

Best Arm Identification in Spectral Bandits

Tomáš Kocák*, Aurélien Garivier

École Normale Supérieure de Lyon

tomas.kocak@gmail.com, aurelien.garivier@ens-lyon.fr

Abstract

We study best-arm identification with fixed confidence in bandit models with graph smoothness constraint. We provide and analyze an efficient gradient ascent algorithm to compute the sample complexity of this problem as a solution of a non-smooth max-min problem (providing in passing a simplified analysis for the unconstrained case). Building on this algorithm, we propose an asymptotically optimal strategy. We furthermore illustrate by numerical experiments both the strategy’s efficiency and the impact of the smoothness constraint on the sample complexity. Best Arm Identification (BAI) is an important challenge in many applications ranging from parameter tuning to clinical trials. It is now very well understood in vanilla bandit models, but real-world problems typically involve some dependency between arms that requires more involved models. Assuming a graph structure on the arms is an elegant practical way to encompass this phenomenon, but this had been done so far only for regret minimization. Addressing BAI with graph constraints involves delicate optimization problems for which the present paper offers a solution.

1 Introduction

This work is devoted to the optimization of a stochastic function on a structured, discrete domain $\mathcal{A} = [K] \triangleq \{1, \dots, K\}$. We consider the noisy black-box model: a call to the function at point $a \in \mathcal{A}$ yields an independent draw of an unknown distribution $\nu_a \in \mathcal{F}$ with mean μ_a , where \mathcal{F} is some family of probability laws. At each time step $t \in \mathbb{N}$, a point $A_t \in \mathcal{A}$ can be chosen (based on past observations) and the corresponding random outcome Y_t of law ν_{A_t} is observed.

The signal $\boldsymbol{\mu} \triangleq (\mu_1, \dots, \mu_K)^\top$ is assumed to be *smooth* in the following sense: \mathcal{A} is equipped with a fixed and *weighted graph structure* G with adjacency matrix $\mathbf{W} = (w_{a,b})_{a,b \in \mathcal{A}}$, and $\boldsymbol{\mu}$ satisfies the graph smoothness property:

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

*Contact Author

for some (known) smoothness parameter R , where \mathcal{L} is the **graph Laplacian** defined as: $\mathcal{L}_{a,b} = -w_{a,b}$ for $a \neq b$ and $\mathcal{L}_{a,a} = \sum_{b \neq a} w_{a,b}$. This enforces values of means $(\mu_a)_a$ at two points $a, b \in \mathcal{A}$ to be close to each other if the weight $w_{a,b}$ is large.

Following a classical framework in statistical testing, a risk parameter δ is fixed, and the goal is to design an algorithm that successfully maximizes μ on \mathcal{A} as fast as possible, with failure probability at most δ . More specifically, the algorithm consists of a *sampling rule* $(\psi_t)_{t \geq 1}$ that chooses to observe $A_t = \psi_t(A_1, Y_1, \dots, A_{t-1}, Y_{t-1})$ at step t , and a stopping rule τ such that whatever the distributions $\nu = (\nu_a)_{a \in \mathcal{A}} \in \mathcal{F}^K$, satisfies the constraint

$$\mathbb{P}_\nu \left(A_{\tau+1} \in a^*(\boldsymbol{\mu}) \triangleq \arg \max_{i \in [K]} \mu_i \right) \geq 1 - \delta.$$

Among all such algorithms, the goal is to minimize the sample complexity $\mathbb{E}_\mu[\tau]$.

This problem is known in the multi-armed bandit literature as *best-arm identification with fixed confidence* [Even-Dar *et al.*, 2006; Gabillon *et al.*, 2012]. In that context, points of the domain \mathcal{A} are called *arms* and the set of distributions ν is called a *model*. Bandit models have raised strong interest in the past years, as multi-armed bandits were gaining popularity thanks to their ability to model a variety of real-world scenarios while providing strong theoretical guarantees for the proposed algorithms. For an introduction to the theory of bandit models and a recent list of references, see [Lattimore and Szepesvári, 2019] and references therein. Best-arm identification, in particular, has many implications in artificial intelligence and data science. For example, in pharmacology, the optimal drug dosage depends on the drug formulation but also on the targeted individuals (genetic characters, age, sex) and environmental interactions. In the industrial context, the shelf life and performance of a manufacturing device depend on its design, but also on its environment and on uncertainties during the manufacturing process. As for targeted recommender systems, their efficiency depends on the ability to understand users’ preferences despite the high variability of choices. Lastly, the performance of a deep neural network depends on its (hyper-)parameters and on the quality of the training set. This short list is of course not comprehensive, but the automation of these representative optimization tasks is one of the great challenges of modern artificial in-

telligence in the fields of personalized medicine, predictive maintenance, online marketing, and autonomous learning.

The problem of best-arm identification has recently received a precise solution for the vanilla version of the multi-armed bandit problem, where the learner gains information only about the arm played. [Garivier and Kaufmann, 2016a; 2019] (see also [Russo, 2016]) have described the information-theoretic barriers of the sample complexity in the form an instance-optimal lower bound on $\mathbb{E}_\mu[\tau]$; furthermore, they have proposed a strategy, called Track-and-Stop, that asymptotically reaches this lower bound.

However, the vanilla bandit model is extremely limiting in many applications, and some additional information has to be incorporated into the learning process. Recently, several papers studied multi-armed bandit problems with the ability to capture the rich structure of real-world problems. In particular, [Degenne and Koolen, 2019; Degenne *et al.*, 2019] have extended the results mentioned above to classes of possibly structured bandits with possibly different goals. In parallel, a fruitful approach to handle large and organized domains is assume the existence of a graph structure on top of arms which provides additional information. There are two most dominant approaches to use this graph structure. The first approach is to use graphs to encode additional information about other arms; playing an arm reveals (fully or partially) the rewards of all the neighbors [Mannor and Shamir, 2011; Alon *et al.*, 2013; Kocák *et al.*, 2014; Alon *et al.*, 2017; Kocák *et al.*, 2016b; 2016a]. The second approach is to encode the similarities between arms by the weights of the edges; the rewards of connected arms are similar [Valko *et al.*, 2014; Kocák *et al.*, 2014; Kocák *et al.*, 2019]. This second approach, which we adopt in this paper, easily permits to capture a lot of information about the problem by the creation of a *similarity graph*, which explains its wide use in signal processing, manifold, and semi-supervised learning. However, most of the work in multi-armed bandits with structure has been done so far under the regret minimization objective, while best arm identification (a practically crucial objective in autoML, clinical trials, A/B testing, etc.) was mostly neglected.

1.1 Our Contributions

In this paper, we aim to fill this gap by combining the best arm identification work of [Garivier and Kaufmann, 2016a] with the spectral setting of [Valko *et al.*, 2014], resulting in spectral best arm identification problem. This setting captures a rich variety of real-world problems that have been neglected so far. We also analyze this setting and provide instance-optimal bounds on the sample complexity. Inspired by [Degenne and Koolen, 2019], we use a game-theoretical point of view on the problem which enables us to obtain two main contributions of this paper:

Contribution 1: simplified and enlightening game-theoretic analysis for the unconstrained case (case without the spectral constraint) of [Garivier and Kaufmann, 2016a] that brings new insights to the problem which enable us to extend the analysis to the constrained case.

Contribution 2: building upon the unconstrained case, we provide and analyze an algorithm that computes the sample complexity as well as the optimal arm pulling proportions in the constrained case. We use this algorithm to propose an asymptotically optimal strategy for fixed-confidence best-arm identification and we provide numerical experiments supporting the theoretical results and showing the interest of the graph constraint for reducing the sample complexity.

1.2 Assumptions

A standard assumption on the family \mathcal{F} of considered probability laws is that they form an exponential family of distributions parametrized by their means so that a bandit problem is fully described by the vector $\boldsymbol{\mu} \triangleq (\mu_1, \dots, \mu_K)^\top$. To avoid technicalities and to emphasize the methodological contributions of this paper (ie. how to perform optimal best-arm identification with graph regularity constraints), we focus here on the case of Gaussian distributions with variance 1. For the same reason, we assume here that the vector $\boldsymbol{\mu}$ has a unique maximum also denoted by $a^*(\boldsymbol{\mu})$. We furthermore denote by $\mathcal{M}_R = \{\boldsymbol{\lambda} \in \mathbb{R}^K : \boldsymbol{\lambda}^\top \mathcal{L} \boldsymbol{\lambda} \leq R\}$ the constrained set of considered signals (bandit problems), and by $\mu_* \triangleq \max_{i \in [K]} \mu_i = \mu_{a^*(\boldsymbol{\mu})}$ the maximum of the signal $\boldsymbol{\mu}$ (we identify the best arm with a singleton $a^*(\boldsymbol{\mu})$).

1.3 Paper Structure

The paper is organized as follows: Section 2 discusses the information-theoretic lower bound and a simple algorithm for the unconstrained case, previously analyzed by [Garivier and Kaufmann, 2016a], with focus on new techniques that considerably simplify the analysis. The constrained case requires more care: we propose in Section 3 a mirror gradient ascent algorithm to compute optimal arm allocation and show non-asymptotic bounds of convergence for this algorithm. Using this procedure permits us in Section 4 to present the **SpectralTaS** algorithm as an extension of the Track-and-Stop devoted to best-arm identification with graph regularity constraint, that reaches the instance-optimal sample complexity. We illustrate its performance and the impact of the smoothness constraint on the sample complexity by reporting some numerical experiments.

2 Sample Complexity

The general information-theoretical lower bound analysis provided by [Garivier and Kaufmann, 2016b] applies to any set of bandit problems, in particular to the set of all the problems with smoothness bounded from above by R .

Proposition 1. *For any δ -correct strategy and any bandit problem $\boldsymbol{\mu}$*

$$\mathbb{E}_\mu[T_\delta] \geq T_R^*(\boldsymbol{\mu})k(\delta, 1 - \delta)$$

where

$$T_R^*(\boldsymbol{\mu})^{-1} \triangleq \sup_{\omega \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \mathcal{A}_R(\boldsymbol{\mu})} \sum_{a \in [K]} \omega_a k(\mu_a, \lambda_a) \quad (1)$$

for Δ_K being K -dimensional simplex, $k(\mu_a, \lambda_a)$ KL-divergence between distributions parametrized by μ_a and λ_a ,

and for the set $\mathcal{A}_R(\boldsymbol{\mu})$ of bandit problems with different best arm than $\boldsymbol{\mu}$, defined as

$$\begin{aligned}\mathcal{A}_{R,i}(\boldsymbol{\mu}) &\triangleq \{\boldsymbol{\lambda} \in \mathcal{M}_R : \lambda_i \geq \lambda_{a^*}(\boldsymbol{\mu})\} \\ \mathcal{A}_R(\boldsymbol{\mu}) &\triangleq \cup_{i \neq a^*}(\boldsymbol{\mu}) \mathcal{A}_{R,i}(\boldsymbol{\mu}).\end{aligned}$$

Indeed, even if the constrained case is not explicitly covered there, this result is easily shown by following the lines of the proof of Theorem 1 in [Garivier and Kaufmann, 2016b], with $\mathcal{A}_R(\boldsymbol{\mu})$ as the set of alternatives to $\boldsymbol{\mu}$.

This lower bound proves to play an important role in designing algorithms for the best arm identification since, roughly speaking, the time needed to distinguish between bandit problems $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, if the number of pulls of arm i is proportional to ω_i , scales with inverse of $\sum_{a \in [K]} \omega_a k(\mu_a, \lambda_a)$. Therefore, the minimization part of problem (1) selects the alternative problem which is the most difficult to distinguish from $\boldsymbol{\mu}$ while playing according to $\boldsymbol{\omega}$ while, on the other hand, the maximization part of problem (1) chooses $\boldsymbol{\omega}$ such that the expected stopping time is as small as possible, even in the worst case scenario chosen by the minimization part of the problem.

Thus, a sampling strategy playing according to $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ that maximizes expression (1) is optimal in the worst case, and having a procedure that computes optimal $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ enables to design a best arm identification algorithm with optimal sample complexity.

In this section, we provide an algorithm to compute this sample complexity $T_R^*(\boldsymbol{\mu})$ as well as the optimal proportion $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ of the arm allocation for the learner. We proceed in three steps.

Step 1: we introduce the *best response oracle* that computes the best response $\boldsymbol{\lambda}^*(\boldsymbol{\omega})$ to fixed $\boldsymbol{\omega}$ by solving the minimization part of problem (1) for chosen $\boldsymbol{\omega}$:

$$\boldsymbol{\lambda}^*(\boldsymbol{\omega}) \in \arg \min_{\boldsymbol{\lambda} \in \mathcal{A}_R(\boldsymbol{\mu})} \sum_{a \in [K]} \omega_a k(\mu_a, \lambda_a).$$

Step 2: we show that substituting the minimization part of problem (1) by $\boldsymbol{\lambda}^*(\boldsymbol{\omega})$, the resulting maximization problem is concave with respect to $\boldsymbol{\omega}$. Moreover, supergradient for this problem is computable using the same best response oracle we used for the minimization part of the problem.

Step 3: using supergradient we can apply any supergradient ascent algorithm to find optimal arm allocation $\boldsymbol{\omega}^*$. The algorithm of our choice is mirror ascent with generalized negative entropy as a mirror map. We chose this algorithm due to its strong guarantees on convergence rate in simplex setup however, even basic gradient-based algorithms could be applied. The choice of mirror ascent algorithm will be apparent later in Section 3, we show that this algorithm enjoys convergence rate proportional to $\sqrt{\log K}$ instead of \sqrt{K} of the basic supergradient ascent algorithm.

As briefly mentioned in the introduction, for the sake of simplicity of presentation, we assume that the distributions ν_{μ_a} are Gaussian. This simplifies some of the calculations while our bounds remain true even in the case of sub-Gaussian random variables, including a wide class of distri-

butions among which any distributions with bounded support or Bernoulli distribution.

Even though using a general exponential family of probability distributions uses similar techniques, all the proofs are more technical and rather out of the scope of a conference paper.

With assumption that the distributions associated with arms are Gaussian with normalized variance, KL-divergence of ν_{μ_a} and ν_{λ_a} can be expressed as

$$k(\mu_a, \lambda_a) = \frac{(\mu_a - \lambda_a)^2}{2}.$$

2.1 Best Arm Identification without Spectral Constraint

This section is dedicated to the vanilla setting of [Garivier and Kaufmann, 2016a]. Even though this problem can be seen as a special case of spectral setting either (for the edgeless graph or for smoothness parameter R being ∞), we have decided to analyze it separately. The reason is that we can demonstrate techniques later used in the constrained case while proving the main result of [Garivier and Kaufmann, 2016a] with more insightful arguments that are significantly more elegant. Also, not having the spectral constraint allows us to have best response oracle with a closed-form solution which implies a simple way of finding optimal arm allocation $\boldsymbol{\omega}^*$. This is very different from the spectral setting where best response oracle does not have a closed-form solution and therefore, we can not avoid a numerical procedure to find $\boldsymbol{\omega}^*$.

Theorem 2. *Without loss of generality, assume that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$ and define $\{x_a(c)\}_{a=2}^K$, for some positive constant c , recursively as*

$$x_K(c) = c$$

$$x_{a-1}(c) = (1 + x_a(c)) \left(\frac{\mu_1 - \mu_{a-1}}{\mu_1 - \mu_a} \right)^2 - 1$$

for $2 < a \leq K$. Let $f(c)$ be a function with parameter c defined as

$$f(c) = \sum_{a=2}^K x_a(c)^2$$

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

$$\omega_1^*(\boldsymbol{\mu}) = \frac{1}{1 + \sum_{a=2}^K x_a(c^*)}$$

$$\omega_a^*(\boldsymbol{\mu}) = x_a(c^*) \omega_1^*(\boldsymbol{\mu})$$

Remark. *Since $\mu_1 > \mu_{a-1} \geq \mu_a$, we see that*

$$\left(\frac{\mu_1 - \mu_{a-1}}{\mu_1 - \mu_a} \right)^2 \leq 1,$$

which in consequence means that. If $x_a(c)$ is negative, $x_{a-1}(c)$ is negative as well. If $x_a(c)$ is positive, $x_{a-1}(c) \leq x_a(c)$. Thus, $f(0) \leq 0$. Moreover, f is an increasing and continuous function with $\lim_{c \rightarrow \infty} f(c) = \infty$. Therefore, the existence of c^* is guaranteed.

In the rest of this section, we build the necessary tools to provide the proof of Theorem 2.

Best Response Oracle - Vanilla Setting

In the case of smoothness $R = \infty$, the spectral constraint is always satisfied which provides a very simple closed form solution for the best response oracle given by the following lemma.

Lemma 3. *Let ω be a fixed vector from Δ_K . Then the best response $\lambda^i(\omega) \in \mathcal{A}_{\infty,i}(\mu)$ with the best arm i is*

$$\lambda^i(\omega) = (t, \mu_2, \dots, \mu_{i-1}, t, \mu_{i+1}, \dots, \mu_K)^\top$$

for t being a weighted average of μ_1 and μ_i with weights ω_1 and ω_i

$$t = \frac{\mu_1 \omega_1 + \mu_i \omega_i}{\omega_1 + \omega_i}.$$

The proof of this Lemma can be obtained by simple calculus.

Game-Theoretical Point of View

Our optimization problem (1) can be seen as a zero sum game where Player 1 plays a vector from simplex Δ_K and Player 2 plays point λ from $\mathcal{A}_R(\mu)$. We would like to have guarantees on the existence of Nash equilibrium however $\mathcal{A}_R(\mu)$ is not a convex set which poses a problem. Also, we can not directly convexify the problem since it would change the value of $T_R^*(\mu)^{-1}$. To get around this obstacle, we define

$$\mathcal{D}_{R,i}(\mu) \triangleq \{(k(\mu_1, \lambda_1), \dots, k(\mu_K, \lambda_K))^\top : \lambda \in \mathcal{A}_{R,i}(\mu)\},$$

$$\mathcal{D}_R(\mu) \triangleq \cup_{i \neq a^*(\mu)} \mathcal{D}_{R,i}(\mu).$$

This enables us to rewrite the optimization problem in the following form

$$T_R^*(\mu)^{-1} \triangleq \sup_{\omega \in \Delta_K} \inf_{\mathbf{d} \in \mathcal{D}_R(\mu)} \omega^\top \mathbf{d} \quad (2)$$

where Player 2, instead of playing λ , plays vectors of divergences $(k(\mu_1, \lambda_1), \dots, k(\mu_K, \lambda_K))^\top$ where the element on position i of the vector is divergence $k(\mu_i, \lambda_i)$ of the distributions corresponding to i -th elements of μ and λ . This still haven't solve our issue with non-convexity of set $\mathcal{D}_R(\mu)$, however, now we can simply optimize over the convex hull of $\mathcal{D}_R(\mu)$

$$T_R^*(\mu)^{-1} \triangleq \sup_{\omega \in \Delta_K} \inf_{\mathbf{d} \in \text{Conv}(\mathcal{D}_R(\mu))} \omega^\top \mathbf{d} \quad (3)$$

and the following lemma guarantees that the values of optimization problems (2) and (3) are the same.

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

$$\inf_{\mathbf{d} \in \mathcal{D}} \omega^\top \mathbf{d} = \inf_{\mathbf{d} \in \text{Conv}(\mathcal{D})} \omega^\top \mathbf{d}$$

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

Now we are left a zero sum game (3) with convex and compact sets of actions for both players. This guarantees existence of Nash equilibrium with the best actions denoted by $\omega^*(\mu)$ and $\mathbf{d}^*(\mu)$. Whole purpose of seeing (3) as a zero-sum game is the following lemma

Lemma 5. *Let $\omega^*(\mu)$ and $\mathbf{d}^*(\mu)$ are vectors for which Nash equilibrium of game (3) is attained then there exist a constant c such that*

$$d_i^*(\mu) = c, \quad \text{for all } i \in [K].$$

Proof of Theorem 2

Now we have all necessary tools to prove Theorem 2. Note that we still assume that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$. Let ω^* be the Nash equilibrium strategy. As we showed in Lemma 3, the best response $\lambda^i(\omega^*) \in \mathcal{A}_{\infty,i}(\mu)$ to ω^* has a particular form

$$\lambda^i(\omega^*) = (t, \mu_2, \dots, \mu_{i-1}, t, \mu_{i+1}, \dots, \mu_K)^\top,$$

where $t = \frac{\mu_1 \omega_1^* + \mu_i \omega_i^*}{\omega_1^* + \omega_i^*}$. This $\lambda^i(\omega^*)$ corresponds to point $\mathbf{d}^i(\omega^*)$ from $\mathcal{D}_{R,i}$ defined as

$$\mathbf{d}^i(\omega^*) = (y_i, 0, \dots, 0, z_i, 0, \dots, 0)^\top$$

for $y_i = (\mu_1 - \mu_i)^2/2$ and $z_i = (\mu_i - \mu_K)^2/2$. It is important to notice that the only two non-zero elements are at positions 1 and i . The Nash equilibrium strategy \mathbf{d}^* of Player 2 can be expressed as a convex combination of optimal points from $\mathcal{D}_R(\mu)$ as showed in Lemma 4. The only candidates are points $\mathbf{d}^i(\omega^*)$ for $i \in \{2, \dots, K\}$ since all the other points from $\mathcal{D}_R(\mu)$ are sub-optimal. Moreover, using Lemma 5 we know that all the elements of \mathbf{d}^* are equal which in particular means that \mathbf{d}^* needs to be a convex combination of all $K - 1$ vectors $\mathbf{d}^i(\omega^*)$ and all of them are equally good and from the definition of $\mathbf{d}^i(\omega^*)$ we have

$$(\omega^*)^\top \mathbf{d}^i(\omega^*) = (\omega^*)^\top \mathbf{d}^j(\omega^*)$$

for every $i, j \geq 2$. This expression can be further modified to obtain

$$\frac{\omega_i^*}{\omega_1^*} = \left(1 + \frac{\omega_j^*}{\omega_1^*}\right) \left(\frac{\mu_1 - \mu_i}{\mu_1 - \mu_j}\right)^2 - 1$$

Denoting $\frac{\omega_i^*}{\omega_1^*}$ by c^* and using $i = j - 1$ we directly obtain

that the recurrence from the theorem holds for $x_i(c^*) = \frac{\omega_i^*}{\omega_1^*}$. Therefore, it is enough to know c^* and ω_1^* to reconstruct ω^* .

Lemma 5 provides condition for \mathbf{d}^* . We know that it can be written as a convex combination of $K - 1$ vectors $\mathbf{d}^i(\omega^*)$. This is possible only if vector $\mathbf{1}$ of ones can be expressed as their linear combination which gives us condition

$$\sum_{i=2}^K \frac{\mathbf{d}^i(\omega^*)}{z_i} = \mathbf{1}.$$

We know that all the elements of the resulting vector are equal to 1 except for the first element which has a value

$$\sum_{i=2}^K \frac{y_i}{z_i} = \sum_{i=2}^K \left(\frac{\omega_i^*}{\omega_1^*}\right)^2 = \sum_{i=2}^K x_i(c^*)^2 = f(c^*).$$

Therefore, this value has to be 1 too, which gives us the second part of the theorem. Now we can recover all the ratios ω_i^*/ω_1^* and we also know that the sum of all ω_i is 1 which provides all the necessary ingredients for the proof of the last part of the theorem.

2.2 Best Response Oracle - Spectral Setting

As we showed in Section 2.1, Lemma 3, finding the best response to ω , in the vanilla setting, is not a difficult problem since the optimization problem involves only linear constraints. The situation in the spectral setting is a little bit more complicated.

In this part, we again assume that $\mu_1 > \mu_2 \cdots \geq \mu_K$ and focus on the best response oracle

$$\lambda^*(\omega) = \arg \min_{\lambda \in \mathcal{A}_R(\mu)} \sum_{a \in [K]} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}$$

in spectral setting, where R is a finite upper bound on smoothness of μ with respect to graph G given by its Laplacian \mathcal{L} . We first find best responses

$$\lambda^i(\omega) = \arg \min_{\lambda \in \mathcal{A}_{R,i}(\mu)} \sum_{a \in [K]} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}$$

with respect to convex sets $\mathcal{A}_{R,i}(\mu)$ and then take the best out of these $K-1$ vectors.

First of all, notice that if the response of vanilla oracle denoted by $\lambda_\infty^i(\omega)$ has smoothness smaller than R , the response of spectral oracle should be the same. In the case where $S_G(\lambda_\infty^i(\omega)) \geq R$ we can restrict our search and look for λ with $S_G(\lambda)$ exactly R thanks to the following lemma.

Lemma 6. *Let $\lambda_\infty^i(\omega)$ be the response of vanilla oracle such that $\lambda_\infty^i(\omega)^\top \mathcal{L} \lambda_\infty^i(\omega) > R$ then the response of spectral oracle $\lambda^i(\omega)$ satisfies $\lambda^i(\omega)^\top \mathcal{L} \lambda^i(\omega) = R$.*

2.3 Best Response Oracle Implementation

Now lets consider the case where vanilla oracle produces a vector that does not satisfy smoothness constraint. In this case we know, using Lemma 6, that the best response $\lambda^i(\omega)$ of spectral oracle has smoothness R . Therefore, we can use standard Lagrange multiplier method to solve this problem.

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

We should not forget that we still need to ensure that $\lambda_1 = \lambda_i$ since we want to find the solution in $\mathcal{A}_{R,i}(\mu)$ with best arm i . Therefore, we can simplify the notation and later calculations by the following definitions

- $\tilde{\lambda}$: created from λ by removing the first component
- $\tilde{\omega}$: created from ω by removing the first component
- $\tilde{\mu}$: created from μ by averaging components 1st and i -th components w.r. to ω and removing the first component

$$\tilde{\mu}_{i-1} = \frac{\omega_1 \mu_1 + \omega_i \mu_i}{\omega_1 + \omega_i}.$$

- $\tilde{\mathcal{L}}$: created from \mathcal{L} by adding the first row and column to i -th row and column, adjusting diagonal element $\mathcal{L}_{i,i}$ so that the sum in the i -th row is zero, removing the first row and column

Using previous definitions we can define

$$\tilde{F}(\tilde{\lambda}, \gamma) \triangleq \sum_{a \in [K-1]} \tilde{\omega}_a \frac{(\tilde{\mu}_a - \tilde{\lambda}_a)^2}{2} + \gamma (\tilde{\lambda}^\top \tilde{\mathcal{L}} \tilde{\lambda} - R)$$

for which $\tilde{F}(\tilde{\lambda}, \gamma) = F(\lambda, \gamma) + c$ holds, assuming that $\lambda_1 = \lambda_i$. The identity can be checked by using previous definitions. The main advantage of this transformation is that

now we don't have any constraint on $\tilde{\lambda}$. Now we can take partial derivatives with respect to $\tilde{\lambda}_j$ and γ and set them equal to 0 in order to find the best $\tilde{\lambda}$ and later reconstruct λ :

$$\begin{aligned} \nabla_{\tilde{\lambda}} \tilde{F}(\tilde{\lambda}, \gamma) &= \tilde{\Omega}(\tilde{\lambda} - \tilde{\mu}) + 2\gamma \tilde{\mathcal{L}} \tilde{\lambda} = \mathbf{0}, \\ \frac{\partial \tilde{F}(\tilde{\lambda}, \gamma)}{\partial \gamma} &= \tilde{\lambda}^\top \tilde{\mathcal{L}} \tilde{\lambda} - R = 0, \end{aligned}$$

where $\tilde{\Omega}$ is a diagonal matrix with $\tilde{\omega}$ on its diagonal. First expression is just a system of $K-1$ linear equations with parameter γ with solution

$$\tilde{\lambda}(\gamma) = (\tilde{\Omega} + 2\gamma \tilde{\mathcal{L}})^{-1} \tilde{\Omega} \tilde{\mu}.$$

To find the best γ so that also the second expression is true, we can use for example bisection method.

The last step is choosing $i > 1$ that minimizes $F(\lambda, \gamma)$.

Properties of Maximization Problem

Before we proceed to the algorithm computing optimal $\omega^*(\mu)$ we need the following two lemmas to be able to compute supergradients and guarantee the convergence of the algorithm.

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

Lemma 8. *Let $f : \Delta_K \rightarrow \mathbb{R}$ be a function such that*

$$f(\omega) = \inf_{\lambda \in \mathcal{A}_R(\mu)} \sum_{i=1}^K \omega_i k(\mu_i, \lambda_i).$$

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

$$L \geq \max_{i,j \in [K]} k(\mu_i, \mu_j).$$

3 Best Arm Allocation Algorithm

Now that we are able to compute best response $\lambda^*(\omega)$ and therefore, supergradient of concave function $f(\omega) = \sum_{a \in [K]} \omega_a k(\mu_a, \lambda_a^*(\omega))$ at ω , we have all the ingredients needed for any supergradient-based algorithm to compute best arm allocation $\omega^* = \arg \max_{\omega \in \Delta_K} f(\omega)$. The algorithm of our choice is mirror ascent algorithm with generalized negative entropy as the mirror map:

$$\Phi(\omega) = \sum_{a \in [K]} (\omega_a \log(\omega_a) - \omega_a).$$

Theorem 9. *Let $\omega_1 = (\frac{1}{K}, \dots, \frac{1}{K})^\top$ and learning rate $\eta = \frac{1}{L} \sqrt{\frac{2 \log K}{t}}$. Then mirror ascent algorithm optimizing function f defined on Δ_K with generalized negative entropy Φ as the mirror map enjoys the following guarantees*

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

for any $L \geq \max_{i,j \in [K]} k(\mu_i, \mu_j)$.

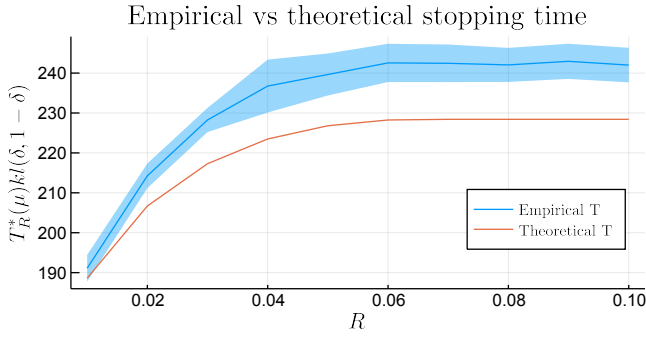


Figure 1: Theoretical vs empirical expected stopping time as a function of smoothness parameter R with optimal value of R being 0.01.

Proof. It is easy to show that Φ is 1-strongly convex w.r.t. $\|\cdot\|_1$ and $\sup_{\omega \in \Delta_K} \Phi(\omega) - \Phi(\omega_1) \leq \log K$ since

$$\sum_{a \in [K]} \left[\omega_a \log \omega_a - \frac{1}{K} \log \frac{1}{K} \right] \leq \sum_{a \in [K]} -\frac{1}{K} \log \frac{1}{K} = \log K$$

holds for any $\omega \in \Delta_K$. Using Lemma 8 we have that f is L -Lipschitz w.r.t. $\|\cdot\|_1$ for any $L \geq \max_{i,j \in [K]} k(\mu_i, \mu_j)$. This gives us all the necessary ingredients mirror ascent guarantees in Theorem 4.2 from [Bubeck, 2015]. \square

3.1 Value of Spectral Constraint

Having an assumption that captures the problem might improve the learning speed significantly. We demonstrate this effect in a simplistic scenario where $\mu = (0.9, 0.5, 0.6)^\top$, graph has only one edge between nodes 2 and 3, and the range of R is from 0.01 to $\mu^\top \mathcal{L} \mu$ to 0.1. $R = 0.1$ is completely non-restrictive and the best response to every ω is the same as in the best arm identification problem without the spectral constraint. This can be seen on Figure 1 where we plot the value of $T_R^*(\mu)$ (red curve) as a function of R which is proportional to the stopping time lower bound.

4 Spectral BAI Algorithm

The algorithm for the best arm identification with the spectral constraint `SpectralTaS` (Algorithm 1) is a variant of Track-and-Stop algorithm introduced in [Garivier and Kaufmann, 2016b]. We discuss the main ingredients of the algorithm in the next part of this section.

Sampling rule. As a by-product of the lower bound analysis, Proposition 1 provides the existence of *optimal sampling weights* $\omega^*(\mu)$ that need to be respected in order to reach the optimal sample complexity. `SpectralTaS` simply tracks, at every timestep, some guess of these optimal proportions that is obtained by solving the sample complexity optimization problem associated with the *current estimates* $\hat{\mu}_t$ of the means μ . In order to capture a possibility of the initial underestimation of an arm, some level of exploration is needed and enforced by the algorithm: for every $\varepsilon \in (0, 1/K]$, let $\omega^{*,\varepsilon}(\mu)$ be an L^∞ projection of $\omega^*(\mu)$ onto Δ_K^ε defined as $\{(w_1, \dots, w_K) \in [\varepsilon, 1]^K : w_1 + \dots + w_K = 1\}$. Then the sampling rule is

$$A_{t+1} \in \arg \max_{a \in [K]} \sum_{s=0}^t \omega_a^{*,\varepsilon_s}(\hat{\mu}(s)) - N_a(t). \quad (4)$$

Algorithm 1 `SpectralTaS`

- 1: **Input and initialization:**
 - 2: \mathcal{L} : graph Laplacian
 - 3: δ : confidence parameter
 - 4: R : upper bound on the smoothness of μ
 - 5: Play each arm a once and observe rewards r_a
 - 6: $\hat{\mu}_1 = (r_1, \dots, r_K)^\top$: empirical estimate of μ
 - 7: **while** Stopping Rule (5) not satisfied **do**
 - 8: Compute $\omega^*(\hat{\mu}_t)$ by mirror ascent
 - 9: Choose A_t according to Sampling Rule (4)
 - 10: Obtain reward r_t of arm A_t
 - 11: Update $\hat{\mu}_t$ according to r_t
 - 12: **end while**
 - 13: Output arm $A_* = \arg \max_{a \in [K]} \hat{\mu}_a$
-

As shown in [Garivier and Kaufmann, 2016a] (Lemma 7 and Proposition 9), the choice $\varepsilon_s = (K^2 + s)^{-1/2}/2$ ensures that the number $N_a(t)$ of draws of arm a converges almost-surely to $\omega_a^*(\mu)$ as t goes to infinity.

Stopping rule. The algorithm should stop as soon as it has gathered sufficient evidence on the superiority of one of the arms with risk δ . The design of an optimal sequential testing procedure for the hypothesis $\mu_a > \max_{b \neq a} \mu_b$ can be traced back to [Chernoff, 1959], and is discussed in detail in [Garivier and Kaufmann, 2019]. We simply recall its form here: for two arms $a, b \in [K]$, denote by $\hat{\mu}_{a,b}(t) = (N_a(t)\hat{\mu}_a(t) + N_b(t)\hat{\mu}_b(t))/(N_a(t) + N_b(t))$ and by $Z_{a,b} = \text{sign}(\hat{\mu}_a(t) - \hat{\mu}_b(t))(N_a(t)(\hat{\mu}_a(t) - \hat{\mu}_{a,b}(t))^2 + N_b(t)(\hat{\mu}_b(t) - \hat{\mu}_{a,b}(t))^2)/2$ the generalized likelihood ratio statistics for the test $\mu_a > \mu_b$. Then the stopping rule is given by

$$\tau = \inf \left\{ t \in \mathbb{N} : \max_{a \in [K]} \min_{b \neq a} Z_{a,b}(t) > \beta(t, \delta) \right\}, \quad (5)$$

where $\beta(\cdot, \cdot)$ is a threshold function to be chosen typically slightly larger than $\log(1/\delta)$. Theorem 10 in [Garivier and Kaufmann, 2016a] 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 $\mathbb{P}_\nu(A_{\tau+1} \notin a^*(\mu)) \leq \delta$.

Empirical evaluation. Using the same experiment setting as in Section 3.1 we have evaluated `SpectralTaS` for 10 different values of R ranging from true $R = \mu^\top \mathcal{L} \mu = 0.01$ to $R = 0.1$ and we plotted theoretical stopping time (red line) as well as average and standard deviation of 20 runs of `SpectralTaS` algorithm (blue line). Figure 1 shows that the algorithm can utilize the spectral constraint and improve stopping time significantly whenever the value of R is close to the real smoothness of the problem, as suggested by theory.

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

- [Alon *et al.*, 2013] Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In *Neural Information Processing Systems*, 2013.
- [Alon *et al.*, 2017] Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.
- [Bubeck, 2015] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.
- [Chernoff, 1959] Herman Chernoff. Sequential design of Experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.
- [Degenne and Koolen, 2019] Rémy Degenne and Wouter M. Koolen. Pure Exploration with Multiple Correct Answers. *technical report*, feb 2019.
- [Degenne *et al.*, 2019] Rémy Degenne, Wouter M. Koolen, and Pierre Ménard. Non-Asymptotic Pure Exploration by Solving Games. *technical report*, jun 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.
- [Gabillon *et al.*, 2012] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Neural Information Processing Systems*, 2012.
- [Garivier and Kaufmann, 2016a] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference On Learning Theory*, 2016.
- [Garivier and Kaufmann, 2016b] Aurélien Garivier and Emilie Kaufmann. Optimal Best Arm Identification with Fixed Confidence. *29th Annual Conference on Learning Theory*, 2016.
- [Garivier and Kaufmann, 2019] Aurélien Garivier and Emilie Kaufmann. Non-Asymptotic Sequential Tests for Overlapping Hypotheses Applied to Near Optimal Arm Identification in Bandit Models. *technical report*, jun 2019.
- [Kocák *et al.*, 2014] Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Neural Information Processing Systems*, 2014.
- [Kocák *et al.*, 2016a] Tomáš Kocák, Gergely Neu, and Michal Valko. Online learning with Erdos-Rényi side-observation graphs. In *Conference on Uncertainty in Artificial Intelligence*, 2016.
- [Kocák *et al.*, 2016b] Tomáš Kocák, Gergely Neu, and Michal Valko. Online learning with noisy side observations. In *International Conference on Artificial Intelligence and Statistics*, 2016.
- [Kocák *et al.*, 2019] Tomáš Kocák, Rémi Munos, Branislav Kveton, Shipra Agrawal, and Michal Valko. Spectral Bandits. *Journal of Machine Learning Research*, 2019.
- [Kocák *et al.*, 2014] Tomáš Kocák, Michal Valko, Rémi Munos, and Shipra Agrawal. Spectral Thompson sampling. In *Proceedings of the National Conference on Artificial Intelligence*, 2014.
- [Lattimore and Szepesvári, 2019] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019.
- [Mannor and Shamir, 2011] Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Neural Information Processing Systems*, 2011.
- [Russo, 2016] Daniel Russo. Simple bayesian algorithms for best arm identification. In *Proceedings of the 29th Conference On Learning Theory*, 2016.
- [Valko *et al.*, 2014] Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, 2014.