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

Antoine Barrier[1,2], Aurélien Garivier[1], Tomáš Kocák[3]
[1] ENS Lyon, [2] Université Paris-Saclay, [3] University of Potsdam

Standard Gaussian Muti-Armed Bandit

A **standard Gaussian multi-armed bandit problem** is a collection of $K \geq 2$ unit Gaussian distributions $(\mathcal{N}(\mu_a, 1))_{a\in[K]}$ indexed by a set of actions $[K] \triangleq \{1, \dots, K\}$ called arms
→ **the bandit problem is characterised by its mean vector**

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^\intercal$$

In the following we consider bandit problems with **means** $\boldsymbol{\mu} \in [0,1]^K$ and having a **unique optimal arm**, denoted by $a^*(\boldsymbol{\mu})$, such that

$$\mu_{a^*(\boldsymbol{\mu})} > \max_{a\in[K]\setminus\{a^*(\boldsymbol{\mu})\}} \mu_a$$

A **learner** interacts sequentially with an *unknown* bandit problem $\boldsymbol{\mu}$. At each round $t \in \mathbb{N}^*$, he

- picks an action $A_t \in [K]$ depending on past observations
- obtains a reward from distribution $\mathcal{N}(\mu_{A_t}, 1)$

Best-Arm Identification with fixed confidence

The **strategy** of the learner consists of

- a sampling strategy that chooses the next action A_t
- a stopping rule τ and a decision rule $\hat{a}_\tau$

The **goal of Best-Arm Identification** (BAI) is

- to find strategies that identify the best action $a^*(\boldsymbol{\mu})$ with probability at least $(1-\delta)$ for any $\boldsymbol{\mu}$, where $\delta \in (0,1)$ is a confidence level, that is

$$\mathbb{P}_{\boldsymbol{\mu}}(\hat{a}_{\tau_\delta} \neq a^*(\boldsymbol{\mu})) \leq \delta$$

→ **such strategies are called δ-correct**

- among all δ-correct strategies, **find one that minimizes the expected number of observations $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$**

Lower bound for BAI [1]

Let $\mathrm{Alt}(\boldsymbol{\mu}) \triangleq \{\boldsymbol{\lambda} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$ be the set of bandit problems which have a different best arm than $a^*(\boldsymbol{\mu})$ and $\Delta_K \triangleq \{\boldsymbol{v} \in [0,1]^K : \sum_{a\in[K]} v_a = 1\}$

Theorem 1. *For any δ-correct strategy one has*

$$\forall \boldsymbol{\mu}, \quad \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T(\boldsymbol{\mu}) \,\mathrm{kl}(\delta, 1-\delta)$$

where

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

Asymptotically, this result yields $\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geq T(\boldsymbol{\mu})$

A δ-correct strategy for which equality holds is called **asymptotically optimal** and should approximately sample arms according to the **optimal weight vector $\boldsymbol{w}(\boldsymbol{\mu}) \in \Delta_K$** realizing the supremum in the definition of $T(\boldsymbol{\mu})$

[1] Garivier, A. and Kaufmann, E. (2016), **Optimal Best Arm Identification with Fixed Confidence**, In *29th Conference On Learning Theory (COLT)*

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.

Track-and-Stop [1]

Let $N_a(t)$ and $\hat{\mu}_a(t)$ respectively denote the number of observations and average reward of arm a after round t

Main idea **Track the current optimal weight vector $\boldsymbol{w}(\hat{\boldsymbol{\mu}}(t))$ and force some minimal exploration rate of order $\sqrt{t}$ to ensure convergence to $\boldsymbol{w}(\boldsymbol{\mu})$**

Algorithm 1: 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 ; $t \leftarrow K$
while $Z(t) \leq \beta(t,\delta)$ **do**
  $\tilde{\boldsymbol{w}}(t) \leftarrow \boldsymbol{w}(\hat{\boldsymbol{\mu}}(t))$
  if $U_t \triangleq \{a \in [K] : N_a(t) < \sqrt{t} - K/2\} \neq \varnothing$ **then**
    Choose $A_{t+1} \in \mathrm{argmin}_{a\in U_t} N_a(t)$
  else
    Choose $A_{t+1} \in \mathrm{argmin}_{a\in[K]} N_a(t) - \sum_{s=K}^{t-1} \tilde{w}_a(s)$
  Observe $Y_{A_{t+1}}$ and increase t by 1
$\tau_\delta \leftarrow t$; $\hat{a}_{\tau_\delta} \leftarrow \mathrm{argmax}_{a\in[K]} \hat{\mu}_a(t)$

Pros and cons

- ✓ δ-correct using threshold $\beta(t,\delta) = \log(Rt^\alpha/\delta)$ for some $\alpha \in [1,2]$ and constant R
- ✓ asymptotically optimal
- ✗ lack of non-asymptotic result (for fixed values of δ)
- ✗ require to force exploration at an arbitrary rate ($\sqrt{t}$ here)

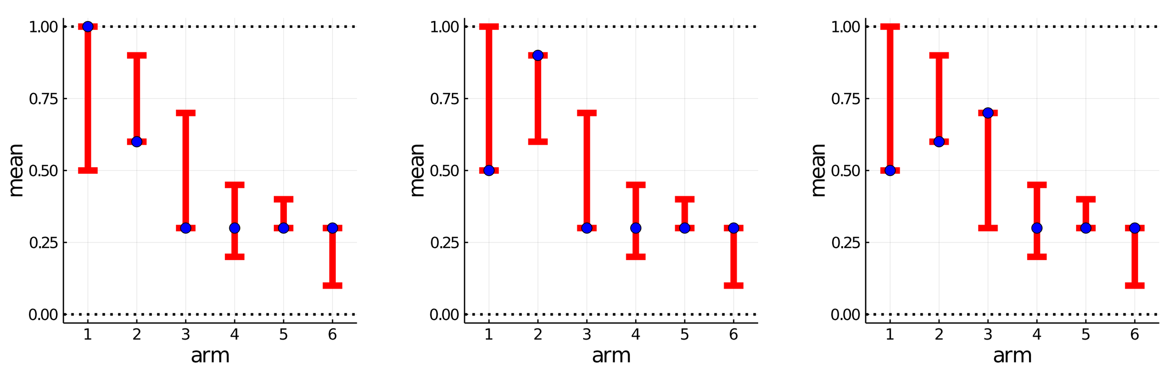
Modifying the sampling strategy

Tracking the estimate vector $\boldsymbol{w}(\hat{\boldsymbol{\mu}}(t))$ is quite hazardous: without forced exploration, a bad estimate can lead to an under-sampling of the worst arms

Improvement **Compute a confidence region $\mathcal{CR}$ for $\boldsymbol{\mu}$ around $\hat{\boldsymbol{\mu}}(t)$ and track the optimal weight associated to some bandit $\tilde{\boldsymbol{\mu}} \in \mathcal{CR}$ that maximizes exploration by satisfying**

$$\min_{a\in[K]} w_a(\tilde{\boldsymbol{\mu}}) = \max_{\boldsymbol{\nu}\in\mathcal{CR}} \min_{a\in[K]} w_a(\boldsymbol{\nu})$$

→ **this bandit $\tilde{\boldsymbol{\mu}}$ is computable**: intuitively, maximizing $w_{\min}$ over $\mathcal{CR}$ requires to increase and equalize all the positive gaps as much as possible, making the identification of the second best arm more challenging ; this principle allows to restrict the search for $\tilde{\boldsymbol{\mu}}$ to only a few candidates, one per potential best arm



The 3 candidates for the example $\mathcal{CR}$ in red

We denote by Optimistic Weights($\mathcal{CR}$) the procedure computing $\boldsymbol{w}(\tilde{\boldsymbol{\mu}})$

Exploration-Biased Sampling

Goal **Improve Track-and-Stop to obtain non-asymptotic bounds and correct the unstability behaviors**

Main idea **Use the modified sampling strategy by computing confidence regions $\mathcal{CR}_{\boldsymbol{\mu}}(t)$ for $\boldsymbol{\mu}$ at each round**

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 ; $t \leftarrow K$
while $Z(t) \leq \beta(t,\delta)$ **do**
  $\mathcal{CR}_{\boldsymbol{\mu}}(t) \leftarrow \prod_{a\in[K]} \left[\hat{\mu}_a(t) \pm 2\sqrt{\frac{\log(4N_a(t)K/\gamma)}{N_a(t)}}\right]$
  $\tilde{\boldsymbol{w}}(t) \leftarrow$ Optimistic Weights($\mathcal{CR}_{\boldsymbol{\mu}}(t)$)
  Choose $A_{t+1} \in \mathrm{argmin}_{a\in[K]} N_a(t) - \sum_{s=K}^{t-1} \tilde{w}_a(s)$
  Observe $Y_{A_{t+1}}$ and increase t by 1
$\tau_\delta \leftarrow t$; $\hat{a}_{\tau_\delta} \leftarrow \mathrm{argmax}_{a\in[K]} \hat{\mu}_a(t)$

Pros and cons

- ✓ δ-correct using same threshold as Track-and-Stop
- ✓ non-asymptotic bound with high probability

Theorem 2. *Fix $\gamma \in (0,1)$, $\eta \in (0,1]$. There exists an event $\mathcal{E}$ of probability at least $1-\gamma$ and $\delta_0 > 0$ such that for any $0 < \delta \leq \delta_0$, algorithm* Exploration-Biased Sampling *satisfies*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta \mathbb{1}_{\mathcal{E}}] \leq (1+\eta) T(\boldsymbol{\mu}) \log(1/\delta) + o_{\delta\to 0}(1)$$

(with an explicit formula for δ_0 and the $o_{\delta\to 0}(1)$)

- ✓ asymptotically optimal
- ✓ natural exploration (no need to force exploration!)
- ✗ the convergence of $\tilde{\boldsymbol{w}}(t)$ to $\boldsymbol{w}(\boldsymbol{\mu})$ is slower than Track-and-Stop

Sample Optimization Problem

We obtained new **quantitative regularity results** for the solution of the optimization problem (1) defining $T(\boldsymbol{\mu})$

Theorem 3. *Let $\boldsymbol{\mu}, \boldsymbol{\mu}'$ having the same optimal arm a^*, and assume that*

$$(1-\varepsilon)(\mu_{a^*} - \mu_a)^2 \leq (\mu'_{a^*} - \mu'_a)^2 \leq (1+\varepsilon)(\mu_{a^*} - \mu_a)^2$$

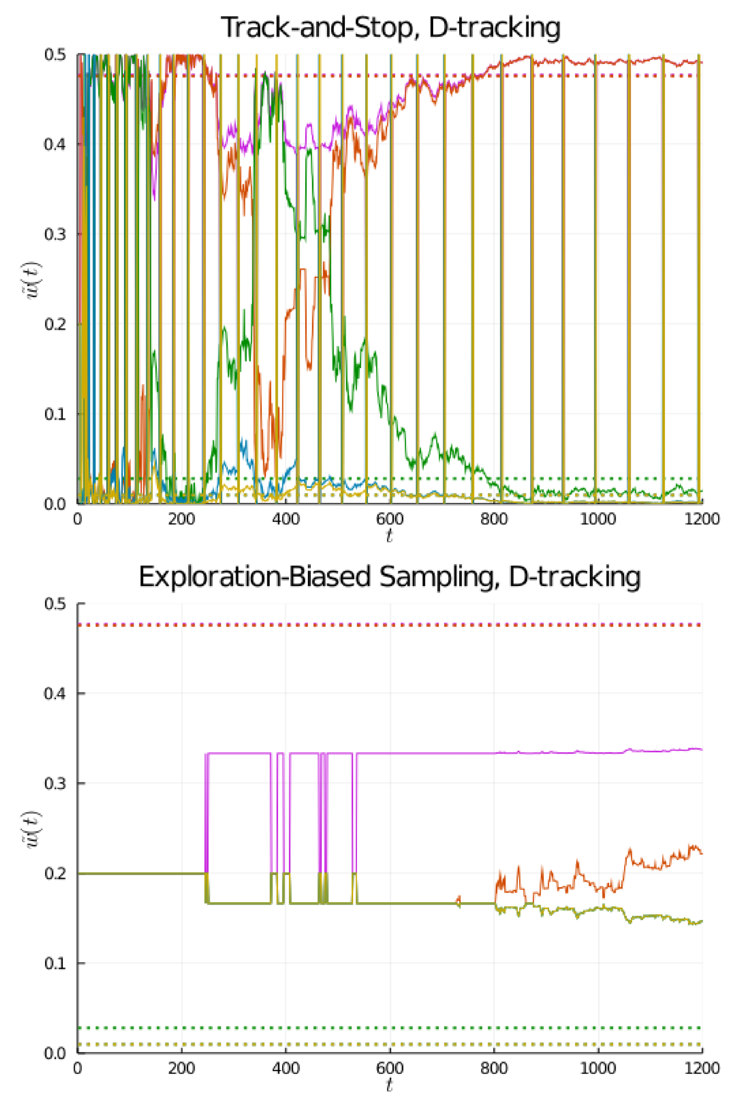
for all $a \in [K] \setminus \{a^*\}$ and some $\varepsilon \in [0, 1/7]$. Then

$$(1-3\varepsilon) T(\boldsymbol{\mu}) \leq T(\boldsymbol{\mu}') \leq (1+6\varepsilon) T(\boldsymbol{\mu})$$

and $\forall a \in [K], \quad (1-10\varepsilon) w_a(\boldsymbol{\mu}) \leq w_a(\boldsymbol{\mu}') \leq (1+10\varepsilon) w_a(\boldsymbol{\mu})$

Numerical experiments

Stability improvement



Evolution of $\tilde{\boldsymbol{w}}(t)$ when running the strategies with $\delta = 0.01$, $\gamma = 0.2$ and $\boldsymbol{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$.
The values of $\boldsymbol{w}(\boldsymbol{\mu})$ are dotted

Track-and-Stop

- ✗ unstability of the weights: red and green weights fluctuates (first estimates are poor in general, leading to unstable tracking weights, whereas intuitively one should pick arms uniformly at the beginning)
- ✗ bad arms would be under-sampled without forced exploration (blue and yellow peaks)

Exploration-Biased Sampling

- ✓ uniform weight vector during first rounds
- ✓ stability of the tracking strategy
- ✓ cautious separation of the weights when a clear distinction of the estimates appears

Efficiency The choice of the weights estimator, biased toward uniform exploration, has a price: for practical values of δ, **Exploration-Biased Sampling samples a little less the best arms than Track-and-Stop and thus requires more observations before taking a decision**