which is summarized in the following theorem.

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 $0 < \delta \le \delta_0$, algorithm* Exploration-Biased Sampling *with the threshold of Equation (4) satisfies*

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\tau_\delta > t \ \cap \ \mathcal{E}\big) \le 2Kt \exp\Big(-\frac{t w_{\min}(\boldsymbol{\mu})}{4T(\boldsymbol{\mu})^2} \frac{1}{\log^{\frac{2}{3}}(1/\delta)}\Big) \tag{5}$$

for any $t > (1+\eta)T(\boldsymbol{\mu})\log(1/\delta)$, and

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta \mathbb{1}_{\mathcal{E}}] \le (1+\eta)T(\boldsymbol{\mu})\log(1/\delta) + \frac{2^7 K T(\boldsymbol{\mu})^4}{w_{\min}(\boldsymbol{\mu})^2} \exp\Big(-\frac{w_{\min}(\boldsymbol{\mu})}{4T(\boldsymbol{\mu})} \log^{\frac{1}{3}}(1/\delta)\Big) \log^2(1/\delta)\,. \tag{6}$$

Note that:

- using the results of Section 3, one can show that $w_{\min}(\boldsymbol{\mu}) \geq \frac{\Delta_{\min}(\boldsymbol{\mu})}{2K}$ for any $\boldsymbol{\mu} \in \mathcal{G}^*$ (see Lemma 28 in Appendix F.1),
- the proof of Theorem 5 provides an explicit expression for δ_0,
- the second term of Bound (6) tends to 0 when δ decreases to 0, and hence negligible with respect to the first term: the sample complexity is therefore arbitrarily close to the lower bound.

We additionally prove that, from an asymptotic point of view, the Exploration-Biased Sampling algorithm presents the same guarantees as Track-and-Stop (see also Theorem 30 in Appendix F.2):

Theorem 6 (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*

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

Appendix D will be devoted to the proof of Theorem 5 while the proof of Theorem 6 can be found in Appendix F.3.

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-sub-Gaussian arms with means in $[0,1]$ (in which case, of course, a better lower bound might hold); indeed, these proofs only rely on sub-Gaussian 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 C.

In this section, 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 $\boldsymbol{v} \in \Sigma_K$, we define:

$$g(\boldsymbol{\mu}, \boldsymbol{v}) = \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [K]} v_a \frac{(\mu_a - \lambda_a)^2}{2} \tag{7}$$

$$= \frac{1}{2} \min_{a \neq a^*} \frac{v_{a^*} v_a}{v_{a^*} + v_a} \Delta_a(\boldsymbol{\mu})^2 \ . \tag{8}$$

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

$$T(\boldsymbol{\mu})^{-1} = g(\boldsymbol{\mu}, \boldsymbol{w}(\boldsymbol{\mu})) \ , \tag{9}$$

$$\text{and} \qquad Z(t) = t\, g\Big(\hat{\boldsymbol{\mu}}(t), \frac{\boldsymbol{N}(t)}{t}\Big) \tag{10}$$

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, let $a^* = a^*(\boldsymbol{\mu})$, $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\mu})$, $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{\mu})$, $w_{\min} = w_{\min}(\boldsymbol{\mu})$ and $T = T(\boldsymbol{\mu})$.

3.1 Solving the Optimization Problem

We define

$$\phi_{\boldsymbol{\mu}} : r \in \left(\frac{1}{\Delta_{\min}^2}, +\infty\right) \longmapsto \sum_{a \neq a^*} \frac{1}{(r\Delta_a^2 - 1)^2} - 1 \,. \tag{11}$$

Lemma 7. *$\phi_{\boldsymbol{\mu}}$ 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 8. *Let $r = r(\boldsymbol{\mu})$ be the solution of $\phi_{\boldsymbol{\mu}}(r) = 0$. Then*

$$w_{a^*} = \frac{1}{1 + \sum_{a \neq a^*} \frac{1}{r\Delta_a^2 - 1}} \,, \tag{12}$$

$$\forall a \neq a^*, \quad w_a = \frac{w_{a^*}}{r\Delta_a^2 - 1} \,, \tag{13}$$

$$\textit{and} \qquad T = 2\frac{r}{w_{a^*}} \,. \tag{14}$$

Besides,

$$w_{a^*} = \sqrt{\sum_{a \neq a^*} {w_a}^2} \,. \tag{15}$$

Recall that in the case of 2 arms, $\boldsymbol{w}(\boldsymbol{\mu}) = (0.5, 0.5)$. Besides, the monotonicity of the optimal weights with respect to the gaps follows from Equation (13).

Corollary 9. *Assume that $K \geq 3$. Then*

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

Equation (13) 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 Δ_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 8.

3.2 Bounds and Computation of the Problem Characteristics

By Proposition 8, it suffices to compute r to obtain the values of both T and $\boldsymbol{w}$. 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 4 of Appendix C.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 $\boldsymbol{w}$ and the characteristic time T.

Proposition 10. *Denoting by* $\overline{\Delta^2} = \frac{1}{K-1}\sum_{a\neq a^*}\Delta_a^2$ *the average squared gap,*

$$\max\left(\frac{2}{\Delta_{\min}^2}, \frac{1+\sqrt{K-1}}{\overline{\Delta^2}}\right) \leq r \leq \frac{1+\sqrt{K-1}}{\Delta_{\min}^2}\,, \tag{16}$$

$$\frac{1}{1+\sqrt{K-1}} \leq w_{\max} \leq \frac{1}{2}\,, \tag{17}$$

$$\max\left(\frac{8}{\Delta_{\min}^2}, 4\frac{1+\sqrt{K-1}}{\overline{\Delta^2}}\right) \leq T \leq 2\frac{\left(1+\sqrt{K-1}\right)^2}{\Delta_{\min}^2}\,. \tag{18}$$

Note that all of these inequalities can be reached for certain parameters $\boldsymbol{\mu}$, as discussed in Appendix C.2 after the proof of Proposition 10.

3.3 Monotonicity of the min-max Problem

We now show monotonicity results of the mappings $\boldsymbol{\nu} \mapsto T(\boldsymbol{\nu})$ and $\boldsymbol{\nu} \mapsto \boldsymbol{w}(\boldsymbol{\nu})$ when moving arm(s). When $K = 2$, the optimization problem is simple and leads to $\boldsymbol{w}(\boldsymbol{\mu}) = (0.5, 0.5)$ and $T(\boldsymbol{\mu}) = 8\Delta_2^2$, so that we assume 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 $\boldsymbol{\Delta}'$, $\boldsymbol{w}'$, $w'_{\min}$, T' and r' similarly to problem $\boldsymbol{\mu}$. The three following lemmas, which are the key ingredients to prove Proposition 1, are shown in Appendix C.4.

Lemma 11. *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 12. *Assume that* $\Delta'_a = \Delta_a + d$ *for every* $a \neq a^*$ *and some* $d > 0$*. Then* $w'_{\min} \geq w_{\min}$*, with strict inequality whenever* $\Delta_a \neq \Delta_b$ *for some* $a, b \neq a^*$.

Lemma 13. *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* $w'_{\min} \geq w_{\min}$.

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

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

Theorem 14. *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 g.

Proposition 15. *Let* $\boldsymbol{v} \in \Sigma_K$*. Then:*

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

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 the proof of the non-asymptotic bounds of Theorem 5.

4 Numerical experiments

In this section, we discuss the behavior and performance of Exploration-Biased Sampling for practical values of confidence δ. We propose a comparison with Track-and-Stop, Chernoff-Racing and LUCB++, and begin with a reminder on those strategies.

Track-and-Stop The strategy tracks the optimal weights $\boldsymbol{w}(\boldsymbol{\mu})$ by estimating it by $\boldsymbol{w}(\hat{\boldsymbol{\mu}}(t))$. Some exploration rate is forced to ensure that bad initial observations does not lead to an under-sampling of some arms (the strategy ensures that each $N_a(t)$ growths at least in $\sqrt{t}$). The stopping rule is the same as the one presented for Exploration-Biased Sampling.

Chernoff-Racing The strategy is divided into rounds during which the arms of a currently active set are sampled once. At the end of each round, a decision is made to keep or eliminate the current worst arm from the active set. 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.

LUCB++ The strategy (Simchowitz et al., 2017) (see also Kalyanakrishnan1 et al., 2012; Howard et al., 2021) samples two arms at each round: the one with the current best estimate and the one in the remaining arms with the highest optimistic indice $U_a(t)$ which is an upper confidence bound:

$$U_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{3}{N_a(t)} \log\Big(\frac{\log(N_a(t)) \times 2K}{\delta}\Big)}$$

(constant $\sqrt{3}$ appeared to be empirically optimal). For the fairness of the comparison we will take the same stopping condition as Track-and-Stop and Exploration-Biased Sampling.

Exploration-Biased Sampling We ran our experiments with confidence lengths $C_\gamma(s) = \sqrt{\frac{\log(s/\gamma)}{s}}$, and for all strategies we used the same threshold

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

These choices are more aggressive than what the theoretical analysis suggests: yet, empirically, they appears to guarantee the desired failure rate. Using the larger intervals of Section 2 would have increased the number of rounds with uniform exploration, and using larger thresholds unnecessarily delays the stopping for all strategies.

We now discuss the numerical pros and cons of Exploration-Biased Sampling.

Improving the Stability of Track-and-Stop In Section 1, we highlighted the weaknesses of Track-and-Stop, especially the forced exploration parameter and the non-interpretable and unstable sampling strategy during the first rounds. On Figures 2 and 3 we see the improvements of Exploration-Biased Sampling concerning those behaviours. During the first rounds, as for a racing algorithm, a uniform sampling is observed as the learner has not collected enough information (the confidence intervals on all arms are not separated), which is the expected behavior. Then the best arms are sampled more and more often, but still in a more cautious way than Track-and-Stop. We observe on Figure 3 the stability of the sampling strategies comparing to Track-and-Stop during the first rounds: 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 forced exploration at regular rounds (giving the yellow and blue peaks), which is unnecessary for Exploration-Biased Sampling as a natural exploration is always performed (Lemma 3).

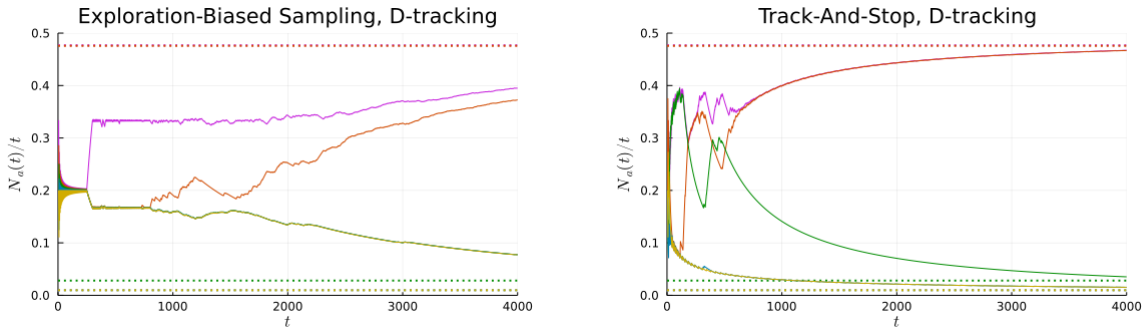

Figure 2: Evolution of the Sampling Frequencies $\boldsymbol{N}(t)/t$ on a Simulation of Exploration-Biased Sampling and Track-and-Stop. ($\delta = 0.01$, $\gamma = 0.2$, and $\boldsymbol{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$; the values of $\boldsymbol{w}(\boldsymbol{\mu}) = (0.477, 0.476, 0.028, 0.010, 0.010)$ are dotted)

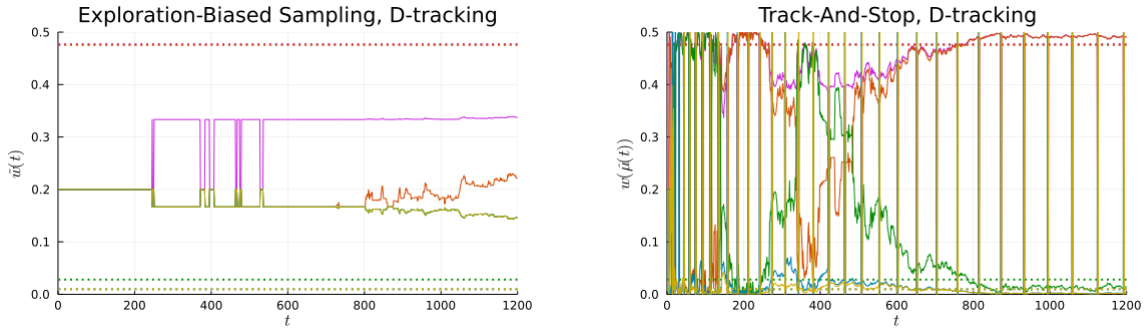

Figure 3: Evolution of the Targeted Weights $\tilde{\boldsymbol{w}}(t)$ (resp. $\boldsymbol{w}(\hat{\mu}(t))$) During the First 1200 Rounds on a Simulation of Exploration-Biased Sampling (resp. Track-and-Stop). ($\delta = 0.01$, $\gamma = 0.2$, $\boldsymbol{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$)

Comparisons of the Strategies The cost of the the cautiousness of the algorithm (the exploration-biased weights) 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 that occurs for every bandit parameter[1]. This can be observed on Table 1, where we present the performances of Exploration-Biased Sampling, Track-and-Stop, Chernoff-Racing and LUCB++ with two scenarios and a set of parameters. Exploration-Biased Sampling globally performs correctly but we see that the other strategies are always a little more efficient. Note that when increasing γ, the confidence intervals reduces so that the targeted weights are closer to $\boldsymbol{w}$, improving the performance of the algorithm. For similar reasons the initial cautiousness of the strategy disappears at long-term, thus when δ is very small the relative performance of Track-and-Stop and Exploration-Biased Sampling gets 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 $w_1 \gg w_2$).

Chernoff-Racing shows great performance with both $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$. This strategy samples the two last arms of the race equally often, thus can be optimal only when $\boldsymbol{w}(\boldsymbol{\mu})$ has its two highest components of similar value, e.g. when the two best arms are well separated from the others : this is the case of bandit $\boldsymbol{\mu}^{(1)}$. For $\boldsymbol{\mu}^{(2)}$ any strategy performs well as the problem is easy. However, Chernoff-Racing (whose theoretical analysis remains to be written) leads to a few more misidentifications in our experiments that might be linked to the stopping rule we chose here; for fairness reasons, it was taken identical to that of the other algorithms. LUCB++ presents similar performance with Chernoff-Racing, which

[1]Note that the cautiousness of our strategy is required to obtain the non-asymptotic bounds of Theorem 5.

Table 1: Empirical Expected Number of Draws $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$, Averaged over 1000 Experiments: $\boldsymbol{\mu}^{(1)} = (0.9, 0.8, 0.6, 0.4, 0.4)$, $\boldsymbol{w}(\boldsymbol{\mu}^{(1)}) = (0.477, 0.476, 0.028, 0.010, 0.010)$; $\boldsymbol{\mu}^{(2)} = (0.9, 0.5, 0.45, 0.4)$, $\boldsymbol{w}(\boldsymbol{\mu}^{(2)}) = (0.375, 0.286, 0.195, 0.144)$

Bandit	δ	γ	$T\,\mathrm{kl}(\delta, 1-\delta)$	EBS C	TaS C	EBS D	TaS D	Racing	LUCB++
$\boldsymbol{\mu}^{(1)}$	0.1	0.05	1476	4727	3597	4191	3477	3124	3353
$\boldsymbol{\mu}^{(1)}$	0.01	0.05	3782	7363	5664	6330	5584	5419	5549
$\boldsymbol{\mu}^{(1)}$	0.01	0.2	3782	7090	5664	6136	5584	5419	5372
$\boldsymbol{\mu}^{(1)}$	10^{-5}	0.2	9669	13801	12181	12376	11439	11557	11644
$\boldsymbol{\mu}^{(2)}$	0.1	0.05	135	476	367	470	322	405	365
$\boldsymbol{\mu}^{(2)}$	0.01	0.05	347	708	588	699	485	542	565

can be explained by the similar behaviour of the strategies: LUCB++ samples half time the best arm asymptotically, and the worst arms are eliminated one by one once their indice fall under the two best estimates.

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 $\boldsymbol{\mu}$ (thus the empirical proportions of draws converge faster to the optimal weight), while the impact is diluted in time with C-tracking. However we did not prove theoretical guarantees for D-tracking.

Additional experiments showing and interpreting the dependence on parameter δ of Exploration-Biased Sampling are postponed to Appendix G.

5 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, observing uniformly before having a high certainty that one arm is better than another.

It would be interesting but it remains out of reach to generalize this approach to non-Gaussian models: this requires to extend our results on the sample-complexity optimization problem, technically challenging task for which the simple and clean arguments developed here are likely to be replaced by much more involved derivations, if this is possible. In addition, it will be necessary to 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 ε-best arm identification.

Acknowledgements

Aurélien Garivier and Antoine Barrier acknowledges the support of the Project IDEXLYON of the University of Lyon, in the framework of the Programme Investissements d'Avenir (ANR-16-IDEX-0005), and Chaire SeqALO (ANR-20-CHIA-0020-01).

References

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

Chen, L., Li, J., and Qiao, M. (2017). Towards Instance Optimal Bounds for Best Arm Identification. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 535–592. PMLR.

Chernoff, H. (1959). Sequential Design of Experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

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, 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, 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(39):1079–1105.

Garivier, A., Hadiji, H., Menard, P., and Stoltz, G. (2019). KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *arXiv:1805.05071*.

Garivier, A. and Kaufmann, E. (2016). Optimal Best Arm Identification with Fixed Confidence. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *Conference on Learning Theory*, volume 49, pages 998–1027. PMLR.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. In Balcan, M. F., Feldman, V., and Szepesvári, C., editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 423–439. PMLR.

Kalyanakrishnan1, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC Subset Selection in Stochastic Multi-armed Bandits. In *Proceedings of the 29th International Conference on Machine Learning*.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost Optimal Exploration in Multi-Armed Bandits. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1238–1246. PMLR.

Kaufmann, E. and Kalyanakrishnan, S. (2013). Information Complexity in Bandit Subset Selection. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 228–251. PMLR.

Kocák, T. and Garivier, A. (2020). Best Arm Identification in Spectral Bandits. In Bessiere, C., editor, *Proceedings of The 29th International Joint Conference on Artificial Intelligence*, volume 3, pages 2220–2226. International Joint Conferences on Artificial Intelligence Organization.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press, Cambridge.

Ménard, P. (2019). Gradient Ascent for Active Exploration in Bandit Problems. *arXiv:1905.08165*.

Russo, D. (2016). Simple Bayesian Algorithms for Best Arm Identification. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *Proceedings of the 2016 Conference on Learning Theory*, volume 49, pages 1417–1418. PMLR.

Shang, X., Heide, R., Menard, P., Kaufmann, E., and Valko, M. (2020). Fixed-confidence guarantees for Bayesian best-arm identification. In Chiappa, S. and Calandra, R., editors, *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 1823–1832. PMLR.

Simchowitz, M., Jamieson, K., and Recht, B. (2017). The Simulator: Understanding Adaptive Sampling in the Moderate-Confidence Regime. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1794–1834. PMLR.

Tirinzoni, A., Pirotta, M., Restelli, M., and Lazaric, A. (2020). An Asymptotically Optimal Primal-Dual Incremental Algorithm for Contextual Linear Bandits. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc.

Appendix outline

The appendix is organized as follows:

A. Precise description of the Track-and-Stop strategy

B. Proof of the time-uniform confidence regions guarantees for $\boldsymbol{\mu}$ (Lemma 2)

C. Proofs of the results on the sample complexity for Gaussian arms (Section 3)

D. Proof of the non-asymptotic result (Theorem 5)

E. Technical results associated to the proof of Theorem 5 (complements to Appendix D)

F. Asymptotic analysis of Exploration-Biased Sampling (Theorems 6 and 30)

G. Additional experiments to see the dependency of Exploration-Biased Sampling in δ

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 set $\hat{\mu}_{a,n} = \frac{1}{n}\sum_{p\in[n]} X_{a,p}$ for all $n \geq 1$ and assume that

$$\forall t \geq K, \quad \hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)} \,. \tag{19}$$

A The Track-and-Stop strategy

We recall the description of the Track-and-Stop strategy in Algorithm 3. We use the notations of Section 2 and algorithm Optimal Weights (Algorithm 4 of Appendix C.3) which efficiently computes the solution of optimization problem (1).

Algorithm 3: Track-and-Stop

Input: confidence level δ
threshold function $\beta(t,\delta)$
Output: stopping time τ_δ
estimated best arm $\hat{a}_{\tau_\delta}$

Observe each arm once ; $t \leftarrow K$
for $s = 0$ **to** $K-1$ **do**
  $\tilde{\boldsymbol{w}}(s) \leftarrow (1/K, \dots, 1/K)$
while $Z(t) \leq \beta(t,\delta)$ **do**
  if $U_t = \{a \in [K] \,:\, N_a(t) < \sqrt{t} - K/2\} \neq \varnothing$ **then**
    Choose $A_{t+1} \in \operatorname{argmin}_{a\in U_t} N_a(t)$ `/* forced exploration */`
  else
    $\tilde{\boldsymbol{w}}(t) \leftarrow$ Optimal Weights$(\hat{\boldsymbol{\mu}}(t))$
    Choose $A_{t+1} \in \operatorname{argmin}_{a\in[K]} N_a(t) - \sum_{s\in[t]} \tilde{w}_a(s)$ `/* C-tracking */`
  Observe $Y_{A_{t+1}}$ and increase t by 1
$\tau_\delta \leftarrow t$; $\hat{a}_{\tau_\delta} \leftarrow \operatorname{argmax}_{a\in[K]} \hat{\mu}_a(t)$

The presented algorithm uses C-tracking (the cumulative sums of the weights are tracked), but one can consider D-tracking for a direct track of the current weight (by replacing $\sum_{s\in[t]} \tilde{w}_a(s)$ by $t\tilde{w}_a(t)$).

B Proof of Lemma 2

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

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

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 (19):

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

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

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

(a) is obtained using the fact that $n \mapsto nC_\gamma(n)$ 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).

C Proofs of results presented in Section 3

In this appendix, we first prove Proposition 8, then we focus on the consequences developed in Section 3.

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

C.1 Solving the Optimization Problem

Proof of Equation (8). Let $\boldsymbol{v} \in \Sigma_K$. One has:

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

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

Proof of Proposition 8. Let us define, for some $v_1 \in [0,1]$:

$$C(v_1) = \max_{v_{2:K} \,:\, \boldsymbol{v}\in\Sigma_K} \min_{a\neq 1} \frac{v_1 v_a}{v_1+v_a}\Delta_a^2 \tag{20}$$

so that

$$T^{-1} = \max_{\boldsymbol{v}\in\Sigma_K} g(\boldsymbol{\mu},\boldsymbol{v}) = \frac{1}{2}\max_{v_1\in[0,1]} C(v_1)\ . \tag{21}$$

Fix $v_1 \in [0,1]$. The maximum in Equation (20) 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:

$$\forall a \neq 1, \quad v_a = \frac{v_1 C}{v_1\Delta_a^2 - C}\ . \tag{22}$$

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

$$\Phi(v_1, C) := v_1 + \sum_{a\neq 1} \frac{v_1 C}{v_1\Delta_a^2 - C} - 1 = 0\ . \tag{23}$$

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

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

Hence $C(v_1)$ is a smooth non-negative function with a continuous derivative. By Equation (21), it vanishes when $v_1 \to 0$ and $v_1 \to 1$, 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

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

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

Equations (12), (13) and (14) can be respectively derived from (23), (22) and (21). It remains to obtain Equation (15) by combining Equation (13) 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 9. When a, b are suboptimal, the result is a direct consequence of Equation (13) of Proposition 8. It remains to see that $w_1 > \max_{a\neq 1} w_a$, which is a direct consequence of Equation (15) and the fact that all weights are positive. □

C.2 Proof of Proposition 10

Defining $q_a = \frac{1}{r\Delta_a^2 - 1}$ for $a \neq 1$, we will use that, as $\phi_{\boldsymbol{\mu}}(r) = 0$, the $(q_a^2)_{a\neq 1}$ are positive and sum to 1, hence for any $a \neq 1$ one has $q_a \leq 1$ (with strict inequality when $K \geq 3$).

Let us begin with Equation (17). As we assume $a^* = 1$, $w_{\max} = w_1$ by Corollary 9. Using Equation (12) of Proposition 8 one has:

- on the one hand

$$\begin{aligned} w_1 &= \Big(1 + \sum_{a\neq 1} \frac{1}{r\Delta_a^2 - 1}\Big)^{-1} && \text{by Equation (12) of Proposition 8} \\ &\leq \Big(1 + \sum_{a\neq 1} \frac{1}{(r\Delta_a^2 - 1)^2}\Big)^{-1} && \text{as } q_a \leq 1 \\ &= \frac{1}{2} && \text{as } \phi_{\boldsymbol{\mu}}(r) = 0 \end{aligned}$$

  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 (16) :

- 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 $x \mapsto \frac{1}{(rx-1)^2}$:

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

  and hence $\phi_{\boldsymbol{\mu}}(\frac{1+\sqrt{K-1}}{\overline{\Delta^2}}) \geq 0$, which by decreasing of $\phi_{\boldsymbol{\mu}}$ (Lemma 7) gives $r \geq \frac{1+\sqrt{K-1}}{\overline{\Delta^2}}$.

- one can also check that

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

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

Finally, combining the obtained inequalities with Equation (14) yields Equation (18).

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

- First note that when $K = 2$, lower and upper bounds match in Inequalities (16), (17) and (18). 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 $T = 2\big(1+\sqrt{K-1}\big)^2/\Delta_{\min}^2$ are reached if and only if $\Delta_{\min} = \Delta_{\max}$, or in other words $\Delta_2 = \cdots = \Delta_K$. 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.

C.3 Computing r

At the sight of Proposition 8, 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 7). Hence, when initialized with a value $r_0 < r$, the iterates of a Newton procedure remain smaller than r. The lower bound of Inequalities (16) of Proposition 10 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 4 for details.

Algorithm 4: Optimal Weights

Input: bandit $\boldsymbol{\mu} \in \mathcal{G}^*$ with best arm 1
tolerance parameter tol (typically 10^{-10})
Output: optimal weight vector $\boldsymbol{w}$
characteristic time T

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

C.4 On the monotonicity of the min-max problem

In this section we prove Lemmas 11, 12 and 13, and then use those Lemmas to prove Proposition 1. We recall that we assume $K \geq 3$ in this section (note that Proposition 1 is trivial when $K = 2$).

Proof of Lemma 11.

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 $a\notin\{1,b\}$ 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 $\phi_{\boldsymbol{\mu}}(r)=0=\phi_{\boldsymbol{\mu}'}(r')$:
$$\frac{1}{r'{\Delta'_b}^2-1}<\frac{1}{r\Delta_b^2-1}$$
or equivalently $r'{\Delta'_b}^2-1>r\Delta_b^2-1$.
Combining those inequalities with Equation (13) of Proposition 8, 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 $\boldsymbol{w}(\boldsymbol{\nu})$ 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 (12) and (13) of Proposition 8. 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:
$$\begin{aligned}
\frac{\partial F_a}{\partial d_b}&=\frac{\partial r}{\partial d_b}d_a+\sum_{c\neq 1}\left[\frac{\frac{\partial r}{\partial d_b}d_a}{rd_c-1}-\frac{rd_a-1}{(rd_c-1)^2}\left(\frac{\partial r}{\partial d_b}d_c\right)\right]-\frac{rd_a-1}{(rd_b-1)^2}r\\
&=\frac{\partial r}{\partial d_b}d_a\Big(1+\sum_{c\neq 1}\frac{1}{rd_c-1}-\frac{rd_c}{(rd_c-1)^2}\Big)+\frac{\partial r}{\partial d_b}\sum_{c\neq 1}\frac{d_c}{(rd_c-1)^2}-\frac{rd_a-1}{(rd_b-1)^2}r\\
&=\frac{\partial r}{\partial d_b}d_a\sum_{c\neq 1}\underbrace{\frac{1+(rd_c-1)-rd_c}{(rd_c-1)^2}}_{=0}+\frac{\partial r}{\partial d_b}\sum_{c\neq 1}\frac{d_c}{(rd_c-1)^2}-\frac{rd_a-1}{(rd_b-1)^2}r\\
&=\frac{\partial r}{\partial d_b}\sum_{c\neq 1}\frac{d_c}{(rd_c-1)^2}-\frac{rd_a-1}{(rd_b-1)^2}r
\end{aligned}$$

(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 Δ_b increases. In fact, we already noticed that by showing that $r' < r$ in the first part of the proof of Lemma 11. 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 (9) and (8):

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

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 $\boldsymbol{w}$ and the fact that $\boldsymbol{w} \neq \boldsymbol{w}'$, as previously obtained.

□

Before proving Lemmas 12 and 13, we show the following result.

Lemma 16. *Assume that there exists $\kappa > 0$ such that $\Delta'_a = \kappa\Delta_a$ for any $a \neq 1$. Then $\boldsymbol{w}' = \boldsymbol{w}$.*

Proof of Lemma 16. As r is the unique solution of $\phi_{\boldsymbol{\mu}}(r) = 0$, one has:

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

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

This implies $r\Delta_a^2 = r'{\Delta'_a}^2$ for any $a \neq 1$, hence $\boldsymbol{w}' = \boldsymbol{w}$ by Equations (12) and (13) of Proposition 8. □

Proof of Lemma 12. 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 $\boldsymbol{w}'' = \boldsymbol{w}(\boldsymbol{\mu}'')$, we have $\boldsymbol{w}'' = \boldsymbol{w}'$ by Lemma 16. 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 apply Lemma 11 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 9

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

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 11 gives a strict increasing of $w_{\min}$. □

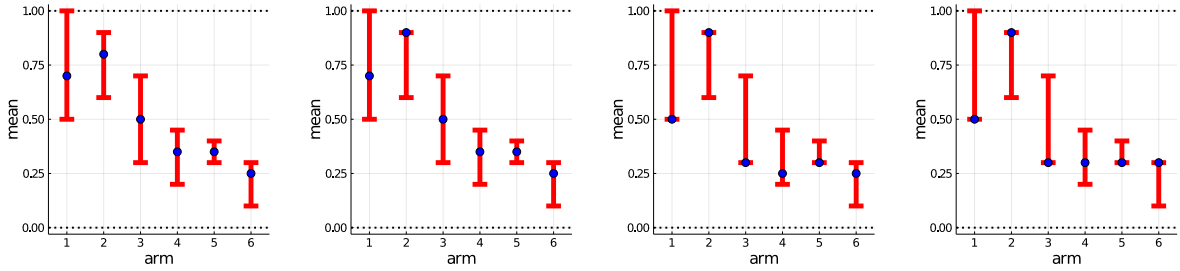

Figure 4: Transformations in the proof of Proposition 1, for some instance bandit $\boldsymbol{\nu}$. From left to right: $\boldsymbol{\nu}$, $\boldsymbol{\nu}^{(1)}$, $\boldsymbol{\nu}^{(2)}$, $\boldsymbol{\nu}^{(3)} = \tilde{\boldsymbol{\mu}}^{\mathrm{test}(2)}$

Proof of Lemma 13. Using scaling argument from Lemma 16, like in the proof of Lemma 12, 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 11, would mean that corresponding $w_{\min}$ increases. □

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 $\boldsymbol{w} = \boldsymbol{w}(\tilde{\boldsymbol{\mu}})$. When minUB $\geq$ maxLB the algorithm returns a constant bandit and $\boldsymbol{w} = (1/K, \dots, 1/K)$ which is its optimal weight vector by convention. As all weight vectors belong to Σ_K, the result is clear.

Now assume that minUB $<$ maxLB and fix $\boldsymbol{\nu} \in \mathcal{CR}$. If $\boldsymbol{\nu}$ as several optimal arms, then $w_{\min}(\boldsymbol{\nu}) = 0$ so that trivially $w_{\min}(\boldsymbol{\nu}) \leq w_{\min}(\tilde{\boldsymbol{\mu}})$. Assume now that $\boldsymbol{\nu}$ has a unique optimal arm denoted by a. Note that $a \in$ PotentialBest, so that we will show that $w_{\min}(\boldsymbol{\nu}) \leq w_{\min}(\tilde{\boldsymbol{\mu}}^{\mathrm{test}(a)})$ by transforming $\boldsymbol{\nu}$ to $\tilde{\boldsymbol{\mu}}^{\mathrm{test}(a)}$ 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 9. The procedure, illustrated in Figure 4, is the following:

1. Transform $\boldsymbol{\nu}$ into $\boldsymbol{\nu}^{(1)}$ by increasing arm a so that $\boldsymbol{\nu}_a^{(1)} = \overline{\mu}_a$. Using Lemma 12, one has $w_{\min}(\boldsymbol{\nu}^{(1)}) \geq w_{\min}(\boldsymbol{\nu})$.
2. Transform $\boldsymbol{\nu}^{(1)}$ into $\boldsymbol{\nu}^{(2)}$ by decreasing, for each arm $b \neq a$, μ_b to $\max(\underline{\mu}_b, \boldsymbol{\nu}_{\min})$. By several applications of Lemma 11, one has $w_{\min}(\boldsymbol{\nu}^{(2)}) \geq w_{\min}(\boldsymbol{\nu}^{(1)})$ (remark that imposing to stay above $\boldsymbol{\nu}_{\min}$ ensures that the associated worst arm stays one of the worst arms at each modification).
3. Transform $\boldsymbol{\nu}^{(2)}$ into $\boldsymbol{\nu}^{(3)}$ by increasing all the worst arms to minUB. By Lemma 13, one has $w_{\min}(\boldsymbol{\nu}^{(3)}) \geq w_{\min}(\boldsymbol{\nu}^{(2)})$.

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

$$\max_{\boldsymbol{\nu} \in \mathcal{CR}} w_{\min}(\boldsymbol{\nu}) = \max_{a \in \mathrm{PotentialBest}} w_{\min}(\tilde{\boldsymbol{\mu}}^{\mathrm{test}(a)}) = w_{\min}(\tilde{\boldsymbol{\mu}}) \ ,$$

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

C.5 Proof of Theorem 14

We have that

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

and

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

hence by monotonicity of $\phi_{\boldsymbol{\mu}'}$ 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

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

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 (16) of Proposition 10), 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

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

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 (14) of Proposition 8, on the one hand:

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

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

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

which yields by Equation (13) of Proposition 8, for any $a \neq 1$:

$$\begin{aligned}(1-10\varepsilon)w_a \le \frac{w_1/(1+5\varepsilon)}{(r\Delta_a^2-1)(1+\frac{2\varepsilon}{1+\varepsilon})} \le w_a' = \frac{w_1'}{r'{\Delta_a'}^2-1} \\ \le \frac{w_1/(1-3\varepsilon)}{(r\Delta_a^2-1)(1-2\frac{2\varepsilon}{1+\varepsilon})} = \frac{1+\varepsilon}{(1-3\varepsilon)^2}w_a \le (1+10\varepsilon)w_a\ .\end{aligned}$$

C.6 Proof of Proposition 15

We will prove Proposition 15 by combining two Lemmas. Note that in this section $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ are general bandits, with possibly more than one best arm.

Lemma 17. *Let $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{G}$ and $\boldsymbol{v} \in \Sigma_K$ be any optimal vector. Then:*

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

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:

$$\begin{aligned}g(\boldsymbol{\mu}', \boldsymbol{v}) - g(\boldsymbol{\mu}, \boldsymbol{v}) &= \frac{1}{2}\min_{a\neq 1}\frac{v_1 v_a}{v_1+v_a}{\Delta_a'}^2 - \frac{1}{2}\min_{b\neq 1}\frac{v_1 v_b}{v_1+v_b}\Delta_b^2 && \text{by Equation (8)} \\ &= \frac{1}{2}\min_{a\neq 1}\max_{b\neq 1}\frac{v_1 v_a}{v_1+v_a}{\Delta_a'}^2 - \frac{v_1 v_b}{v_1+v_b}\Delta_b^2 \\ &\ge \frac{1}{2}\min_{a\neq 1}\frac{v_1 v_a}{v_1+v_a}\left({\Delta_a'}^2 - \Delta_a^2\right) && \text{taking } b = a\ .\end{aligned}$$

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

$$|\Delta_a - \Delta_a'| = |(\mu_1 - \mu_1') - (\mu_a - \mu_a')| \le |\mu_1 - \mu_1'| + |\mu_a - \mu_a'| \le 2\varepsilon$$

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

$$\left|\Delta_a^2 - {\Delta_a'}^2\right| = |\Delta_a - \Delta_a'|\,(\Delta_a + \Delta_a') \le 4\varepsilon\ .$$

  As $\boldsymbol{v}$ is an optimal vector, we have $0 \le v_a \le v_1 \le \frac{1}{2}$ using Equation (17), so that:

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

  hence

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

- 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

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

Select the subdivision $0 = t_0 < t_1 < \cdots < t_N = 1$ 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 $n \in [\![1, N-1]\!]$, $\boldsymbol{\mu}^{(t_n)}$ has at least two best arms so that $g(\boldsymbol{\mu}^{(t_n)}, \boldsymbol{v}) = 0$,
- $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}$ have a common best arm,
- $\boldsymbol{\mu}^{(N-1)}$ and $\boldsymbol{\mu}'$ have a common best arm.

Thus

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

□

Lemma 18. *Let $\boldsymbol{\mu}' \in \mathcal{G}$ be a Gaussian bandit and $\boldsymbol{u}, \boldsymbol{v} \in \Sigma_K$ 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:

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

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:

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

□

Proof of Proposition 15. The result follows directly by applying Lemmas 18 and 17 with $\boldsymbol{u} = \boldsymbol{w}(\boldsymbol{\mu})$:

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

□

D 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 $T = T(\boldsymbol{\mu})$.

Recall that the confidence regions are defined, for $t \in [\![K, \tau_\delta]\!]$, by

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

where $\ell_a(t) = C_{\gamma/K}(N_a(t)) = 2\sqrt{\frac{\log(4KN_a(t)/\gamma)}{N_a(t)}}$.

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

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

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 19 and proved in Appendix E.1.

Lemma 19. *On event $\mathcal{E}$ one has:*

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

This inequality directly implies the following lower bound:

$$\forall t \geq \frac{2K}{w_{\min}}, \quad \min_{a\in[K]} N_a(t) \geq \frac{t w_{\min}}{2} . \tag{24}$$

Proof outline

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 δ, 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 E.

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 (24) implies that

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

$L(t)$ is an arm-independent bound on the half-length of the confidence interval of each μ_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 14 allows to control the difference between $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}(t)$ for t large enough, as the following Lemma claims.

Lemma 20. *Let*

$$T_0 = \max\left(\frac{224^2}{\Delta_{\min}^2 w_{\min}} \log\Big(\frac{2 \times 224^2 eK}{\Delta_{\min}^2 w_{\min}\gamma}\Big), \frac{2K}{w_{\min}}\right). \tag{25}$$

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

$$\forall a \in [K], \quad w_a(1-\varepsilon_t) \leq \tilde{w}_a(t) \leq w_a(1+\varepsilon_t)\,. \tag{26}$$

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 $\boldsymbol{\mu} \in \mathcal{CR}_{\boldsymbol{\mu}}(t) = \prod_{a\in[K]}[\underline{\mu}_a(t), \overline{\mu}_a(t)]$, hence for any $a \neq 1$:

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

so that the confidence interval for μ_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

$$\tilde{\Delta}_a(t)^2 \leq (\Delta_a + 2L(t))^2 = \Delta_a^2\Big(1 + \frac{4L(t)}{\Delta_a} + \frac{4L(t)^2}{\Delta_a^2}\Big) \leq \Delta_a^2\Big(1 + \frac{8L(t)}{\Delta_{\min}}\Big)$$

$$\text{and} \quad \tilde{\Delta}_a(t)^2 \geq (\Delta_a - 2L(t))^2 = \Delta_a^2\Big(1 - \frac{4L(t)}{\Delta_a} + \frac{4L(t)^2}{\Delta_a^2}\Big) \geq \Delta_a^2\Big(1 - \frac{8L(t)}{\Delta_{\min}}\Big)\,.$$

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 14 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:

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

and this inequality is satisfied, by Lemma 26, for

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

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

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 25), and as $\varepsilon_t \to_{t\to+\infty} 0$, we are able to control the difference between $\boldsymbol{w}$ and $\boldsymbol{N}(t)/t$ after a deterministic time T_1.

Lemma 21. *Fix $\eta \in (0,1)$ and let*

$$T_1 = \frac{\max(640^2, 8K)}{\eta^2 \Delta_{\min}^2 w_{\min}^2} \log\Big(\frac{2 \times 640^2 eK}{\eta^2 \Delta_{\min}^2 w_{\min}\gamma}\Big)\,. \tag{27}$$

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

$$\forall a \in [K], \quad w_a(1-\eta) \leq \frac{N_a(t)}{t} \leq w_a(1+\eta)\,. \tag{28}$$

Proof. Let T_0 be defined by Equation (25). Let $t > T_0$ and $a \in [K]$. Equation (26) of Lemma 20 gives:

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

By definition of ε_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 25:

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

Thus the conclusion of the Lemma holds when:

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

and this inequality is satisfied, using Lemma 26, when:

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

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

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

and T_1 is greater than this lower bound. □

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 δ small enough, we have the following inclusion of events.

Lemma 22. *Fix $\eta \in (0, 0.15]$ and let δ be such that*

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

where T_1 is defined by Equation (27) and

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

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

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

Remark 23. *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 21, one has

$$\max_{a\in[K]} \left| \frac{w_a - N_a(t)/t}{w_a} \right| \leq \eta \,. \tag{29}$$

Then using Proposition 15 and Equation (9):

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

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

$$\forall C \in (0,1], \quad f\Big((1+C)\frac{(1+\eta)^2}{(1-\eta)^2}T\log(1/\delta)\Big) \leq \frac{1}{(1+C)T} \,. \tag{30}$$

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:

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

hence Inequality (30) is satisfied if

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

which is the case, by Lemma 26, when:

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

Finally when Inequality (30) holds we have for $t \geq (1+C)\frac{(1+\eta)^2}{(1-\eta)^2}T\log(1/\delta)$:

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

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

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

Fix $\eta \in (0,1]$ and assume in the following that conditions (C1) and (C2) of Lemma 22 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

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

Lemmas 22 and 27 – a consequence of Hoeffding's inequality – (note that Condition (C1) ensures that $t \geq \frac{2K}{w_{\min}}$) give the bound:

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\tau_\delta > t \ \cap \ \mathcal{E}\big) \leq \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_t) \leq 2Kt \exp\Big(- \frac{t w_{\min}}{4T^2} C^2 \Big) \ . \tag{31}$$

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

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

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

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

the condition holds as

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

It remains to focus on the bound of $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta \mathbb{1}_{\mathcal{E}}]$. Using Equation (31) we have:

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

Define

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

With some technical calculations (see Appendix E.4), one can obtain that:

Lemma 24. *One has*

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

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

$$S(C) \leq \frac{32T^4}{w_{\min}^2} \exp\Big(- \frac{w_{\min}}{4T} \log^{\frac{1}{3}}(1/\delta)\Big)\Big(\log^{\frac{5}{3}}(1/\delta) + \log^{\frac{4}{3}}(1/\delta)\Big) \leq \frac{64T^4}{w_{\min}^2} \exp\Big(- \frac{w_{\min}}{4T} \log^{\frac{1}{3}}(1/\delta)\Big) \log^2(1/\delta)$$

thus

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

Under Condition (C3) we get

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

and obtain the Bound (6) claimed in the theorem. Combining conditions (C1), (C2) and (C3) together, one can define δ_0 satisfying:

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

with some simplifications allowed by Equation (18) of Proposition 10.

E Technical details for the proof of Appendix D

E.1 Proof of Lemma 19

We will use the following deterministic Lemma:

Lemma 25. *One has:*

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

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

The claim is true for $t\in[\![0,K]\!]$ as Equation (17) of Proposition 10 gives

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

Otherwise, fix $t\in[\![K+1,\tau_\delta]\!]$ and $a\in[K]$. For any $s\in[\![0,K-1]\!]$, 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 $\tilde{w}_a(s)\geq w_{\min}$ ($\boldsymbol{w}\in\Sigma_K$ implies $w_{\min}\leq\frac{1}{K}$). For any $s\in[\![K,\tau_\delta-1]\!]$, one has by Proposition 1 :

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

as $\boldsymbol{\mu}\in\mathcal{CR}_{\boldsymbol{\mu}}(s)$ on event $\mathcal{E}$. Hence by Lemma 25

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

E.2 A technical lemma

Lemma 26. *For any $c_1,c_2>0$,*

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

is such that $c_1x\geq\log(c_2x)$.

This is a direct consequence of Garivier and Kaufmann (2016, Lemma 18).

E.3 Deviation bound

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

Lemma 27. *For any $t \geq \frac{2K}{w_{\min}}$ and $x > 0$, one has*

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

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

$$\begin{aligned}
\mathbb{P}\Big(|\hat{\mu}_a(t) - \mu_a| > x \ \cap \ \mathcal{E} \Big) &= \sum_{s=T}^{t} \mathbb{P}\Big(|\hat{\mu}_a(t) - \mu_a| > x \ \cap \ \mathcal{E} \ \cap \ N_a(t) = s \Big) && \text{by Equation (24)} \\
&\leq \sum_{s=T}^{t} \mathbb{P}\Big(|\hat{\mu}_{a,s} - \mu_a| > x \Big) && \text{by Equation (19)} \\
&\leq \sum_{s=T}^{t} 2 \exp\Big(- \frac{s}{2} x^2 \Big) && \text{by Hoeffding's inequality} \\
&\leq 2t \exp\Big(- \frac{T}{2} x^2 \Big)
\end{aligned}$$

giving the desired bound by union bound. □

E.4 Proof of Lemma 24

We have

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

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 $n_0 = \lfloor \frac{1}{A} \rfloor$. We will show that $S(C) \leq 2 \int_B^{+\infty} f(t) \, \mathrm{d}t$.

- If $B > n_0$ then f is decreasing on $[B, +\infty[$ and one has $S(C) \leq \int_B^{+\infty} f(t) \, \mathrm{d}t$.
- Otherwise, one has:

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

  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

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

We have:

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

where in the last inequality, we used the simple calculation

$$\int_Y^{+\infty} t \exp(-tX) \mathrm{d}t = \exp(-YX)\Big(\frac{Y}{X} + \frac{1}{X^2} \Big)$$

for $X, Y > 0$.

In both cases we have:

$$S(C) \leq 2 \int_{(1+C)\zeta T \log(1/\delta)}^{\infty} t \exp\Big(-\frac{t w_{\min}}{4T^2} C^2 \Big)\ \mathrm{d}t$$

and using the same calculation as before

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

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

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

F Proof of asymptotic results

F.1 Proof of Lemma 3

We will need the two following lemmas. The first gives a lower bound of $w_{\min}(\boldsymbol{\mu})$ and the second provides a lower bound on the minimal gap of the optimistic bandit computed by Algorithm 1.

Lemma 28. *For any $\boldsymbol{\mu} \in \mathcal{G}^*$ one hase $w_{\min}(\boldsymbol{\mu}) \geq \frac{\Delta_{\min}(\boldsymbol{\mu})}{2K}$.*

Proof. Let $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{\mu})$, $w_{\min} = w_{\min}(\boldsymbol{\mu})$ and $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\mu})$. We have

$$\begin{aligned} w_{\min} &= \frac{w_{\max}}{r\Delta_{\max} - 1} && \text{by Equation (13) of Proposition 8} \\ &\geq \frac{1}{\sqrt{K-1}+1} \times \frac{1}{\frac{\sqrt{K-1}+1}{\Delta_{\min}} \Delta_{\max} - 1} && \text{by Inequalities (16) and (17)} \\ &\geq \frac{\Delta_{\min}}{(\sqrt{K-1}+1)^2} && \text{as } \Delta_{\max}(t) \leq 1 \\ &\geq \frac{\Delta_{\min}}{2K}\ . \end{aligned}$$

□

Lemma 29. *Let $\mathcal{CR} = \prod_{a\in[K]}[\underline{\mu}_a, \overline{\mu}_a]$ be a confidence region such that $\underline{\mu}_a < \overline{\mu}_a$ for $a \in [K]$ and $\max_{a\in[K]} \underline{\mu}_a = maxLB > minUB = \min_{a\in[K]} \overline{\mu}_a$, and $(\tilde{\boldsymbol{\mu}}, \boldsymbol{v}) \leftarrow$ OptimisticWeights($\mathcal{CR}$). Then*

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

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

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

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 $\tilde{\mu}_1 > \tilde{\mu}_2 \geq \tilde{\mu}_3 \geq \cdots \geq \tilde{\mu}_K$. 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 $\tilde{\mu}_1 = \overline{\mu}_1$. In addition, our assumption implies $\Delta_{\min}(\tilde{\boldsymbol{\mu}}) < \overline{\mu}_1 - \underline{\mu}_1$, giving $\tilde{\mu}_2 > \underline{\mu}_1$. Recall that $\tilde{\mu}_2 = \max(\underline{\mu}_2, \text{minUB})$, so that we split our analysis to the two possible cases:

- if $\tilde{\mu}_2 = \underline{\mu}_2$, then we cannot have $\overline{\mu}_2 \leq \overline{\mu}_1 = \tilde{\mu}_1$ otherwise $\Delta_{\min}(\tilde{\boldsymbol{\mu}}) > \overline{\mu}_2 - \underline{\mu}_2$, which is impossible. Then $\overline{\mu}_2 > \overline{\mu}_1$. By defining $\boldsymbol{\nu} = (\tilde{\mu}_2, \overline{\mu}_2, \tilde{\mu}_3, \ldots, \tilde{\mu}_K)$, one has $\boldsymbol{\nu} \in \mathcal{CR}$ and $w_{\min}(\boldsymbol{\nu}) > w_{\min}(\tilde{\boldsymbol{\mu}})$ by Lemma 12. Thus $\tilde{\boldsymbol{\mu}}$ cannot maximize $w_{\min}$ over $\mathcal{CR}$ which is in contradiction with Proposition 1.
- if $\tilde{\mu}_2 = \text{minUB}$, then $\tilde{\mu}_2 = \tilde{\mu}_3 = \cdots = \tilde{\mu}_K$ and thus all confidence intervals share a common point equal to $\tilde{\mu}_2$ (recall that $\tilde{\mu}_2 \in [\underline{\mu}_1, \overline{\mu}_1]$), which is a contradiction with maxLB > minUB.

□

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

- If at time t one has $\tilde{\boldsymbol{w}}(t) = (1/K, \ldots, 1/K)$ 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 and Lemma 28 gives

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

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

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

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 25:

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

F.2 Almost sure asymptotic bound

Theorem 30 (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*

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

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 31. *For any choice of parameters and $\boldsymbol{\mu} \in \mathcal{G}^*$, the sampling rule of* Exploration-Biased Sampling *satisfies:*

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

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

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

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

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

so that we also have

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

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

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

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

$$\begin{aligned}
\left| \frac{N_a(t)}{t} - w_a(\boldsymbol{\mu}) \right| &\leq \frac{1}{t} \left| N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s) \right| + \left| \frac{1}{t} \sum_{s=0}^{t-1} (\tilde{w}_a(s) - w_a(\boldsymbol{\mu})) \right| \\
&\leq \frac{K-1}{t} + \left| \frac{1}{t} \sum_{s=0}^{t-1} (\tilde{w}_a(s) - w_a(\boldsymbol{\mu})) \right| \qquad \text{by Lemma 25} \\
&\to_{t \to +\infty} 0
\end{aligned}$$

(using the Cesaro Lemma for the second term). □

F.3 Proof of Theorem 6

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 shown in this section are satisfied.

Let us recall the notations of Garivier and Kaufmann (2016). We assume that 1 is the best arm of $\boldsymbol{\mu}$. Fix $\varepsilon > 0$. By continuity of $\boldsymbol{w}$ in $\boldsymbol{\mu}$, let $\xi \leq \Delta_{\min}(\boldsymbol{\mu})/4$ be such that

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

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 32. *There exist two positive constants B, C (that depend on $\boldsymbol{\mu}$ and ε) 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

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\mu}}(|\hat{\mu}_a(t) - \mu_a| > \xi) &= \sum_{s=\frac{2}{K}\sqrt{t}-K}^{t} \mathbb{P}_{\boldsymbol{\mu}}(|\hat{\mu}_a(t) - \mu_a| > \xi \ \cap \ N_a(t) = s) && \text{by Lemma 3} \\
&\leq \sum_{s=\frac{2}{K}\sqrt{t}-K}^{t} \mathbb{P}(|\hat{\mu}_{a,s} - \mu_a| > \xi) && \text{by Equation (19)} \\
&\leq 2 \sum_{s=\frac{2}{K}\sqrt{t}-K}^{t} \exp\Big(-s\frac{\xi^2}{2}\Big) && \text{by Hoeffding's inequality} \\
&\leq 2\frac{\exp(-(\frac{2}{K}\sqrt{t}-K)\xi^2/2)}{1-\exp(-\xi^2/2)} \ .
\end{aligned}$$

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 33. *There exists a constant T_ε such that for $T \geq T_\varepsilon$, il holds that on $\mathcal{E}_T$*

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

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

$$\begin{aligned}
\Big|\frac{N_a(t)}{t} - w_a(\boldsymbol{\mu})\Big| &\leq \frac{1}{t}\Big|N_a(t) - \sum_{s=0}^{t-1}\tilde{w}_a(s)\Big| + \Big|\frac{1}{t}\sum_{s=0}^{t-1}(\tilde{w}_a(s) - w_a(\boldsymbol{\mu}))\Big| \\
&\leq \frac{K-1}{t} + \frac{h(T)}{t} + \Big|\frac{1}{t}\sum_{s=h(T)}^{t-1}(\tilde{w}_a(s) - w_a(\boldsymbol{\mu}))\Big| && \text{by Lemma 25} \\
&\leq \frac{K-1}{T^{1/2}} + \frac{1}{T^{1/4}} + \varepsilon && \text{by definition of } \mathcal{E}_T \\
&\leq \frac{K}{T^{1/4}} + \varepsilon \leq 3\varepsilon
\end{aligned}$$

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

G Additional experiments

In this section we present numerical experiments to compare the dependence on parameter δ of three strategies, namely Exploration-Biased Sampling, Track-and-Stop and Uniform Sampling (that samples arms uniformly).

On Figure 5, we plot for each strategy and several bandit parameters the estimate of $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ for different values of δ (using the same threshold β as in the experiments of Section 4 and $\gamma = 0.1$ for Exploration-Biased Sampling). We also plot in black the lower bound of Garivier and Kaufmann (2016) ($\sim_{\delta\to 0} T(\boldsymbol{\mu})\log(1/\delta)$).

In term of performance, we observe that Exploration-Biased Sampling is always between Uniform Sampling and Track-and-Stop (which is always quite close to the lower bound). More precisely there are different behaviours:

- when the problem is difficult (with small gaps), Exploration-Biased Sampling behaves almost like Track-and-Stop. Indeed for those parameters the uniform sampling phase of Exploration-Biased Sampling is relatively small comparing to the required number of samples so that Exploration-Biased Sampling has time to shrink its confidence regions close to parameter $\boldsymbol{\mu}$ and thus behaves like Track-and-Stop (see bandit $\boldsymbol{\mu}^{(1)}$),
- when the problem is easier (with large gaps), Exploration-Biased Sampling behaves like Uniform Sampling, as in almost all simulations the strategy does not have enough confidence to leave the uniform sampling phase before the stopping condition is satisfied (see bandits $\boldsymbol{\mu}^{(2)}$ and $\boldsymbol{\mu}^{(3)}$). When δ decreases, there is a separation between Exploration-Biased Sampling and Uniform Sampling as more and more simulations reach the non-uniform sampling phase of our strategy. If we continue to check for smaller values of δ, one can expect that Exploration-Biased Sampling will come closer to Track-and-Stop than Uniform Sampling, for the same reasons as before: the confidence regions of Exploration-Biased Sampling have more time to shrink. This is what we observe we bandit $\boldsymbol{\mu}^{(4)}$, for which Exploration-Biased Sampling has the behaviour of Uniform Sampling for moderate values of δ and then the behaviour of Track-and-Stop for small values of δ.

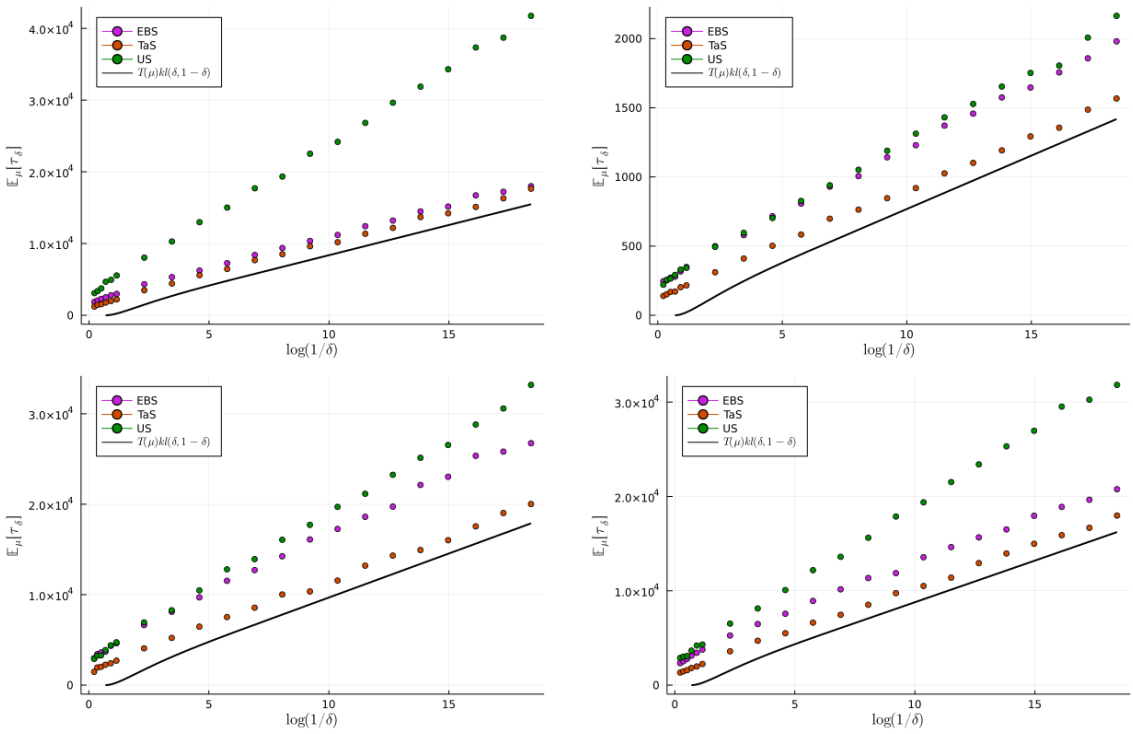

Figure 5: Empirical Expected Number of Draws $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$, Averaged over 500 Experiments. Top left: $\boldsymbol{\mu}^{(1)} = (0.9, 0.8, 0.6, 0.4, 0.4)$. Top Right: $\boldsymbol{\mu}^{(2)} = (0.9, 0.5, 0.45, 0.4)$. Bottom Left: $\boldsymbol{\mu}^{(3)} = (0.9, 0.8, 0.75, 0.7)$. Bottom Right: $\boldsymbol{\mu}^{(4)} = (0.9, 0.8, 0.7, 0.6)$