Epsilon Best Arm Identification in Spectral Bandits

Tomáš Kocák and **Aurélien Garivier**
Unité de Mathématiques Pures et Appliquées et Laboratoire de l'Informatique du Parallélisme
École Normale Supérieure de Lyon, Université de Lyon
tomas.kocak@gmail.com, aurelien.garivier@ens-lyon.fr

Abstract

We propose an analysis of Probably Approximately Correct (PAC) identification of an ε-best arm in graph bandit models with Gaussian distributions. We consider finite but potentially very large bandit models where the set of arms is endowed with a graph structure, and we assume that the arms' expectations $\boldsymbol{\mu}$ are smooth with respect to this graph. Our goal is to identify an arm whose expectation is at most ε below the largest of all means. We focus on the fixed-confidence setting: given a risk parameter δ, we consider sequential strategies that yield an ε-optimal arm with probability at least $1-\delta$. All such strategies use at least $T^*_{R,\varepsilon}(\boldsymbol{\mu})\log(1/\delta)$ samples, where R is the smoothness parameter. We identify the complexity term $T^*_{R,\varepsilon}(\boldsymbol{\mu})$ as the solution of a min-max problem for which we give a game-theoretic analysis and an approximation procedure. This procedure is the key element required by the asymptotically optimal Track-and-Stop strategy.

1 Introduction

A *bandit model* (see [Lattimore and Szepesvári, 2019] and references therein) is a set of probability distributions $\boldsymbol{\nu} = \{\nu_a : a \in \mathcal{A}\}$. These distributions are called *arms*, and the statistician can sample one of them at each time step $t \geq 1$. *Best-arm identification* (BAI) consists of using those samples so as to find which arm has the highest expectation $\mu_a = \mathbb{E}(\nu_a)$, while ε-BAI aims at identifying an arm a such that $\mu_a \geq \max_b \mu_b - \varepsilon$. A *fixed-confidence* algorithm for a given risk δ consists in a sampling rule A_t choosing thanks to past observations which arm is sampled at each time step t, and of a stopping rule τ (a stopping time): it is called δ-correct if $A_{\tau+1}$ is ε-optimal with probability at least $1-\delta$. The efficiency of this algorithm is measured by the mean number $\mathbb{E}_{\boldsymbol{\nu}}[\tau]$ of samples needed.

Since the work of [Mannor and Tsitsiklis, 2004] and [Even-Dar *et al.*, 2006], best-arm identification has received considerable interest. It has been proved that good strategies require no more than $A(\boldsymbol{\nu}) + B(\boldsymbol{\nu})\log(1/\delta)$ samples, for some functions A and B of the model that were progressively improved. While, for example, [Karnin *et al.*, 2013] investigated more on term A, other authors insisted on the fact that B is the dominant term when δ is small and focused on the best possible term B. For BAI (with $\varepsilon = 0$), the latter was identified by [Garivier and Kaufmann, 2016] and [Russo, 2016]. The first of those articles provided a generic analysis that reduces BAI to the identification of an information-theoretic complexity term that gives at the same time a lower bound on the performance of any algorithm, and a key ingredient of an asymptotically optimal strategy called Track-and-Stop. This term appears to be the solution of a min-max optimization program, for which an ad-hoc solution was given in the aforementioned article.

While these first works were limited to the δ-correct identification of the best arm (which was assumed to exist and be unique), [Garivier and Kaufmann, 2019] proposed an extension to the problem of identifying ε-optimal arms. Simultaneously, [Degenne and Koolen, 2019] leveraged the game-theoretic nature of the complexity term to encompass even more general objectives.

In parallel with this progress, vanilla bandit models have shown limitations in settings (such as recommendation systems) where the number of arms is huge. It is then not uncommon that the set of arms is naturally endowed with some *structure*. One simple way to take this structure into account is to assume the existence of some notion of *similarity*: some arms are "close" from one another, in the sense that their outcomes are expected to have similar distributions. *Graph bandits* are meant to provide a theoretical framework for this setting: the set of arms is provided with a graph structure where the weight $w_{a,b}$ of the link from arm a to arm b measures their similarity. The set of means $\boldsymbol{\mu} = (\mu_a : 1 \leq a \leq K)$ is assumed to be *smooth* with respect to this graph in the sense that

$$\|\boldsymbol{\mu}\|^2_{\mathcal{L}} \triangleq \sum_{a,\, b\in\mathcal{A}} w_{a,b}\frac{(\mu_a - \mu_b)^2}{2} = \boldsymbol{\mu}^{\mathsf{T}}\mathcal{L}\boldsymbol{\mu} \leq R\,, \qquad (1)$$

where $\mathcal{L}$ denotes the graph's Laplacian and R is some known upper-bound. This means that two arms connected by an edge with significant weight should have similar expectations.

Recently, [Kocák and Garivier, 2020] showed how to optimally, δ-correctly identify the best arm in a graph bandit model if it exists. But the consideration of very large sets of arms, and the fact that many of them might be close to optimal, suggest that it is often more relevant to identify (more quickly) any $\varepsilon-$optimal arm instead of the unique best.

In the present work, we address the problem of PAC identification of an ε-best arm in graph bandit models, hence encompassing all the aforementioned difficulties. We focus on finite bandit models with K arms: $\mathcal{A} = \{1, \ldots, K\}$. In order to avoid technicalities, we consider *Gaussian* arms $\nu_a = \mathcal{N}(\mu_a, 1)$ and identify the bandit problem with vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ of arm expectations.

Following the approach of [Garivier and Kaufmann, 2016], we start in Section 2 by identifying the complexity of the problem as the optimal ratio $\mathbb{E}_{\boldsymbol{\nu}}[\tau]/\log(1/\delta)$ of any (ε, δ)-PAC algorithm when $\delta \to 0$. This complexity term appears to be the solution of an interesting max-min progam. Taking great benefit from its game-theoretic structure, we propose a solution that is radically different from [Garivier and Kaufmann, 2019] even in the case where $R = \infty$, a case that we treat separately in Section 3. The structured case $R < \infty$, which is the main contribution of this paper, is treated in Section 4.

2 The Complexity of ε-BAI for Graph Bandits

2.1 Characteristic Time

Several factors are contributing to the complexity of identifying one of the ε-best arms in bandit problems. The difficulty is related to the quantity $T^*_{R,\varepsilon}(\boldsymbol{\mu})$ called **characteristic time**.

Definition 1. Characteristic time $T^*_{R,\varepsilon}(\boldsymbol{\mu})$ is defined as

$$T^*_{R,\varepsilon}(\boldsymbol{\mu})^{-1} \triangleq \max_{\substack{\boldsymbol{\omega}\in\Delta_K \\ i\in\mathcal{A}^*_\varepsilon(\boldsymbol{\mu})}} \min_{\substack{j\neq i \\ \boldsymbol{\lambda}\in\mathcal{M}^{i,j}_{R,\varepsilon}}} \sum_{a\in[K]} \omega_a \frac{(\mu_a-\lambda_a)^2}{2} \qquad (2)$$

where Δ_K is the K-dimensional simplex, $\mathcal{A}^*_\varepsilon(\boldsymbol{\mu})$ is the set of all ε-best arms of bandit problem $\boldsymbol{\mu}$

$$\mathcal{A}^*_\varepsilon(\boldsymbol{\mu}) \triangleq \{a \in \mathcal{A} : \mu_a \geq \max_{b\in\mathcal{A}}(\mu_b) - \varepsilon\},$$

and $\mathcal{M}^{i,j}_{R,\varepsilon}$ is a set of bandit problems with smoothness at most R and with arm j being better than arm i by at least ε margin

$$\mathcal{M}^{i,j}_{R,\varepsilon} \triangleq \{\boldsymbol{\lambda} \in \mathbb{R}^K : \boldsymbol{\lambda}^\intercal \mathcal{L}\boldsymbol{\lambda} \leq R,\ \lambda_i \leq \lambda_j - \varepsilon\}. \qquad (3)$$

$\boldsymbol{\lambda}^\intercal \mathcal{L}\boldsymbol{\lambda} \leq R$ is often called a *spectral constraint*, hence the name *spectral bandits*.

Remark. *This definition is backward compatible with previous papers. By setting R to infinity, every problem satisfies the spectral constraint and we obtain the setting of [Garivier and Kaufmann, 2019]. By setting ε to zero, we are identifying only the best arm which leads to the setting of [Kocák and Garivier, 2020]. By setting R to infinity and ε to zero at the same time we obtain the original setting of [Garivier and Kaufmann, 2016].*

The starting point for this paper is stated in the following proposition. This proposition can be obtained along the lines of a recent paper by [Garivier and Kaufmann, 2019]. It shows the connection between the expected stopping time of any δ-correct algorithm that identifies ε-best arms and the characteristic time defined previously.

Proposition 1. *For any δ-correct strategy and any bandit problem $\boldsymbol{\mu}$, the expectation of stopping time τ_δ of the strategy is lower bounded as*

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geq T^*_{R,\varepsilon}(\boldsymbol{\mu})$$

where the characteristic time $T^*_{R,\varepsilon}(\boldsymbol{\mu})$ is defined in Eq. (2).

The most significant part of this proposition is that it provides a lower bound that scales with the characteristic time. The proof is assuming that the learner plays according to strategy $\boldsymbol{\omega}$ from the definition of $T^*_{R,\varepsilon}(\boldsymbol{\mu})$, i.e. playing arm a with probability ω_a, while the environment chooses some bandit problem $\boldsymbol{\lambda}$ for the learner. By choosing the best possible strategy $\boldsymbol{\omega}$ for the learner while the environment chooses the hardest possible bandit problem $\boldsymbol{\lambda}$ we obtain the lower bound from the proposition.

However, this works also the other way around. If the learner plays according to the optimal strategy $\boldsymbol{\omega}^*$, the strategy that maximizes expression in the definition of $T^*_{R,\varepsilon}(\boldsymbol{\mu})$, the expected stopping time of the learner is also proportional to the characteristic time and therefore matching the lower bound (possibly up to some multiplicative constant). Therefore, the main focus of this paper is on analyzing the characteristic time and finding a way to compute optimal weight $\boldsymbol{\omega}^*$ for the learner and provide an algorithm that utilizes these weights.

2.2 Game-Theoretical Point of View

As we hinted in the previous section, computing the inverse of characteristic time, $T^*_{R,\varepsilon}(\boldsymbol{\mu})^{-1}$, can be seen as a game between the learner and the environment where:

- The first player (learner) chooses one of the ε-best arms $i \in \mathcal{A}^*_\varepsilon(\boldsymbol{\mu})$ and $\boldsymbol{\omega} \in \Delta_K$ while trying to maximize the value of the optimization problem.
- The second player (environment) chooses alternative arm $j \neq i$ and bandit problem $\boldsymbol{\lambda} \in \mathcal{M}^{i,j}_{R,\varepsilon}$ while trying to minimize the value of the optimization problem.

In fact, the optimization function in the definition of $T^*_{R,\varepsilon}(\boldsymbol{\mu})^{-1}$ is very simple; linear in $\boldsymbol{\omega}$ and quadratic in $\boldsymbol{\lambda}$. To make it more apparent, we use the following definition and rewrite the optimization problem in a slightly different way.

Definition 2. Let $\mathcal{M}^{i,j}_{R,\varepsilon}$ be the set of the problems with smoothness at most R and arm j being better than arm i by at least ε (expression (3)). Define the set of elementwise divergences from $\boldsymbol{\mu}$ to $\mathcal{M}^{i,j}_{R,\varepsilon}$ as

$$\mathcal{D}^{i,j}_{R,\varepsilon} \triangleq \left\{ \boldsymbol{d} \in \mathbb{R}^K : \exists \boldsymbol{\lambda} \in \mathcal{M}^{i,j}_{R,\varepsilon} \text{ s.t. } d_a \triangleq \frac{(\mu_a-\lambda_a)^2}{2} \right\}$$

This definition enables us to rewrite $T^*_{R,\varepsilon}(\boldsymbol{\mu})^{-1}$ in a more compact way as

$$T^*_{R,\varepsilon}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega}\in\Delta_K} \max_{i\in\mathcal{A}^*_\varepsilon(\boldsymbol{\mu})} \min_{j\neq i} \inf_{\boldsymbol{d}\in\mathcal{D}^{i,j}_{R,\varepsilon}} \boldsymbol{\omega}^\intercal \boldsymbol{d} \qquad (4)$$

Thanks to this reparametrization we obtained an optimization problem that is linear with independent $\boldsymbol{\omega}$ and $\boldsymbol{d}$. We use this form later to simplify the presentation of some ideas and proofs. We approach this optimization problem in two steps:

1. Given i and $\boldsymbol{\omega}$ of the first player (learner), we compute **the best response $\boldsymbol{d}$** (resp. $\boldsymbol{\lambda}$) of the second player
2. Having the best response, we can find optimal $\boldsymbol{\omega}^*$ (either directly or numerically, depending on the problem)

In any case, it is important to be able to compute the best response of the second player. When finding the best response, usually it is helpful to split the optimization problem into several smaller problems and solve them separately. We do it by considering only a subproblem with fixed i.

Definition 3. By fixing i in $T^*_{R,\,\varepsilon}(\boldsymbol{\mu})^{-1}$ we can define

$$T^i_{R,\,\varepsilon}(\boldsymbol{\mu})^{-1} \triangleq \sup_{\boldsymbol{\omega}\in\Delta_K} \min_{j\neq i} \inf_{\boldsymbol{d}\in\mathcal{D}^{i,\,j}_{R,\,\varepsilon}} \boldsymbol{\omega}^\intercal \boldsymbol{d}$$

This definition enables us to compute $T^*_{R,\,\varepsilon}(\boldsymbol{\mu})^{-1}$ as

$$T^*_{R,\,\varepsilon}(\boldsymbol{\mu})^{-1} = \max_{i\in\mathcal{A}^*_\varepsilon(\boldsymbol{\mu})} T^i_{R,\,\varepsilon}(\boldsymbol{\mu})^{-1}.$$

3 BAI Problem Without Structure

In this section, we focus on the setting without structure. This setting was previously studied by [Garivier and Kaufmann, 2019] and can be obtained simply by setting the smoothness parameter R to ∞ which would make any bandit problem satisfy the smoothness constraint (Expression (1)). The main contribution of this section is to significantly simplify the proofs of [Garivier and Kaufmann, 2019] by using a previously mentioned game-theoretical approach while providing the necessary ideas and reasoning later used in the more difficult spectral case. As mentioned earlier, this game-theoretic approach was initiated in [Degenne and Koolen, 2019].

Finding values of individual $T^i_{\infty,\,\varepsilon}(\boldsymbol{\mu})^{-1}$ is reminiscent of the problems solved by [Garivier and Kaufmann, 2016] and [Kocák and Garivier, 2020]. However, this time we assume that i is not necessarily the optimal arm but it is at most ε away from the optimal arm. The following theorem shows the main result of the unconstrained case and a convenient way of computing optimal weights $\boldsymbol{\omega}^*(\boldsymbol{\mu})$.

Theorem 2. *Assume that i is one of the ε optimal arms of bandit problem $\boldsymbol{\mu}$, i.e. $\mu_i > \mu_j - \varepsilon$ for every $j \in [K]$. Let I be any arm different from i and define sequence $\{x_a(c)\}_{a\in[K]/\{i\}}$ as*

$$x_I(c) = c$$
$$x_j(c) = \left[\left(1 + x_I(c)^{-1}\right)\frac{\delta^{i,\,j}_\varepsilon}{\delta^{i,\,I}_\varepsilon} - 1\right]^{-1}$$

for any $j \in [K]/\{i,\, I\}$, constant c, and $\delta^{i,\,j}_\varepsilon = (\mu_i - \mu_j + \varepsilon)$. Let $f(c)$ be a function with parameter c defined as

$$f(c) = \sum_{j\in[K]/\{i\}} x_j(c)^2.$$

Then there exist $c^* \in \mathbb{R}^+$ such that $f(c^*) = 1$ and we obtain optimal $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ as

$$\omega^*_i(\boldsymbol{\mu}) = \frac{1}{1 + \sum_{j\in[K]/\{i\}} x_j(c^*)}$$
$$\omega^*_j(\boldsymbol{\mu}) = x_j(c^*)\,\omega^*_i(\boldsymbol{\mu}) \qquad \textit{for } j \in [K]/\{i\}$$

The rest of this section is dedicated to building necessary tools for the proof of Theorem 2 later in Section 3.2.

3.1 Best Response Oracle - Setting Without Constraint

The following lemma shows us the form of the best response of Player 2 in $T^i_{\infty,\,\varepsilon}(\boldsymbol{\mu})^{-1}$ game.

Lemma 3. *Let $\boldsymbol{\omega}$ be a vector from Δ_K and $\boldsymbol{\mu}$ a bandit problem. Then the best response $\boldsymbol{\lambda}^{i,\,j}_{\infty,\,\varepsilon}(\boldsymbol{\omega}) \in \mathcal{M}^{i,\,j}_{\infty,\,\varepsilon}$ to $\boldsymbol{\omega}$, with arm j being better than arm i by at least ε, is in form*

$$\boldsymbol{\lambda}^{i,\,j}_{\infty,\,\varepsilon}(\boldsymbol{\omega}) = (\mu_1,\,\ldots,\,\mu_{i-1},\,\boldsymbol{t},\,\ldots,\,\boldsymbol{t}+\varepsilon,\,\mu_{j+1},\,\ldots,\,\mu_K)^\intercal$$

for

$$t = \frac{\mu_i\omega_i + \mu_j\omega_j}{\omega_i + \omega_j} - \varepsilon\left(\frac{\omega_j}{\omega_i + \omega_j}\right).$$

Proof. Assuming that j-th position of $\boldsymbol{\lambda}^{i,\,j}_{\infty,\,\varepsilon}(\boldsymbol{\omega})$ is exactly $\overline{\varepsilon}$ larger than i-th position of $\boldsymbol{\lambda}^{i,\,j}_{\infty,\,\varepsilon}(\boldsymbol{\omega})$, for some $\overline{\varepsilon} \geq \varepsilon$. Using simple calculus we obtain that

$$t = \frac{\mu_i\omega_i + \mu_j\omega_j}{\omega_i + \omega_j} - \overline{\varepsilon}\left(\frac{\omega_j}{\omega_i + \omega_j}\right)$$

while the element on the j-th position has value

$$\frac{\mu_i\omega_i + \mu_j\omega_j}{\omega_i + \omega_j} + \overline{\varepsilon}\left(\frac{\omega_i}{\omega_i + \omega_j}\right).$$

The first part of both expressions is a weighted average of μ_i and μ_j which is always in interval $[\mu_j,\,\mu_i]$ and therefore, by increasing $\overline{\varepsilon}$ we increase our objective function. This makes $\overline{\varepsilon} = \varepsilon$ the optimal choice. □

Corollary 1. *Let $\boldsymbol{\omega}$ be a vector from Δ_K and $\boldsymbol{\mu}$ a bandit problem. Then the best response $\boldsymbol{d}^{i,\,j}_{\infty,\,\varepsilon}(\boldsymbol{\omega}) \in \mathcal{D}^{i,\,j}_{\infty,\,\varepsilon}$ to $\boldsymbol{\omega}$, with arm j being better than arm i by at least ε, is zero everywhere except for the i-th and j-th position*

$$\boldsymbol{d}^{i,\,j}_{\infty,\,\varepsilon}(\boldsymbol{\omega}) = (0,\,\ldots,\,0,\,\underbrace{\frac{\overline{\omega_j}^2\delta^{i,\,j}_\varepsilon}{2}}_{\text{position } i},\,\ldots,\,\underbrace{\frac{\overline{\omega_i}^2\delta^{i,\,j}_\varepsilon}{2}}_{\text{position } j},\,0,\,\ldots,\,0)^\intercal$$

for

$$\overline{\omega_j} = \frac{\omega_j}{\omega_i + \omega_j},\quad \overline{\omega_i} = \frac{\omega_i}{\omega_i + \omega_j},\quad \delta^{i,\,j}_\varepsilon = (\mu_i - \mu_j + \varepsilon)^2\,.$$

Proof. The proof is obtained by combining Lemma 3 with the definition of $\mathcal{D}^{i,\,j}_{\infty,\,\varepsilon}$. □

3.2 Proof of Theorem 2

Now that we know the exact form of the best response provided by the oracle, we are ready to prove the statement of Theorem 2 for the non-spectral setting.

Using Corollary 1 problem $T^i_{\infty,\,\varepsilon}(\boldsymbol{\mu})^{-1}$ transforms into

$$T^i_{\infty,\,\varepsilon}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega}\in\Delta_K} \min_{j\neq i} \boldsymbol{\omega}^\intercal \boldsymbol{d}^{i,\,j}_{\infty,\,\varepsilon}(\boldsymbol{\omega})\,,$$

where Player 2 now chooses only $j \neq i$. The following lemma shows that Player 2 can play a mixed strategy (a convex combination of pure strategies where the player plays only one arm j) while not changing the value of the game.

Lemma 4. *Let $\boldsymbol{\omega}$ be a vector in $\mathbb{R}^K$ and $\mathcal{D}$ be a compact subset of $\mathbb{R}^K$ then*

$$\inf_{\boldsymbol{d}\in\mathcal{D}} \boldsymbol{\omega}^\intercal \boldsymbol{d} = \inf_{\boldsymbol{d}\in Conv(\mathcal{D})} \boldsymbol{\omega}^\intercal \boldsymbol{d}\ ,$$

where $Conv(\mathcal{D})$ is the convex hull of $\mathcal{D}$.

Proof. Since $\mathcal{D}$ is a compact set, there exist vector $\boldsymbol{d}^*$ such that

$$\boldsymbol{\omega}^\intercal \boldsymbol{d}^* = \inf_{\boldsymbol{d}\in Conv(\mathcal{D})} \boldsymbol{\omega}^\intercal \boldsymbol{d}\ .$$

We also know that $\boldsymbol{d}^*$ is a vector from the convex hull of $\mathcal{D}$ therefore, $\boldsymbol{d}^*$ can be expressed as a convex combination $\sum_{j\in[K]} q_j \boldsymbol{d}^j$ for $\boldsymbol{q} = (q_1, \ldots, q_K) \in \Delta_K$ of at most K points $\boldsymbol{d}^j \in \mathcal{D}$. Therefore, we have

$$\boldsymbol{\omega}^\intercal \boldsymbol{d}^* = \boldsymbol{\omega}^\intercal \sum_{j\in[K]} q_j \boldsymbol{d}^j \geq \sum_{j\in[K]} q_j \boldsymbol{\omega}^\intercal \boldsymbol{d}^* = \boldsymbol{\omega}^\intercal \boldsymbol{d}^*\ .$$

The inequality in the previous expression holds since $\boldsymbol{d}^*$ is the minimizer of the expression from the lemma statement. However, the first term in the previous expression is the same as the last term and therefore the inequality should achieve equality. This can occur only if $\boldsymbol{\omega}^\intercal \boldsymbol{d}^j = \boldsymbol{\omega}^\intercal \boldsymbol{d}^*$ for every j where $q_j \neq 0$. Since at least one q_j is strictly positive, corresponding $\boldsymbol{d}^j$ is one of the minimizers of the minimization problem over $\mathcal{D}$. □

As we mentioned, Lemma 4 allows Player 2 to play a mixed strategy while the value of the game stays the same. The main benefit of this change is that the game now has a Nash equilibrium. In order to be in the equilibrium, any change to the strategy of the first player should result in same value of the game as long as the second player plays the optimal mixed strategy. This also means that in the mixed strategy of the second player, all the elements of played vector should be the same. Therefore, there has to be a convex combination with coefficients $\{q_j \geq 0\}_{j\in[K]/\{i\}}$ such that

$$\boldsymbol{d}^{i,*}_{\infty,\varepsilon}(\boldsymbol{\omega}) = \sum_{\substack{j\in[K]\\ j\neq i}} q_j \boldsymbol{d}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega}) \quad \text{and} \quad \boldsymbol{d}^{i,*}_{\infty,\varepsilon}(\boldsymbol{\omega}) = r\mathbb{1} \tag{5}$$

for some constant r. Since for any $j \neq i$ only $\boldsymbol{d}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})$ is not 0 at position j, value of q_j can not be 0 and

$$q_j \boldsymbol{d}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega}) = q_k \boldsymbol{d}^{i,k}_{\infty,\varepsilon}(\boldsymbol{\omega}) \quad \text{for any } j,k \in [K]/\{i\}\ . \tag{6}$$

Now we have everything necessary to find the convex combination for a given $\boldsymbol{\omega}$ as well as the way to find optimal $\boldsymbol{\omega}^*$ for the first player.

3.3 Finding Weights q_j and Optimal $\boldsymbol{\omega}^*$

Instead of finding the convex combination we can look for a linear combination that gives us $\mathbb{1}$ and then renormalize it to obtain a convex combination. Since $\boldsymbol{d}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})$ is the only member contributing to the j-th element, we can divide it by $d^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})_j$ to obtain 1 at the j-th position. Therefore, q_j is proportional to $d^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})_j^{-1}$. Now we have a vector of ones except for the i-th element for which we do not have any guarantees. In order to be in Nash equilibrium, i-th element should be 1 as well. That means that

$$d^{i,*}_{\infty,\varepsilon}(\boldsymbol{\omega})_i = \sum_{\substack{j\in[K]\\ j\neq i}} \frac{d^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})_i}{d^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})_j} = \sum_{\substack{j\in[K]\\ j\neq i}} \left(\frac{\omega_j}{\omega_i}\right)^2 = 1\ . \tag{7}$$

From Lemma 4 and the fact that both $\boldsymbol{d}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})$ and $\boldsymbol{d}^{i,k}_{\infty,\varepsilon}(\boldsymbol{\omega})$ contribute to the optimal mixed strategy, they need to be equally good when the first player chooses optimal $\boldsymbol{\omega}^*$. Therefore, we have

$$\omega^*_i d^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega}^*)_i + \omega^*_j d^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega}^*)_j = \omega^*_i d^{i,k}_{\infty,\varepsilon}(\boldsymbol{\omega}^*)_i + \omega^*_k d^{i,k}_{\infty,\varepsilon}(\boldsymbol{\omega}^*)_k$$

which, after a few steps, leads to

$$x_j(c^*) = \left[\left(1 + x_k(c^*)^{-1}\right)\ , \frac{\delta^{i,j}_\varepsilon}{\delta^{i,k}_\varepsilon} - 1\right]^{-1}\ .$$

This provides us a way to express $x_j(c^*)$ using $x_k(c^*)$ for any combination of arms j and k. In particular, setting $k = I$ we recover Theorem 2

4 BAI Problem with Structure

The main complexity of spectral setting comes from the fact that the best response for the second player of game $T^i_{R,\varepsilon}(\boldsymbol{\mu})^{-1}$ does not have a closed form and therefore, it is impossible to compute $\boldsymbol{\omega}^*$ directly as in the case without structure. We solve this problem in several steps:

- Computing best response to $\boldsymbol{\omega}$ numerically.
- Restating $T^i_{R,\varepsilon}(\boldsymbol{\mu})^{-1}$ as a function of $\boldsymbol{\omega}$.
- Computing a supergradient for this function.
- Applying a gradient algorithm to compute optimal $\boldsymbol{\omega}^*$.

4.1 Best Response Oracle - Spectral Setting

The oracle needs to find $\boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega})$ that minimizes

$$\inf_{\boldsymbol{\lambda}\in\mathcal{M}^{i,j}_{R,\varepsilon}} \sum_{a\in[K]} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}\ .$$

In the case where the oracle for the setting without structure returns $\boldsymbol{\lambda}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})$ that satisfies spectral constraint ($\boldsymbol{\lambda}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})^\intercal \mathcal{L} \boldsymbol{\lambda}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega}) \leq R$), we are done and the spectral oracle should return value $\boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega}) = \boldsymbol{\lambda}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})$. On the other hand, we can restrict our search for the problems $\boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega})$ with smoothness exactly R, thanks to the following lemma.

Lemma 5. *Let $\boldsymbol{\lambda}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})$ be the response of non-spectral oracle such that $\boldsymbol{\lambda}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega})^\intercal \mathcal{L} \boldsymbol{\lambda}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega}) > R$ then the response of spectral oracle $\boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega})$ satisfies $\boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega})^\intercal \mathcal{L} \boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega}) = R$.*

Proof idea. Suppose that the smoothness of $\boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega})$ is smaller than R, i.e. $\boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega})^\intercal \mathcal{L} \boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega}) < R$. Define $\boldsymbol{\lambda}$ as a convex combination of non-spectral and spectral oracle responses with parameter $\alpha \in (0, 1)$.

$$\boldsymbol{\lambda} = \alpha \boldsymbol{\lambda}^{i,j}_{R,\varepsilon}(\boldsymbol{\omega}) + (1-\alpha) \boldsymbol{\lambda}^{i,j}_{\infty,\varepsilon}(\boldsymbol{\omega}).$$

For small enough α, we can show that $\boldsymbol{\lambda}$ improves the optimization problem while still satisfying the spectral constraint. □

Knowing that the smoothness of the oracle response is exactly R, we use the Lagrange multiplier method to solve the problem

$$F(\boldsymbol{\lambda}, \gamma) \triangleq \sum_{a\in[K]} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} + \gamma(\boldsymbol{\lambda}^\intercal \mathcal{L} \boldsymbol{\lambda} - R) \; ,$$

where $\lambda_i = \lambda_j - \varepsilon$ and γ is the Lagrange multiplier. Before solving this problem we need to eliminate coupling between i-th and j-th elements of λ and introduce some notation to simplify some of the expressions in the next lemma. Lets fix indices i and j for the moment and define:

Any tilde index $\tilde{a}$ is equal to a if $a < j$ or $a - 1$ if $a > j$. This means that by removing element on the j-position of some vector, we can refer to the original element on the a-th position using $\tilde{a}$.

- $\widetilde{\boldsymbol{\omega}}$ is a vector such that

$$\begin{aligned} \widetilde{\omega}_{\tilde{a}} &= \omega_a \qquad \text{for all} \qquad a \in [K]/\{i,\, j\} \\ \widetilde{\omega}_{\tilde{i}} &= \omega_i + \omega_j \end{aligned}$$

- $\widetilde{\boldsymbol{\Omega}}$ is a diagonal matrix with $\widetilde{\boldsymbol{\omega}}$ on diagonal.
- $\widetilde{\boldsymbol{\lambda}}$ is a vector such that

$$\widetilde{\lambda}_{\tilde{a}} = \lambda_a \qquad \text{for all} \qquad a \in [K]/\{j\}$$

- $\widetilde{\boldsymbol{\mu}}$ is a vector such that

$$\begin{aligned} \widetilde{\mu}_{\tilde{a}} &= \mu_a \qquad \text{for all} \qquad a \in [K]/\{i,\, j\} \\ \widetilde{\mu}_{\tilde{i}} &= \frac{\omega_i\mu_i + \omega_j\mu_j}{\omega_i + \omega_j} - \frac{\varepsilon\omega_j}{\omega_i + \omega_j} \end{aligned}$$

- $\widetilde{\mathcal{L}}$ is a matrix created from $\mathcal{L}$ by adding j-th row and column to i-th row and column and updating diagonal entries to have a zero sum on every row and column and then removing t-th row and column from the matrix.

$$\begin{aligned} \widetilde{\mathcal{L}}_{\tilde{a},\, \tilde{b}} &= \mathcal{L}_{a,\, b} \qquad \text{for all} \qquad a,\, b \in [K]/\{i,\, j\} \\ \widetilde{\mathcal{L}}_{\tilde{i},\, \tilde{a}} &= \widetilde{\mathcal{L}}_{\tilde{a},\, \tilde{i}} = \mathcal{L}_{i,\, a} + \mathcal{L}_{j,\, a} \quad \text{for all} \quad a \in [K]/\{i,\, j\} \\ \widetilde{\mathcal{L}}_{\tilde{i},\, \tilde{i}} &= \sum_{a\in[K]/\{i,\, j\}} -\widetilde{\mathcal{L}}_{i,\, \tilde{a}} \end{aligned}$$

- $\widetilde{\mathcal{L}}_j$ is a vector created from the j-th column of $\mathcal{L}$ by setting i-th element to 0, updating j-th element to have zero sum of elements, and removing i-th element.

Lemma 6. *Let i and j are fixed and $\widetilde{\mathcal{L}}_j$ be the j-th column of $\widetilde{\mathcal{L}}$. Then we define $\widetilde{\boldsymbol{\lambda}}(\gamma)$ as*

$$\widetilde{\boldsymbol{\lambda}}(\gamma) \triangleq \big(\widetilde{\boldsymbol{\Omega}} + 2\gamma\widetilde{\mathcal{L}}\big)^{-1}\big(\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\mu}} + 2\gamma\varepsilon\widetilde{\mathcal{L}}_j\big)$$

There exists γ^* such that $\boldsymbol{\lambda}(\gamma^*)^\intercal \mathcal{L} \boldsymbol{\lambda}(\gamma^*) = R$ and $\boldsymbol{\lambda}(\gamma^*)$ is the best response vector that corresponds to $\widetilde{\boldsymbol{\lambda}}(\gamma^*)$ such that element on the j-th position is larger than the element on the i-th position by exactly ε.

Proof idea. The statement of the lemma can be obtain taking partial derivatives of $F(\boldsymbol{\lambda},\, \gamma)$ and solving the resulting system of equations. □

4.2 Supergradient as the Best Response

Now that we have a way to compute the best response to Player 1, we are ready to restate $T^i_{R,\,\varepsilon}(\boldsymbol{\mu})^{-1}$ as a function of $\boldsymbol{\omega}$ and provide a lemma that shows the form of a supergradient for this function. Define $f^i(\boldsymbol{\omega})$ as

$$f^i(\boldsymbol{\omega}) \triangleq \min_{j\neq i} \inf_{\boldsymbol{d}\in\mathcal{D}^{i,\,j}_{R,\,\varepsilon}} \boldsymbol{\omega}^\intercal \boldsymbol{d} \; .$$

Note that $T^i_{R,\,\varepsilon}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega}\in\Delta_K} f^i(\boldsymbol{\omega})$. The following lemma gives us a convenient way to compute a supergradient of f^i at $\boldsymbol{\omega}$. In fact, the best response $\boldsymbol{d}^{i,\,j}(\boldsymbol{\omega})$, computed by the best response oracle from Section 4.1, is a supergradient of f^i thanks to the following lemma.

Lemma 7. *Let $\mathcal{D} \subseteq \mathbb{R}^K$ be a compact set. Then function $f : \Delta_K \to \mathbb{R}$ defined as $f(\boldsymbol{\omega}) = \inf_{\boldsymbol{d}\in\mathcal{D}} \boldsymbol{\omega}^\intercal \boldsymbol{d}$ is a concave function and $\boldsymbol{d}^*(\boldsymbol{\omega}) = \arg\min_{\boldsymbol{d}\in\mathcal{D}} \boldsymbol{\omega}^\intercal \boldsymbol{d}$ is a supergradient of f at $\boldsymbol{\omega}$.*

Proof. Let $\boldsymbol{d}^*(\boldsymbol{\omega}) \in \mathcal{D}$ be a vector that realizes the infimum from the definition of $f(\boldsymbol{\omega})$. Such a vector is well defined since $\mathcal{D}$ is compact. First, we prove that $\boldsymbol{d}^*(\boldsymbol{\omega})$ is a supergradient of f at any point $\boldsymbol{\omega}$ since the existence of supergradient implies the concavity of the function.

Let $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ be any two points from the domain of f. From the definition of $\boldsymbol{d}^*(\boldsymbol{\omega})$ we have

$$\begin{aligned} \boldsymbol{\omega}_2^\intercal \boldsymbol{d}^*(\boldsymbol{\omega}_1) &\geq \boldsymbol{\omega}_2^\intercal \boldsymbol{d}^*(\boldsymbol{\omega}_2) \\ \boldsymbol{\omega}_1^\intercal \boldsymbol{d}^*(\boldsymbol{\omega}_1) + (\boldsymbol{\omega}_2 - \boldsymbol{\omega}_1)^\intercal \boldsymbol{d}^*(\boldsymbol{\omega}_1) &\geq \boldsymbol{\omega}_2^\intercal \boldsymbol{d}^*(\boldsymbol{\omega}_2) \end{aligned}$$

Which, using the definition of f, can be further rewritten as

$$f(\boldsymbol{\omega}_1) + \boldsymbol{d}^*(\boldsymbol{\omega}_1)^\intercal(\boldsymbol{\omega}_2 - \boldsymbol{\omega}_1) \geq f(\boldsymbol{\omega}_2) \; .$$

Thus, $\boldsymbol{d}^*(\boldsymbol{\omega}_1)$ is a supergradient of f at $\boldsymbol{\omega}_1$ and function f is concave. □

Now that we have a supergradient for function $f^i(\boldsymbol{\omega})$, we are ready to apply a gradient-based algorithm to find the optimal $\boldsymbol{\omega}^*$ and therefore, the value of $T^i_{R,\,\varepsilon}(\boldsymbol{\mu})^{-1}$. The algorithm of our choice is the mirror ascent algorithm that provides strong guarantees.

Theorem 8. *Let $\boldsymbol{\omega}_1 = (\frac{1}{K}, \ldots, \frac{1}{K})^\intercal$ and learning rate $\eta = \frac{1}{L}\sqrt{\frac{2\log K}{t}}$. Then mirror ascent algorithm optimizing L-Lipschitz function f, with respect to $\|\cdot\|_1$, defined on Δ_K with generalized negative entropy Φ as the mirror map enjoys the following guarantees*

$$f(\boldsymbol{\omega}^*) - f\left(\frac{1}{t}\sum_{s=1}^{t} \boldsymbol{\omega}_s\right) \leq L\sqrt{\frac{2\log K}{t}} \; .$$

Proof. This result can be adapted from [Bubeck, 2015]. □

Now, the last step is using the geometry of the problem and the form of the best response oracle to show that $f^i(\boldsymbol{\omega})$ is Lipschitz for some constant L. This is captured in the following lemma.

Lemma 9. *Let $f^i : \Delta_K \to \mathbb{R}$ be a function such that*

$$f^i(\boldsymbol{\omega}) = \min_{j \neq i} \inf_{\boldsymbol{\lambda} \in \mathcal{M}^{i,j}_{R,\varepsilon}} \sum_{a \in [K]} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \, .$$

Then function f^i is L-Lipschitz with respect to $\| \cdot \|_1$ for any

$$L \geq \max_{a,b \in [K]} \frac{(\mu_a - \mu_b + \varepsilon)^2}{2} \, .$$

5 Algorithm

Solving the complexity problem (3) permits to rely on the SpectralTaS (Algorithm 1) by [Kocák and Garivier, 2020] which is a variation of the asymptotically optimal Track-and-Stop algorithm introduced in [Garivier and Kaufmann, 2016]. Essentially, this algorithm tracks optimal sampling distribution $\boldsymbol{\omega}^*$ with respect to the current estimate of unknown bandit problem $\boldsymbol{\mu}$ and playing accordingly. By playing an arm, the estimate of $\boldsymbol{\mu}$ gets progressively better over time which, in consequence, leads to a more precise sampling distribution $\boldsymbol{\omega}^*$.

The last two ingredients for the algorithm are sampling and stopping rules. We recall them for self-containment, and refer to [Garivier and Kaufmann, 2016; Kocák and Garivier, 2020] for more details.

5.1 Sampling Rule

In order to capture and correct possible arm underestimation, the algorithm introduces extra small amount of exploration. For every $\gamma \in (0, 1/K]$, let $\boldsymbol{\omega}^{*,\gamma}(\boldsymbol{\mu})$ be an L^∞ projection of $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ onto Δ^γ_K defined as $\big\{(\omega_1, \dots, \omega_K) \in [\gamma, 1]^K : \omega_1 + \dots + \omega_K = 1\big\}$. Then the sampling rule is

$$A_{t+1} \in \arg\max_{a \in [K]} \sum_{s=0}^{t} \omega^{*,\gamma_s}_a\big(\hat{\boldsymbol{\mu}}(s)\big) - N_a(t) \, . \qquad (8)$$

where γ_s of order of $1/\sqrt{s}$ which provides as much exploration as possible while not influencing the bounds significantly.

5.2 Stopping Rule

The algorithm should stop as soon as it has gathered sufficient evidence on the superiority of one of the arms with probability $1 - \delta$: for two arms $i \in \mathcal{A}^*_\varepsilon(\hat{\boldsymbol{\mu}})$ and $j \in [K]$, denote by

$$Z_{i,j}(t) = \inf_{\boldsymbol{\lambda} \in \mathcal{M}^{i,j}_{R,\varepsilon}} \sum_{a \in [K]} \frac{1}{2} N_a(t)(\mu_a - \lambda_a)^2$$

the generalized likelihood ratio statistics for the test $\mu_i > \mu_j$. Then the stopping rule is given by

$$\tau = \inf \left\{ t \in \mathbb{N} : \max_{i \in \mathcal{A}^*_\varepsilon(\hat{\boldsymbol{\mu}})} \min_{j \neq i} Z_{i,j}(t) > \beta(t, \delta) \right\} \, , \qquad (9)$$

where $\beta(\cdot, \cdot)$ is a threshold function to be chosen typically slightly larger than $\log(1/\delta)$. Theorem 10 in [Garivier and Kaufmann, 2016] shows that the choice $\beta(t, \delta) = \log(2t(K-1)/\delta)$ and $A_{\tau+1} = \arg\max_{a \in [K]} \hat{\mu}_a(\tau)$ yields a probability of failure $\mathbb{P}_\nu\left(A_{\tau+1} \notin a^*(\boldsymbol{\mu})\right) \leq \delta$.

Algorithm 1 SpectralTaS

```
Input and initialization:
    $\mathcal{L}$ : graph Laplacian
    $\varepsilon, \delta$ : tolerance and confidence parameters
    $R$ : upper bound on the smoothness of $\boldsymbol{\mu}$
    Play each arm $a$ once and observe rewards $r_a$
    $\hat{\boldsymbol{\mu}}_1 = (r_1, \dots, r_K)^\intercal$ : empirical estimate of $\boldsymbol{\mu}$
while Stopping Rule (9) not satisfied do
    Compute $\boldsymbol{\omega}^*(\hat{\boldsymbol{\mu}}_t)$ by mirror ascent
    Choose $A_t$ according to Sampling Rule (8)
    Obtain reward $r_t$ of arm $A_t$
    Update $\hat{\boldsymbol{\mu}}_t$ according to $r_t$
end while
    Output arm $A_* = \arg\max_{a \in [K]} \hat{\mu}_a$
```

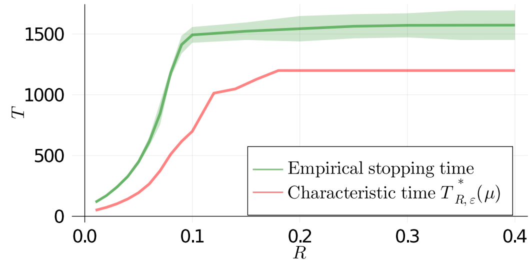

Figure 1: Effect of R on the characteristic and stopping time.

6 Experiments

For the experiments, we used bandit problem

$$\boldsymbol{\mu} = (0,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.4,\ 0.3,\ 0.2,\ 0.1,\ 0)$$

with $K = 11$ arms, a graph that connects all the neighboring actions a and $a + 1$ for every $a \in [K-1]$, $\varepsilon = 0.05$, and different values of R. The following plot demonstrates the effect of smoothness parameter R on both theoretical and empirical stopping times. The green curve represents the average stopping time of 10 runs of SpectralTaS while the red curve represents the characteristic time.

7 Conclusion and Open Problems

We identified the characteristic time of fixed-confidence ε-best arm identification in bandit models with graph smoothness. It appears as a delicate min-max optimization problem, but thanks to a game-theoretic analysis of this complexity we could provide an efficient algorithm for its computation, leading to an asymptotically optimal algorithm. While this provides a complete treatment of the fixed-confidence-setting, the dual fixed-budget setting is still not understood. How good is a strategy following the estimated optimal weights, but stopping at a given time n and not at a chosen stopping time? Are some improvements using the budget n possible? These natural questions are open for further investigations.

Acknowledgements

Aurélien Garivier (ANR chaire SeqALO) and Tomáš Kocák acknowledge the support of the Project IDEXLYON of the University of Lyon, in the framework of the Programme Investissements d'Avenir (ANR-16-IDEX-0005).

References

[Bubeck, 2015] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 2015.

[Degenne and Koolen, 2019] Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, 2019.

[Even-Dar *et al.*, 2006] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

[Garivier and Kaufmann, 2016] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. *Proceedings of the 29th Conference On Learning Theory*, 2016.

[Garivier and Kaufmann, 2019] Aurélien Garivier and Emilie Kaufmann. Non-asymptotic sequential tests for overlapping hypotheses and application to near optimal arm identification in bandit models. *preprint ArXiv:1905.03495*, 2019.

[Karnin *et al.*, 2013] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, 2013.

[Kocák and Garivier, 2020] Tomáš Kocák and Aurélien Garivier. Best arm identification in spectral bandits. *International Joint Conference on Artificial Intelligence*, 2020.

[Lattimore and Szepesvári, 2019] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019.

[Mannor and Tsitsiklis, 2004] Shie Mannor and John N Tsitsiklis. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, 2004.

[Russo, 2016] Daniel Russo. Simple bayesian algorithms for best arm identification. In *Proceedings of the 29th Conference On Learning Theory*, 2016.