Dissimilarity Bandits

Paolo Battellani **Alberto Maria Metelli** **Francesco Trovò**

Dipartimento di Elettronica, Informazione e Bioingegneria,
Politecnico di Milano, Milano
paolo.battellani@mail.polimi.it, {albertomaria.metelli, francesco1.trovo}@polimi.it

Abstract

We study a novel sequential decision-making setting, namely the *dissimilarity bandits*. At each round, the learner pulls an arm that provides a stochastic d-dimensional observation vector. The learner aims to identify the pair of arms with the maximum dissimilarity, where such an index is computed over pairs of expected observation vectors. We propose `Successive Elimination for Dissimilarity` (`SED`), a fixed-confidence best-pair identification algorithm based on sequential elimination. `SED` discards individual arms when there is statistical evidence that they cannot belong to a pair of most dissimilar arms and, thus, effectively exploits the structure of the setting by reusing the estimates of the expected observation vectors. We provide results on the sample complexity of `SED`, depending on $H\!P$, a novel index characterizing the complexity of identifying the pair of the most dissimilar arms. Then, we provide a sample complexity lower bound, highlighting the challenges of the identification problem for dissimilarity bandits, which is almost matched by our `SED`. Finally, we compare our approach over synthetically generated data and a realistic environmental monitoring domain against classical and combinatorial best-arm identification algorithms for the cases $d = 1$ and $d > 1$.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

The classical Multi-Armed Bandit framework (MAB, Lattimore and Szepesvári, 2020) has been developed through the years to deal with the problem of exploration in sequential decision processes with partial feedback. In such a model, at each round, a learner has to select an option, a.k.a. arm, in a finite set and receives a noisy reward corresponding to their choice. Commonly, in the bandit literature, algorithms for pursuing two different objectives have been designed. First, *regret minimization* (Auer et al., 2002) algorithms aim to minimize the cumulative loss due to the learning process w.r.t. playing the optimal arm. Second, *Best-Arm Identification* (BAI, Audibert et al., 2010) algorithms aim to identify with high probability the arm providing the largest expected reward. In this paper, we focus on BAI for a novel setting called *dissimilarity bandits*, in which the learner's expected reward is associated with pairs of arms through a known dissimilarity function.

Over the years, many variants of the MAB setting have been designed and analyzed to model different sequential decision problems occurring in real-world settings. They either extend the standard MAB setting for rewards having different nature than the scalar one, e.g., dueling bandits (Yue et al., 2012; Sui et al., 2018), or they make use of the specific structure of the analyzed problem to speed up the learning process, e.g., Combinatorial MABs (Cesa-Bianchi and Lugosi, 2012; Chen et al., 2013) or linear bandits (Abbasi-Yadkori et al., 2011). In the present work, we analyze a newly defined setting in which selecting a specific arm provides a d-dimensional noisy observation vector, and the learner reward is provided by a dissimilarity function that applies to pairs of expected observation vectors. The learner aims to identify, with high probability (a.k.a. *fixed-confidence* BAI), the pair of arms with the maximum dissimilarity.

Motivation This setting has been inspired and mo-

tivated by *environmental monitoring* applications for potable water (Gabrielli et al., 2021). In such a scenario, each sample of water provides a set of measures consisting of a vector of values representing the response of the water to the excitation using energy beams with different frequencies (a.k.a. flourescence analysis). The response over different frequencies provides information about the bacterial population in the water. This analysis is repeated for water samples collected during different hours throughout the day for a specific location. The bacteria base concentrations are due to environmental factors and vary significantly over the day due to human activities related to potable water. However, thanks to the routine present in everyday human activities, daily patterns are present over the weeks/months. This allows the modeling of measurements taken at the same hour on different days as samples taken from the same stochastic distribution. In such a setting, the values corresponding to a single hour provide no information about the increase/decrease of a specific bacterium (i.e., biological stability). Conversely, pairs of measurements highlight if anomalous changes are occurring in specific hours. The monitoring campaign aims to identify with the smallest number of samples possible the pairs of hours presenting the most dissimilar fluorescence response since their dissimilarity corresponds to the influence of human activities in the water stream. Currently, these campaigns are structured using naïve schemes, e.g., sampling each hour with the same frequency over a given period, which may result in a suboptimal sampling strategy. Conversely, using the modeling approach offered by the dissimilarity bandits and with the algorithm developed here, we provide a more efficient sampling scheme and, given predefined confidence, a stopping time for the monitoring campaign. For further discussion on related works, see Section 6.

Original Contributions This paper provides:

- the definition of the *dissimilarity bandit* setting for the first time in the bandit literature (Section 2);
- the design of the `Successive Elimination for Dissimilarity` (`SED`) algorithm to perform fixed-confidence BAI specifically crafted for the dissimilarity bandit setting (Section 3), which is provided with a computationally efficient implementation (Section 3.1);
- results on the sample complexity of `SED`, depending on HP, a novel index characterizing the complexity of identifying the pair of the most dissimilar arms (Section 3.2);
- an expected sample complexity lower bound that almost matches the upper bound for `SED` and highlights the challenges of our problem (Section 4).
- comparison of our approach over synthetically generated data and a realistic environmental monitoring domain in comparison with classical and combinatorial best-arm identification algorithms for both the cases $d = 1$ and $d > 1$ (Section 7).

The proofs of the results reported in the main paper are deferred to Appendix A for space reasons.

2 PROBLEM FORMULATION

Notation We consider a vector-valued Multi-Armed Bandit (MAB) problem $\underline{\boldsymbol{\nu}} = (\boldsymbol{\nu}_i)_{i\in[\![K]\!]}$ made of $K \in \mathbb{N}$ arms. At each round $t \in \mathbb{N}$, the agent pulls an arm $I_t \in [\![K]\!] \coloneqq \{1, \dots, K\}$ and receives a vector-valued feedback $\boldsymbol{x}_t \sim \boldsymbol{\nu}_{I_t}$ (a.k.a. observation vector) belonging to $\mathbb{R}^d$. For every arm $i \in [\![K]\!]$ we have that $\boldsymbol{x} = (x_1, \dots, x_d)^\top$ is the realization sampled from the distribution $\boldsymbol{\nu}_i = (\nu_{i,1}, \dots, \nu_{i,d})^\top$, with expectation $\boldsymbol{\mu}_i = (\mu_{i,1}, \dots, \mu_{i,d})^\top$ (a.k.a. expected observation vector). The components x_j are assumed to be independent and σ^2-subgaussian, formally:

Assumption 2.1 (Subgaussian Random Vector with Independent Components)**.** *For every arm $i \in [\![K]\!]$, every component x_j for $j \in [\![d]\!]$ of the random vector $\boldsymbol{x} \sim \boldsymbol{\nu}_i$ is independent of the others and σ^2-subgaussian, i.e.:*

$$\mathop{\mathbb{E}}_{x_j\sim\nu_{i,j}}[\exp(\lambda(x_j - \mu_{i,j}))] \leqslant \exp\left(\frac{\sigma^2\lambda^2}{2}\right), \quad \forall\lambda \in \mathbb{R}.$$

However, we note that while the independence between vector components is required for our analysis, a similar result, with a higher constant factor in the complexity bound, can be achieved even if this assumption does not hold (see comments after Lemma 3.1).

Optimality Let $w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a known symmetric dissimilarity function, the goal of the agent is to find an *optimal (unsorted) pair of arms* $\{i^*, j^*\}$, with $i^*, j^* \in [\![K]\!]$ and $i^* \neq j^*$, i.e., that maximizes the function w, formally:

$$\{i^*, j^*\} \coloneqq \mathop{\arg\max}_{i,j\in[\![K]\!], i\neq j} w(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j),$$

and we denote with $w^* \coloneqq w(i^*, j^*)$ the value of the dissimilarity for the optimal pair $\{i^*, j^*\}$. As commonly done in other BAI settings (Garivier and Kaufmann, 2016), we assume that the optimal solution is unique; otherwise, the problem of BAI would become ill-posed. For a pair $\{i, j\} \neq \{i^*, j^*\}$, we define the sub-optimality gap as follows:

$$\Delta_{\{i,j\}} \coloneqq w^* - w(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j).$$

Intuitively, $w(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ quantifies the *dissimilarity* between the two selected arms i and j that tells how

Algorithm 1 SED

Input: Confidence $\delta \in (0,1)$

1: $\mathcal{S} \leftarrow \{\{i,j\} \mid i,j \in [\![K]\!], i \neq j\}$
2: $t \leftarrow 1$
3: Pull each arm $i \in [\![K]\!]$ once
4: Compute the confidence sets $\mathcal{C}_{i,t}^{\delta}$ for each $i \in [\![K]\!]$
5: **while** $|\mathcal{S}| > 1$ **do**
6: $\quad \widetilde{w}_t \leftarrow \max_{\{i,j\} \in \mathcal{S}} \min_{\widetilde{\boldsymbol{\mu}}_i \in \mathcal{C}_{i,t}^{\delta}, \widetilde{\boldsymbol{\mu}}_j \in \mathcal{C}_{j,t}^{\delta}} w(\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\mu}}_j)$
7: $\quad \mathcal{S} \leftarrow \mathcal{S} \setminus \left\{ \{i,j\} \in \mathcal{S} : \max_{\widetilde{\boldsymbol{\mu}}_i \in \mathcal{C}_{i,t}^{\delta}, \widetilde{\boldsymbol{\mu}}_j \in \mathcal{C}_{j,t}^{\delta}} w(\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\mu}}_j) \leqslant \widetilde{w}_t \right\}$
8: $\quad \mathcal{K}_t \leftarrow \{i \in [\![K]\!] : \exists \{h,j\} \in \mathcal{S} \wedge i \in \{h,j\}\}$
9: $\quad$ Pull each arm in $i \in \mathcal{K}_t$ once
10: $\quad t \leftarrow t+1$
11: $\quad$ Update the confidence sets $\mathcal{C}_{i,t}^{\delta}$ for each $i \in [\![K]\!]$
12: **end while**
13: **return** $\{\widehat{I}, \widehat{J}\} \in \mathcal{S}$

much they are far apart, and, consequently, $\Delta_{\{i,j\}}$ describes how much the arm pair $\{i,j\}$ is far from the optimal arm pair in terms of dissimilarity.

Best-Arm Identification Framework Let $\mathcal{F}_{t-1} = \sigma(I_1, \boldsymbol{x}_1, \dots, I_{t-1}, \boldsymbol{x}_{t-1}, I_t)$ be the σ-algebra generated by the observations up to round $t-1$. A BAI strategy for dissimilarity bandits is defined by: (i) a sampling rule $(I_t)_{t \in \mathbb{N}}$, where I_t is $\mathcal{F}_{t-1}$-measurable, telling the learner which arm to pull at round t; (ii) a stopping rule τ, which is a stopping time w.r.t. $\mathcal{F}_t$, telling the learner when to stop the learning procedure; and (iii) a recommendation rule $\{\widehat{I}_\tau, \widehat{J}_\tau\}$ that is $\mathcal{F}_\tau$-measurable, providing a guess on the optimal pair at round τ. In the *fixed-confidence* setting, given a confidence threshold $\delta \in (0,1)$, we want to minimize the *sample complexity* τ, while guaranteeing that the probability of recommending a sub-optimal pair is bounded by δ, formally we want to find an upper bound $\overline{\tau}$ over τ such that:

$$\mathbb{P}\left(\{\widehat{I}_\tau, \widehat{J}_\tau\} = \{i^*, j^*\} \wedge \tau \leqslant \overline{\tau}\right) \geqslant 1-\delta. \qquad (1)$$

3 ALGORITHM

This section proposes a novel algorithm, `Successive Elimination for Dissimilarity` (`SED`), that employs a successive elimination procedure to identify the optimal pair $\{i^*, j^*\}$ with high probability.

The pseudocode for the proposed algorithm is presented in Algorithm 1 and requires the confidence $\delta \in (0,1)$ as input. At first, it initializes the set of admissible arm pairs $\mathcal{S}$ with all pairs $\{i,j\}$, with $i,j \in [\![K]\!]$ and $i \neq j$, sets the phase counter t, and pulls each arm $i \in [\![K]\!]$ once (Lines 1-3). The algorithm assumes to have access to a way of computing confidence sets $\mathcal{C}_{i,t}^{\delta}$ for the expected observation vectors $\boldsymbol{\mu}_i$ holding with high probability $1-\delta$, for all arms $i \in [\![K]\!]$ and uniformly over the phases $t \in \mathbb{N}$ (Line 4). Formally:

$$\mathbb{P}\left(\forall t \in \mathbb{N},\, \forall i \in [\![K]\!] : \boldsymbol{\mu}_i \in \mathcal{C}_{i,t}^{\delta}\right) \geqslant 1-\delta. \qquad (2)$$

The specific form of the confidence sets $\mathcal{C}_{i,t}^{\delta}$ depends on the estimator $\widehat{\boldsymbol{\mu}}_{i,t}$ for the expected observation vectors $\boldsymbol{\mu}_i$. In Section 3.1, we provide an explicit form for the case in which the estimator $\widehat{\boldsymbol{\mu}}_{i,t}$ is the sample mean.

During the learning process, if the set of admissible arm pairs is non-singleton, i.e., $|\mathcal{S}| > 1$ (Line 5), `SED` computes $\widetilde{w}_t$ representing the maximum lower bound over the dissimilarity between the pair of arms compatible with the current confidence sets (Line 6). Notice that $\widetilde{w}_t$ is selected as the minimum value of the dissimilarity among the observation vectors contained in the confidence sets of the arms i and j. This lower bound is employed to determine if any arm pair can be excluded from the set of admissible pairs $\mathcal{S}$ (Line 7). Indeed, if the maximum dissimilarity computed over vectors in the confidence sets of the arms i and j is lower than the $\widetilde{w}_t$, with high probability, the pair $\{i,j\}$ cannot be the optimal one and can be excluded from the search.

Finally, the algorithm pulls all the arms that are still present in at least one pair in the set of admissible arms $\mathcal{S}$, i.e., those belonging to the set $\mathcal{K}_t$ (Lines 9-8), updates the phase count (Line 10), and the confidence sets to be evaluated in the next phase (Line 11). We remark that the set $\mathcal{K}_t$ will become smaller over time, i.e., when an arm is no longer contained in any pair of the admissible set $\mathcal{S}$.

Algorithm 1 has been intentionally presented without specifying the form of the confidence sets $\mathcal{C}_{i,t}^{\delta}$ and the form of the dissimilarity function w. The choice of these elements has a crucial impact on the computational and statistical properties of the algorithm. Indeed, Lines 6 and 7 require solving maximization and minimization problems involving both $\mathcal{C}_{i,t}^{\delta}$ (as domains) and w (as the objective function) which might become computationally intractable.[1] In the next section, we show that for a particular (still notable) subclass of dissimilarity functions w and for the sample mean as an estimator, these optimization problems can be tackled in a computationally efficient way.

3.1 Efficient Implementation

In the following, we provide a computationally-efficient implementation of Algorithm 1 which makes use of the

[1] For instance, when $w(\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\mu}}_j) = \|\widetilde{\boldsymbol{\mu}}_i - \widetilde{\boldsymbol{\mu}}_j\|_2$ is the Euclidean norm of the difference of vectors, $\mathcal{C}_{i,t}^{\delta} = \{\mathbf{0}\}$, and $\mathcal{C}_{j,t}^{\delta}$ is a polytope, Line 7 reduces to computing the Euclidean diameter of $\mathcal{C}_{j,t}^{\delta}$, which is known to be an NP-hard problem (Brieden, 2002).

sample mean $\widehat{\boldsymbol{\mu}}_{i,t}$ for estimating the expected observation vectors $\boldsymbol{\mu}_i$ and the dissimilarity functions w that can be expressed by means of seminorms.

Sample Mean Estimator Concentration To estimate the expected observation vectors $\boldsymbol{\mu}_i$, we use the sample mean of the random observed vectors:

$$\widehat{\boldsymbol{\mu}}_{i,t} \coloneqq \frac{1}{t} \sum_{l \in [\![t]\!]} \boldsymbol{x}_{i,l}, \tag{3}$$

where $\boldsymbol{x}_{i,l}$ is the vector observed when pulling arm $i \in [\![K]\!]$ in the t-th phase. The following result shows that, under the assumption that the components of the observation vector are independent and σ^2-subgaussian (Assumption 2.1), the sample mean enjoys a desirable concentration rate.

Lemma 3.1 (Sample Mean Concentration)**.** *Let $\delta \in (0,1)$. Under Assumption 2.1, the sample mean $\widehat{\boldsymbol{\mu}}_{i,t}$ in Equation (3) fulfills Equation (2) with:*

$$\mathcal{C}_{i,t}^{\delta} \coloneqq \left\{ \widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d : \|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_{i,t}\|_2 \leqslant 4\sqrt{\frac{\sigma^2 \max\left\{d, \log \frac{2Kt^2}{\delta}\right\}}{t}} \right\}.$$

The result is derived by applying the Bernstein's inequality and a union bound over all arms and time instants. It is worth noting that under Assumption 2.1 (thanks to the independence of the components x_j of the observation vector), the straightforward sample mean displays optimal concentration rate (Lugosi and Mendelson, 2019). To achieve the same optimal concentration rate when relaxing the component independence assumption, it is well-known that more complex estimators are needed which come with a more computationally demanding procedure (e.g., *median of means*, Lugosi and Mendelson, 2019).

Seminorm Dissimilarity Function To make the optimization problems at Lines 6 and 7 of Algorithm 1 computationally tractable, we restrict the class of the dissimilarity functions w to those that can be expressed as the seminorm of the difference of the expected observation vectors. This requirement is formalized in the following assumption.

Assumption 3.1 (Seminorm Dissimilarity Function w)**.** *Let $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}_{\geqslant 0}$ be a seminorm, i.e., it fulfills for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$:*

- *(Subadditivity) $\|\boldsymbol{x} + \boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$;*
- *(Absolute homogeneity) $\|\alpha \boldsymbol{x}\| = |\alpha| \|\boldsymbol{x}\|$.*

The dissimilarity function w is s.t. for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$:

$$w(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{4}$$

It is worth noting that Assumption 3.1 allows to comfortably embed a large class of dissimilarity functions, including the widely used p-norms. This assumption, in combination with the concentration result of Lemma 3.1, allows replacing Lines 6 and 7 of Algorithm 1 with a condition that can be evaluated efficiently, as shown in the following lemma.

Lemma 3.2 (Elimination with Seminorms)**.** *Let $\delta \in (0,1)$, $t \in \mathbb{N}$, and let $\partial \mathcal{B}^{d-1}$ be the surface of the unit sphere in $\mathbb{R}^d$. Let us define:*

$$U_t^{\delta} \coloneqq 4 \left(\max_{\boldsymbol{x} \in \partial \mathcal{B}^{d-1}} \|\boldsymbol{x}\| \right) \cdot \sqrt{\frac{\sigma^2 \max\left\{d, \log \frac{2Kt^2}{\delta}\right\}}{t}}. \tag{5}$$

Then, using the sample mean as estimator and under Assumptions 2.1 and 3.1, if the arm pair $\{i,j\}$, with $i,j \in [\![K]\!]$ and $i \neq j$, fulfills:

$$w(\widehat{\boldsymbol{\mu}}_{i,t}, \widehat{\boldsymbol{\mu}}_{j,t}) + 2U_t^{\delta} \leqslant \max_{\{i',j'\} \in \mathcal{S}} w(\widehat{\boldsymbol{\mu}}_{i',t}, \widehat{\boldsymbol{\mu}}_{j',t}) - 2U_t^{\delta}, \tag{6}$$

then, $\{i,j\} \notin \mathcal{S}$.

Thus, Lemma 3.2 provides a sufficient condition for eliminating the candidate pair $\{i,j\}$. Indeed, if the condition of Equation (6) is satisfied, Algorithm 1 excludes pair $\{i,j\}$ from the set of admissible pairs $\mathcal{S}$. Thus, if we replace Lines 6 and 7 of Algorithm 1 with the elimination condition in Equation (6), we are guaranteed not to discard potentially optimal pairs at the price, possibly, of postponing their elimination. Indeed, this drawback is compensated by obtaining a condition that can be evaluated more efficiently (i.e., with linear time in the cost needed to evaluate the seminorm).

3.2 Sample Complexity Analysis

We are now ready to provide the analysis of the sample complexity of `SED`, i.e., Algorithm 1 instanced with the sample mean as the estimator for the expected observation vectors. To this end, we construct a suitable complexity index in which each arm $i \in [\![K]\!]$ contributes with the minimum sub-optimality gap $\Delta_{\{i,j\}}$ in which arm i appears, i.e.:

$$\Delta_i^* \coloneqq \min_{j \in [\![K]\!] \setminus \{i\}} \Delta_{\{i,j\}} = w^* - \max_{j \in [\![K]\!] \setminus \{i\}} w(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j). \tag{7}$$

Similarly to traditional BAI (Kaufmann et al., 2016), we characterize the complexity of the identification problem with an appropriate *complexity index*:

$$HP \coloneqq \sum_{i \in [\![K]\!] \setminus \{i^*, j^*\}} \frac{\sigma^2}{(\Delta_i^*)^2}. \tag{8}$$

The following result provides a high-probability upper bound to the sample complexity of the `SED` algorithm.

Theorem 3.3. *If the estimator is the sample mean and under Assumptions 2.1 and 3.1, with probability at least $1-\delta$, `SED` returns the optimal pair $\{i^*, j^*\}$ with a sample complexity bounded by:*

$$\tau \leqslant O\left(\sum_{i \in [\![K]\!] \setminus \{i^*, j^*\}} \frac{\sigma^2}{(\Delta_i^*)^2} \max\left\{ d, \log\left(\frac{\sigma^2}{(\Delta_i^*)^2} \sqrt{\frac{K}{\delta}} \right) \right.\right.$$

$$+ \log\log\left(\frac{\sigma^2}{(\Delta_i^*)^2}\sqrt{\frac{K}{\delta}}\right)\Bigg\}\Bigg).$$

Thus, we observe that, apart from logarithmic terms, we can bound the sample complexity as follows:

$$\tau \leqslant \widetilde{O}\left(HP \cdot \max\left\{d, \log\left(\frac{1}{\delta}\right)\right\}\right). \tag{9}$$

We would like to point out that these results hold even without Assumption 2.1 if appropriate mean estimators are used. Moreover, it is worth remarking that the complexity index *HP* involves *one term for every arm* and not *one term for every pair of arms*. This is a notable improvement with respect to a full search on the space of the arm pairs, which has cardinality $O(K^2)$. This property can be explained by the fact that SED exploits the structure of the problem by reusing the samples collected from each arm $i \in [\![K]\!]$ for computing the dissimilarity for all the pairs $\{i, j\}$ in which arm i is involved. Thus, arm i will continue to be pulled as long as at least one pair of the form $\{i, j\}$ is considered admissible, i.e., it belongs to set $\mathcal{S}$. This is also the reason why the complexity is governed by the minimum sub-optimality gaps Δ_i^* that determines the last round in which arm i will be pulled by SED.

4 LOWER BOUND

In this section, we derive a lower bound to the expected sample complexity that any algorithm that outputs the optimal pair $\{i^*, j^*\}$ with high probability satisfies in the case the dissimilarity function is a seminorm.

Theorem 4.1. *There exists a class of Gaussian dissimilarity bandits fulfilling Assumptions 2.1 and 3.1 such that for any fixed-confidence BAI algorithm that fulfills* $\mathbb{P}(\{\widehat{I}_\tau, \widehat{J}_\tau\} = \{i^*, j^*\}) \geqslant 1 - \delta$*, there exists a Gaussian dissimilarity bandit* $\underline{\boldsymbol{\nu}}$ *such that:*

$$\mathbb{E}_{\underline{\boldsymbol{\nu}}}[\tau] \geqslant \Omega\left(\sum_{i \in [\![K]\!] \setminus \{i^*, j^*\}} \frac{\sigma^2}{(\Delta_i^*)^2}\log\left(\frac{1}{\delta}\right)\right) \tag{10}$$

$$= \Omega\left(HP \cdot \log\left(\frac{1}{\delta}\right)\right). \tag{11}$$

The construction of the lower bound follows the well-established construction for the BAI problem and the change of measure arguments presented by Kaufmann et al. (2016). Since the involved distributions are over vectors, the main technical challenge consists in finding the appropriate direction along which to alter the expected observation vectors when constructing the alternative instance.

Comparing the lower bound of Theorem 4.1 with the sample complexity upper bound holding for the SED algorithm, we observe the same dependence on the complexity index *HP* and on the confidence term $\log(1/\delta)$. However, the upper bound presents a dependence on d which might become significant for a large value of δ. Nevertheless, the lower bound containing the same complexity term *HP* suggests that our algorithm SED is effectively addressing the challenges of the BAI problem for the dissimilarity bandits.

5 DISCUSSION

In this section, we elaborate on particular instances of the BAI problem for dissimilarity bandits and discuss whether they can be addressed with standard BAI in MABs and we show the corresponding sample complexity results. Table 1 summarizes all the results.

5.1 One-Dimensional Case

When the observation vector is one-dimensional (i.e., $d = 1$) and $w(\mu_i, \mu_j) = |\mu_i - \mu_j|$ is the absolute value of the difference, the problem of finding the most dissimilar arm pair can be reduced to standard fixed-confidence BAI in suitably defined MABs. Specifically, we can adopt two approaches.

BAI on Pairs We map this setting to a standard MAB over the pairs of arms. Let us define the MAB having $[\![K]\!] \times [\![K]\!]$ as the arm set and with the expected rewards defined as $\widetilde{\mu}_{(i,j)} := \mu_i - \mu_j$. In this specific case, we have that $\max_{(i,j)\in[\![K]\!]\times[\![K]\!]} \widetilde{\mu}_{(i,j)} = \max_{(i,j)\in[\![K]\!]\times[\![K]\!]} |\mu_i - \mu_j| = w^*$. Notice that using such a modeling strategy, the arm set has been enlarged by a factor of K; therefore, we can expect the BAI procedure to become more complex. Indeed, let us define $\widetilde{\Delta}_{(i,j)} := w^* - \widetilde{\mu}_{(i,j)}$, a standard Successive Elimination (SE, Even-Dar et al., 2002) analysis leads to:

$$\tau \leqslant \widetilde{O}\left(\sum_{\{i,j\} \neq \{i^*, j^*\}} \frac{\sigma^2}{\widetilde{\Delta}_{(i,j)}^2}\log\left(\frac{1}{\delta}\right)\right). \tag{12}$$

It is worth noting that $\widetilde{\Delta}_{(i,j)} \geqslant \Delta_{\{i,j\}}$ since, by definition $\widetilde{\mu}_{(i,j)} = \mu_i - \mu_j \leqslant |\mu_i - \mu_j| = w(\mu_i, \mu_j)$. Nevertheless, the summation over the pairs in Equation (12) will contain (but is not limited to) the terms Δ_i^* of Equation (7) and, therefore, for sufficiently small δ and apart from constants, the sample complexity of the *BAI on Pairs* approach is larger than that of SED.

MaxBAI + MinBAI Alternatively, we can decompose our problem into two classical BAI ones, in which the goal is to identify the arms with the maximum and the minimum expected reward, respectively. Indeed, we can alternate the identification of the two and recommend, as an outcome, the pair composed of the

Approach	1-*dimensional*	*d-dimensional*
SED (**ours**)	$\boldsymbol{HP} \cdot \mathbf{log}\left(\frac{\mathbf{1}}{\boldsymbol{\delta}}\right)$	$\boldsymbol{HP} \cdot \mathbf{max}\left\{\boldsymbol{d}, \mathbf{log}\left(\frac{\mathbf{1}}{\boldsymbol{\delta}}\right)\right\}$
BAI on Pairs	$\sum_{\{i,j\}\neq\{i*,j*\}} \frac{\sigma^2}{\widetilde{\Delta}^2_{(i,j)}} \log\left(\frac{1}{\delta}\right)$	Not applicable
MaxBAI+MinBAI	$\left(\sum_{\boldsymbol{i:\Delta_i^+>0}} \frac{\boldsymbol{\sigma^2}}{\boldsymbol{(\Delta_i^+)^2}} + \sum_{\boldsymbol{i:\Delta_i^->0}} \frac{\boldsymbol{\sigma^2}}{\boldsymbol{(\Delta_i^-)^2}}\right)\mathbf{log}\left(\frac{\mathbf{1}}{\boldsymbol{\delta}}\right)$	Not applicable

Table 1: Orders (in terms of $\widetilde{O}(\cdot)$) of the upper bounds on the stopping times of different approaches to solving the BAI dissimilarity bandits problem. The best orders of complexity are reported in boldface.

arms provided as guesses by the two identification procedures. Formally, we use the fact that $w^* = \mu^+ - \mu^-$, where $\mu^+ \coloneqq \max_{i\in[\![K]\!]} \mu_i$ and $\mu^- \coloneqq \min_{j\in[\![K]\!]} \mu_j$. Let us define $\Delta_i^+ \coloneqq \mu^+ - \mu_i$ and $\Delta_i^- \coloneqq \mu_i - \mu^-$, the analysis of this double SE algorithm leads to:

$$\tau \leqslant \widetilde{O}\left(\left(\sum_{i:\Delta_i^+>0} \frac{\sigma^2}{(\Delta_i^+)^2} + \sum_{i:\Delta_i^->0} \frac{\sigma^2}{(\Delta_i^-)^2}\right)\log\left(\frac{1}{\delta}\right)\right). \quad (13)$$

We observe that, remarkably, for sufficiently small δ and apart from constant terms, the sample complexity in Equation (13) has the same order as the sample complexity of SED (Theorem 3.3). Indeed, we have:

$$\begin{aligned}\Delta_i^* &= \mu^+ - \mu^- - \max_{j\in[\![K]\!]\setminus\{i\}} |\mu_i - \mu_j| \\ &= \min\{\mu^+ - \mu^- - (\mu_i - \mu^-), \mu^+ - \mu^- - (\mu^+ - \mu_i)\} \\ &= \min\{\Delta_i^+, \Delta_i^-\}.\end{aligned}$$

It follows that the *MaxBAI + MinBAI*, from a theoretical perspective, displays performance comparable to our SED.

5.2 Multi-Dimensional Case

The more challenging case of d-dimensional observation vectors cannot be addressed using the two abovementioned approaches. Indeed, both are not viable since the involved observation vectors would be d-dimensional on which there is no clear definition of maximum and minimum, preventing the use of both the *BAI on Pairs* and *MaxBAI + MinBAI*.

Another approach would be to design a MAB over the pair of arms in which the learner selects a pair of arms $\{i, j\}$ to be pulled, obtains the pair of observation vectors $\{\boldsymbol{x}, \boldsymbol{y}\}$ and uses them to directly compute the dissimilarity function $w(\boldsymbol{x}, \boldsymbol{y})$. This term is regarded as a surrogate of the dissimilarity over the expected observation vectors $w(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ and employed as the reward in the BAI procedure. However, this approach would introduce a bias in estimating the dissimilarities. Indeed, the expected dissimilarity $\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\nu}_i, \boldsymbol{y}\sim\boldsymbol{\nu}_j}[w(\boldsymbol{x}, \boldsymbol{y})]$ does not correspond to the dissimilarity of the expected observation vectors $w(\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\nu}_i}[\boldsymbol{x}], \mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{\nu}_j}[\boldsymbol{y}]) = w(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ in general. Thus, using such an approach with a SE algorithm may fail to identify the optimal arm pair.

Example 5.1. *Let us consider a 1-dimensional 3-armed Gaussian dissimilarity bandit with expected observations $\underline{\mu} = (0, 1/2, 1)^\top$, variances $\underline{\sigma}^2 = (0, b^2, 0)^\top$, and $w(x, y) = |x-y|$ as dissimilarity. The optimal arm pair is $\{i^*, j^*\} = \{1, 3\}$. However, if we consider the expectation of the dissimilarity of the observations, we obtain:*

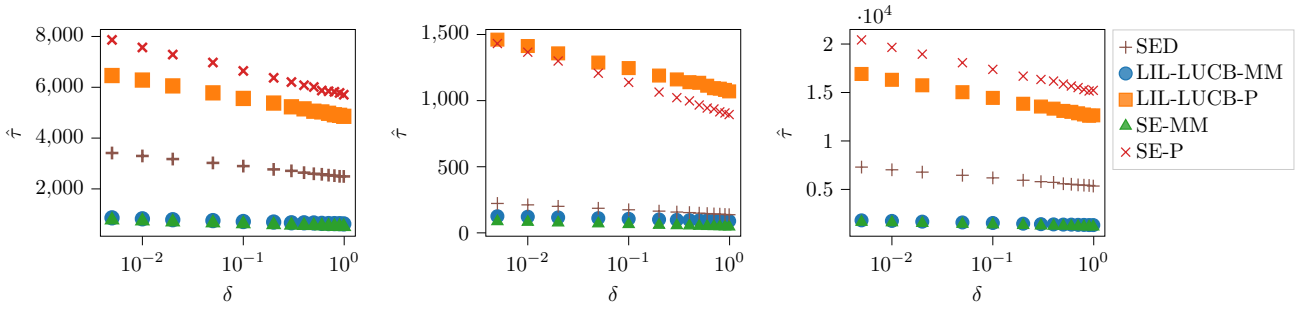
$$\mathbb{E}_{x\sim\nu_1, y\sim\nu_2}[w(x, y)] = \mathbb{E}_{x\sim\nu_2, y\sim\nu_3}[w(x, y)] \geqslant \sqrt{2/\pi} e^{-\frac{1}{8b^2}} b,$$
$$\mathbb{E}_{x\sim\nu_1, y\sim\nu_3}[w(x, y)] = 1.$$

Thus, we can make the first expression arbitrarily large by setting the value of b, leading this approach to wrongly believe that the optimal pairs are $\{1, 2\}$ and $\{2, 3\}$.

6 RELATED WORKS

Combinatorial Bandits The setting we analyze in this work is related to that of Combinatorial MAB (CMAB), defined originally for the regret minimization task by Chen et al. (2013); Cesa-Bianchi and Lugosi (2012) and, successively, extended to BAI (Chen et al., 2014). In such a setting, the learner is allowed to select a subset of the available arms (a.k.a. superarm), and a set of constraints determines the formation of the subset. This setting could be, in principle, adapted to our problem by using the pair of arms as a superarm in the CMAB setting and carrying out the learning using the BAI algorithm designed for such a scenario (e.g., Gabillon et al., 2016; Chen et al., 2016a; Rejwan and Mansour, 2020; Jourdan et al., 2021).

However, such works usually require more demanding assumptions that do not allow their direct application in our setting. Indeed, CMAB works require either that the constraints enforced on the selection of the arms for the superarm definition are matroidal (Chen et al.,



(a) Env. #1 (b) Env. #2 (c) Env. #3

Figure 1: Results for the 1-dimensional environments.

2016a; Jourdan et al., 2021), or that the superarm's reward is a linear combination of the arms' reward (Chen et al., 2014; Gabillon et al., 2016; Öner et al., 2018; Rejwan and Mansour, 2020), or, more generally, that the superarm's reward is monotone in the arms' reward with respect to the partial order of vectors (Chen et al., 2016b; Huang et al., 2017). However, none of the above can be adequately applied to our setting. The fact that we are only considering pairs (and, for instance, we cannot have single arms as superarms) makes the nature of our constraints non-matroidal. On the other hand, adopting a dissimilarity function, which can be interpreted as an indicator of how far two arms are from each other, as the superarm's reward can break the monotonicity assumption. We note that this is not the case when the reward of each arm is a scalar, as the only reasonable dissimilarity function between two arms would be the absolute difference of their expected observations, which can be reduced to a valid CMAB setting by creating a copy of each arm with a negative reward. Indeed, this is equivalent to identifying the minimum and maximum arms adopting standard BAI algorithms. While we compare our results with such methods for the one-dimensional case, we mainly focus on the case where the arms' observations are multi-dimensional, where there is no analog modeling of minimum and maximum problems independently, and the monotonicity assumption does not hold even with the simple choice of the Euclidean distance as the dissimilarity function.

Environmental Monitoring Regarding the application that inspired this work, i.e., environmental monitoring, bandit works have been applied and provided significant results for the environmental field. In particular, a BAI algorithm to identify the maximum concentration of a contaminant in potable water has been designed by Gabrielli et al. (2022, 2024). This work focuses on the exploitation of the temporal dependency of the arms in this setting, and, being of a practical nature, it provides no theoretical result on the sample complexity of the algorithm therein defined. Therefore, it cannot be directly compared with what we will present. Another work, by Martin and Johnson (2020), focused on the regret minimization task and aimed at assessing the performances of classical MAB techniques for designing *smart* sampling techniques. This work showed that applying adaptive sampling strategies can outperform traditional equal probability allocation strategies. Even in this case, the nature of the work was purely experimental, with no novel insights into theoretical results or new algorithms.

7 NUMERICAL SIMULATIONS AND REAL-WORLD EXAMPLE

In this section, we present numerical simulations to complement the theoretical results we provided in the previous sections. At first, we compare the `SED` algorithm in the scenario of 1-dimensional observation vectors, in which we can provide significant and strong baselines. Subsequently, we address the d-dimensional case, and, finally, we provide an example of the performance of `SED` on a simulated potable water scenario.[2]

7.1 1-dimensional Case

In the following, we compare SED with the performance of two BAI algorithms: SE (Even-Dar et al., 2002) and LIL-LUCB (Jamieson et al., 2014) in the 1-dimensional setting using as similarity $w(\mu_i, \mu_j) = |\mu_i - \mu_j|$. For each of the above baseline methods, we applied one of the modeling approaches we mentioned before, i.e., *BAI on Pairs* and *MaxBAI+MinBAI*. We will denote the two approaches using the suffixes -P and -MM, respectively.

We experimented on synthetically generated bandit

[2]Details about the numerical simulations are reported in Appendix B. The code used for the experimental section is available at `https://github.com/paolob2/sed`.

environments with $K = 10$ Gaussian arms having uniform variance $\sigma^2 = 0.01$. We considered the following three scenarios differing for the arm mean rewards:

- Env. #1: $\mu_i = 0.1i \quad \forall i \in [\![K]\!]$;
- Env. #2: $\mu_1 = 1$, $\mu_K = 0$, $\mu_i = 0.5 \quad \forall i \in [\![K]\!] \backslash \{1, K\}$;
- Env. #3: $\mu_1 = 1, \mu_K = 0$, $\mu_i = 0.9 \quad \forall i : 1 < i \leqslant \frac{K}{2}$, and $\mu_i = 0.1 \quad \forall i : \frac{K}{2} < i < K$.

Notice that the complexity indexes of the three environments are $HP \simeq 2.85$, $HP = 0.32$, and $HP = 8$, respectively, which makes Env #3 the one requiring (most likely) the largest number of samples to reach the stopping time.

We tested the considered algorithms over different confidence values, specifically $\delta \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, \ldots, 0.9, 0.99\}$. In the following, we report the average stopping time $\hat{\tau}$ of the different algorithms, averaged over 1000 independent runs.[3]

Results The results are displayed in Figure 1. First, let us notice that despite the guarantees being in high probability, all the analyzed algorithms could always identify the optimal pair of arms. The proportional dependence of the stopping time $\hat{\tau}$ on $\log(1/\delta)$ is apparent for all the algorithms, confirming the theoretical results. Similarly, the algorithms are faster in detecting the optimal pair as the complexity term HP gets smaller. In all the examined scenarios, the -MM approach provides significantly better results, improving the stopping time of a factor at least $\times 7$ than the corresponding -P one. However, this approach is not viable in $d > 1$. The second best option after the -MM approach in all three environments is the `SED` approach, which differs by a factor of approximately $\times 3{-}4$ across all environments. We think the information provided by the dissimilarity used in `SED` can only partially close the gap in performances w.r.t. the -MM, specifically crafted for this setting.

7.2 d-Dimensional Case

The comparison of the d-dimensional case is more complex due to the lack of solid baselines, as discussed earlier. We select as dissimilarity $w(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2$ the Euclidean norm of the vector difference and synthetically generate bandit environments with $K = 10$ arms, where each arm is a vector of $d = 16$ independent Gaussian variables having uniform variance $\sigma^2 = 0.01$. The arms expected observation vector components $\mu_{i,j}$ were randomly generated in the range $(0, 1)$ in such a way that $\min_{i \in [\![K]\!]} \Delta_i^* \geqslant 0.1$.

We compare the `SED` with the *BAI on Pairs* approach for the multi-dimensional case, as discussed in Section 5.2, i.e., by computing the dissimilarity over the observation vectors $\|\boldsymbol{x} - \boldsymbol{y}\|_2$, where $\boldsymbol{x} \sim \boldsymbol{\nu}_i$ and $\boldsymbol{y} \sim \boldsymbol{\nu}_j$.[4] We applied this approach to the LIL-LUCB and the SE algorithms. We tested the above algorithms for different confidences $\delta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ The results show the average stopping time $\hat{\tau}$ over 1000 independent runs and the corresponding 90% confidence intervals as vertical bars.

Results Figure 2a shows that the proposed `SED` approach can deliver the correct answer with a sample complexity of ≈ 2 order of magnitude smaller than the -P approach. We recall that the -P approach introduces a bias in the estimation (Section 5.2) and this may influence their performances. Overall, the results are in line with what has been observed in the 1-dimensional case and strengthen the idea that what we proposed outperforms the currently available algorithms for the dissimilarity bandit setting.

7.3 Environmental Monitoring

Finally, we look at a more practical multi-dimensional setting, where each arm represents the average fluorescence response output at a certain hour of the day. This output can be represented as a grey-scale image, which we resize down to 4×4 ($d = 16$) and where we normalize the values in the $[0, 1]$ range. The environment was generated from the dataset in Gabrielli et al. (2021) as follows. First, we identified a suitable 30-day period (June 2019) and used arms corresponding to even hours of the day ($K = 12$). For each vector element of each arm, we compute the sample mean and variance of the corresponding pixel. Then, each vector element is treated as a Gaussian variable with the computed mean and a common variance equal to the 90% percentile of all the variance estimators (i.e., $\sigma \simeq 0.004$). The resulting value for the complexity index is $HP \simeq 3.98$. We tested for $\delta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$, running 100 runs for each value. We used the same algorithms of the d-dimensional case and a Naive-RR approach equivalent to a uniform sampling strategy (as described in Martin and Johnson (2020)).

Results The results are reported in Figure 2b. Even for the realistic setting, the results are in line with the ones provided before. In this case, the `SED` algorithm provides stopping times that are $\approx 60\%$ smaller than those of the other approaches. This shows empirical evidence of the savings in terms of samples on an environmental monitoring campaign.

[3] We also report the 90% confidence intervals as bars. However, they are hard to spot in the pictures since their width is ≈ 10.

[4] We approximate them with $2d\sigma^2$ subgaussians.

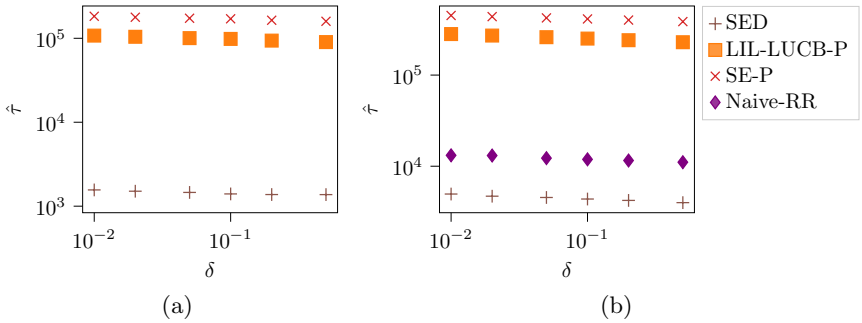


Figure 2: Results for (a) the 16-dimensional environments and (b) the realistic environmental monitoring case.

8 CONCLUSION

In this paper, motivated by real-world environmental monitoring applications, we have introduced the novel setting of the *dissimilarity bandits*. On top of it, we have formulated the fixed-confidence BAI problem of finding the pair of the most dissimilar arms. For this problem, we have presented a novel algorithm `Successive Elimination for Dissimilarity` that, under the assumption that the dissimilarity can be expressed through a seminorm and using the sample mean as an estimator, enjoys desirable computational and statistical properties. Specifically, we have provided a sample complexity analysis that highlights the challenges of the identification problem using a novel complexity index $H\!P$. Furthermore, we have derived a sample complexity lower bound almost matched by our `SED`. Finally, we have conducted an experimental campaign on both synthetic and real-world domains showing the advantages of the proposed method over the considered baselines, especially in the multi-dimensional case.

Future works include the extensions of the proposed approach and the corresponding analysis to different BAI strategies, such as lower-upper approaches and track and stop strategies, as well as the possibility of generalizing the analysis for estimators other than the sample mean. Concerning the dissimilarity functions, further investigations should include the case in which particular known functions are considered (for obtaining more effective algorithms) and the challenging scenario in which the dissimilarity is not known but has to be learned from some environmental feedback. Lastly, we consider the possibility of relaxing the independence assumption between arms.

ACKNOWLEDGMENTS

This paper is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in neural information processing systems (NeurIPS)*, pages 2312–2320.

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010). Best arm identification in multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 41–53.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256.

Brieden (2002). Geometric optimization problems likely not contained in a p x. *Discrete & Computational Geometry*, 28:201–209.

Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422.

Chen, L., Gupta, A., and Li, J. (2016a). Pure exploration of multi-armed bandit under matroid constraints. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 49, pages 647–669.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. (2014). Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems (NeurIPS)*, 27.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. (2016b). Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems*, 29.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 151–159.

Even-Dar, E., Mannor, S., and Mansour, Y. (2002). Pac bounds for multi-armed bandit and markov de-

cision processes. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 255–270.

Gabillon, V., Lazaric, A., Ghavamzadeh, M., Ortner, R., and Bartlett, P. (2016). Improved learning complexity in combinatorial pure exploration bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51, pages 1004–1012.

Gabrielli, M., Antonelli, M., and Trovò, F. (2024). Adapting bandit algorithms for settings with sequentially available arms. *Engineering Applications of Artificial Intelligence*, 131:107815.

Gabrielli, M., Trovò, F., and Antonelli, M. (2022). Automatic optimization of temporal monitoring schemes dealing with daily water contaminant concentration patterns. *Environmental Science: Water Research & Technology*, 8(10):2099–2113.

Gabrielli, M., Turolla, A., and Antonelli, M. (2021). Bacterial dynamics in drinking water distribution systems and flow cytometry monitoring scheme optimization. *Journal of Environmental Management*, 286:112151.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 49, pages 998–1027.

Honorio, J. and Jaakkola, T. (2014). Tight bounds for the expected risk of linear classifiers and pac-bayes finite-sample guarantees. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 384–392.

Huang, R., Ajallooeian, M. M., Szepesvári, C., and Müller, M. (2017). Structured best arm identification with fixed confidence. In *International Conference on Algorithmic Learning Theory*, pages 593–616. PMLR.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 423–439.

Jourdan, M., Mutný, M., Kirschner, J., and Krause, A. (2021). Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, volume 132, pages 805–849.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42.

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

Lugosi, G. and Mendelson, S. (2019). Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794.

Martin, D. M. and Johnson, F. A. (2020). A multiarmed bandit approach to adaptive water quality management. *Integrated Environmental Assessment and Management*, 16(6):841–852.

Öner, D., Karakurt, A., Eryılmaz, A., and Tekin, C. (2018). Combinatorial multi-objective multi-armed bandit problem. *arXiv preprint arXiv:1803.04039*.

Rejwan, I. and Mansour, Y. (2020). Top-k combinatorial bandits with full-bandit feedback. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, volume 117, pages 752–776.

Rigollet, P. (2015). 18. s997: High dimensional statistics. *Lecture Notes), Cambridge, MA, USA: MIT Open-CourseWare*.

Rivasplata, O. (2012). Subgaussian random variables: An expository note. *Internet publication, PDF*, 5.

Sui, Y., Zoghi, M., Hofmann, K., and Yue, Y. (2018). Advancements in dueling bandits. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5502–5510.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. (2012). The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, Sections 3 and 4
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, Section 4.1
   (c) (Optional) Anonymized source code, with a specification of all dependencies, including external libraries. Yes, we added the source code to the supplementary material

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, in the statements of Section 4 and 5
   (b) Complete proofs of all theoretical results. Yes, they are present in Appendix A
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, we added the source code to the supplementary material
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, we used vertical bars in the figures, and mentioned if they are not visible due to the fact that they are negligible
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, they are present in Appendix B

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Yes
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Dissimilarity Bandits: Supplementary Material

A Proofs

We start by presenting some definitions and properties of subgaussian random variables that will be needed for the proofs of the subsequent lemmas Rigollet (2015).

- Let X be a σ^2-subgaussian zero-mean random variable. Then, it holds for every $k \geqslant 1$ that Rivasplata (2012):

$$\mathbb{E}[|X|^k] \leqslant k(2\sigma^2)^{k/2}\Gamma\left(\frac{k}{2}\right), \tag{14}$$

  where $\Gamma(\cdot)$ is the Gamma function.

- A zero-mean random variable X is (ξ^2, β)-subexponential if:

$$\mathbb{E}[\exp(\lambda X)] \leqslant \exp\left(\frac{\xi^2\lambda^2}{2}\right), \quad \forall\ |\lambda| \leqslant \frac{1}{\beta}. \tag{15}$$

- Let X be a σ^2-subgaussian zero-mean random variable. Then, it holds that $X^2 - \mathbb{E}[X^2]$ is a $(32\sigma^4, 4\sigma^2)$-subexponential random variable Honorio and Jaakkola (2014).

- Let $X_1, \ldots, X_n$ be n independent (ξ^2, β)-subexponential random variables. Then, $S_n \coloneqq X_1 + \cdots + X_n$ is a $(n\xi^2, \beta)$-subexponential random variable.

- (Bernstein's inequality) Let X be a zero-mean (ξ^2, β)-subexponential random variable. Then, it holds that:

$$\mathbb{P}\left(X > \epsilon\right) \leqslant \exp\left(-\frac{1}{2}\min\left\{\frac{\epsilon^2}{\xi^2}, \frac{\epsilon}{\beta}\right\}\right).$$

Let us start by showing the concentration rate of the sample mean estimator under Assumption 2.1.

Lemma A.1. *Let $\widehat{\boldsymbol{\mu}}_t$ be the sample mean of t d-dimensional i.i.d. random vectors with independent components drawn from a σ^2-subgaussian distribution with expected value $\boldsymbol{\mu}$ (Assumption 2.1). Then, for every $\delta \in (0,1)$ it holds that:*

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\right\|_2 > 4\sqrt{\frac{\sigma^2 \max\left\{d, \log\frac{1}{\delta}\right\}}{t}}\right) \leqslant \delta. \tag{16}$$

Proof. Under Assumption 2.1, each element $\widehat{\mu}_{j,t}$ of the vector $\widehat{\boldsymbol{\mu}}_t$ is an independent $\frac{\sigma^2}{t}$-subgaussian variable.

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\right\|_2 \geqslant \epsilon\right) = \mathbb{P}\left(\sum_{j\in[\![d]\!]}(\widehat{\mu}_{j,t} - \mu_j)^2 \geqslant \epsilon^2\right) \tag{17}$$

$$= \mathbb{P}\left(\sum_{j\in[\![d]\!]}\left((\widehat{\mu}_{j,t} - \mu_j)^2 - \mathbb{E}[(\widehat{\mu}_{j,t} - \mu_j)^2]\right) \geqslant \epsilon^2 - 4d\frac{\sigma^2}{t}\right) \tag{18}$$

$$\leqslant \exp\left(-\frac{1}{2}\min\left\{\frac{\left(\epsilon^2 - 4d\frac{\sigma^2}{t}\right)^2}{64d\frac{\sigma^4}{t^2}}, \frac{\epsilon^2 - 4d\frac{\sigma^2}{t}}{8\frac{\sigma^2}{t}}\right\}\right), \tag{19}$$

where line (18) follows from Equation (14) applied to $\mathbb{E}[(\widehat{\mu}_{j,t} - \mu_j)^2]$ setting $k = 2$. Line (19) is obtained by Equation (16), recalling that $(\widehat{\mu}_{j,t} - \mu_j)^2 - \mathbb{E}[(\widehat{\mu}_{j,t} - \mu_j)^2]$ is a $\left(\frac{32\sigma^4}{t^2}, \frac{4\sigma^2}{t}\right)$-subexponential random variable and,

consequently, $\sum_{j\in[\![d]\!]}\left((\widehat{\mu}_{j,t}-\mu_j)^2-\mathbb{E}[(\widehat{\mu}_{j,t}-\mu_j)^2]\right)$ is a $\left(\frac{32d\sigma^4}{t^2},\frac{4\sigma^2}{t}\right)$-subexponential random variable, recalling that the d components are independent. Then, by applying Bernstein's inequality, under the assumption that $\epsilon^2\geqslant 4d\frac{\sigma^2}{t}$, we obtain the result.

By setting Equation (19) to be equal to δ, we find the appropriate value for ϵ. Let $\lambda\coloneqq\epsilon^2-4d\frac{\sigma^2}{t}$. Then:

$$\exp\left(-\min\left\{\frac{\lambda^2}{64d\frac{\sigma^4}{t^2}},\frac{\lambda}{8\frac{\sigma^2}{t}}\right\}\right)=\delta \tag{20}$$

$$\implies \lambda=\frac{\sigma^2}{t}\max\left\{\sqrt{64d\log\frac{1}{\delta}},8\log\frac{1}{\delta}\right\} \tag{21}$$

$$\implies \epsilon^2=8\frac{\sigma^2}{t}\max\left\{\sqrt{d\log\frac{1}{\delta}},\log\frac{1}{\delta}\right\}+4d\frac{\sigma^2}{t}\leqslant 16\frac{\sigma^2}{t}\max\left\{d,\log\frac{1}{\delta}\right\} \tag{22}$$

$$\implies \epsilon\leqslant 4\sqrt{\frac{\sigma^2\max\left\{d,\log\frac{1}{\delta}\right\}}{t}}, \tag{23}$$

where Equation (22) is obtained by observing that:

$$8\max\left\{\sqrt{d\log\frac{1}{\delta}},\log\frac{1}{\delta}\right\}+4d\leqslant 16\max\left\{\sqrt{d\log\frac{1}{\delta}},d,\log\frac{1}{\delta}\right\}=16\max\left\{d,\log\frac{1}{\delta}\right\},$$

and Equation (21) follows since $\min(g(\lambda),h(\lambda))=z$ is equivalent to $\lambda=\max(g^{-1}(z),h^{-1}(z))$ when g and h are both invertible. □

Lemma 3.1 (Sample Mean Concentration)**.** *Let $\delta\in(0,1)$. Under Assumption 2.1, the sample mean $\widehat{\boldsymbol{\mu}}_{i,t}$ in Equation* (3) *fulfills Equation* (2) *with:*

$$\mathcal{C}_{i,t}^{\delta}\coloneqq\left\{\widetilde{\boldsymbol{\mu}}\in\mathbb{R}^d:\left\|\widetilde{\boldsymbol{\mu}}-\widehat{\boldsymbol{\mu}}_{i,t}\right\|_2\leqslant 4\sqrt{\frac{\sigma^2\max\left\{d,\log\frac{2Kt^2}{\delta}\right\}}{t}}\right\}.$$

Proof. Let us consider the probability of the real mean vector to belong to the defined confidence intervals:

$$\mathbb{P}\left(\exists t\in\mathbb{N},\,\exists i\in[\![K]\!]\,:\,\boldsymbol{\mu}_i\notin\mathcal{C}_{i,t}^{\delta}\right)\leqslant\sum_{t\in\mathbb{N}}\sum_{i\in[\![K]\!]}\mathbb{P}\left(\boldsymbol{\mu}_i\notin\mathcal{C}_{i,t}^{\delta}\right) \tag{24}$$

$$\leqslant\sum_{t\in\mathbb{N}}\sum_{i\in[\![K]\!]}\mathbb{P}\left(\|\boldsymbol{\mu}_{i,j}-\widehat{\boldsymbol{\mu}}_{i,t}\|_2>4\sqrt{\frac{\sigma^2\max\left(d,\log\frac{2Kt^2}{\delta}\right)}{t}}\right) \tag{25}$$

$$\leqslant\sum_{t\in\mathbb{N}}\sum_{i\in[\![K]\!]}\frac{\delta}{2Kt^2}\leqslant\delta, \tag{26}$$

where a union bound over the time instants t and over the arms $[\![K]\!]$ in Equation (25), and the first inequality in Equation (26) follows from Lemma A.1. □

Lemma 3.2 (Elimination with Seminorms)**.** *Let $\delta\in(0,1)$, $t\in\mathbb{N}$, and let $\partial\mathcal{B}^{d-1}$ be the surface of the unit sphere in $\mathbb{R}^d$. Let us define:*

$$U_t^{\delta}\coloneqq 4\left(\max_{\boldsymbol{x}\in\partial\mathcal{B}^{d-1}}\|\boldsymbol{x}\|\right)\cdot\sqrt{\frac{\sigma^2\max\left\{d,\log\frac{2Kt^2}{\delta}\right\}}{t}}. \tag{5}$$

Then, using the sample mean as estimator and under Assumptions 2.1 and 3.1, if the arm pair $\{i,j\}$, with $i,j\in[\![K]\!]$ and $i\neq j$, fulfills:

$$w(\widehat{\boldsymbol{\mu}}_{i,t},\widehat{\boldsymbol{\mu}}_{j,t})+2U_t^{\delta}\leqslant\max_{\{i',j'\}\in\mathcal{S}}w(\widehat{\boldsymbol{\mu}}_{i',t},\widehat{\boldsymbol{\mu}}_{j',t})-2U_t^{\delta}, \tag{6}$$

then, $\{i,j\}\notin\mathcal{S}$.

Proof. Let us start by proving that the confidence set for the 2-norm induces a confidence set over the seminorm. We apply the definition of U_t^δ and the properties of seminorms:

$$\mathcal{C}_{i,t}^\delta = \left\{ \widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d \,:\, \|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_{i,t}\|_2 \leqslant 4\sqrt{\frac{\sigma^2 \max\left\{d, \log\frac{2Kt^2}{\delta}\right\}}{t}} \right\} \tag{27}$$

$$\subseteq \left\{ \widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d \,:\, \|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_{i,t}\| \leqslant 4 \max_{\widetilde{\boldsymbol{\mu}}' \in \mathcal{C}_{i,t}^\delta} \frac{\|\widetilde{\boldsymbol{\mu}}' - \widehat{\boldsymbol{\mu}}_{i,t}\|}{\|\widetilde{\boldsymbol{\mu}}' - \widehat{\boldsymbol{\mu}}_{i,t}\|_2} \sqrt{\frac{\sigma^2 \max\left\{d, \log\frac{2Kt^2}{\delta}\right\}}{t}} \right\} \tag{28}$$

$$= \left\{ \widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d \,:\, \|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_{i,t}\| \leqslant 4 \max_{\widetilde{\boldsymbol{\mu}}' \in \mathcal{C}_{i,t}^\delta} \left\| \frac{\widetilde{\boldsymbol{\mu}}' - \widehat{\boldsymbol{\mu}}_{i,t}}{\|\widetilde{\boldsymbol{\mu}}' - \widehat{\boldsymbol{\mu}}_{i,t}\|_2} \right\| \sqrt{\frac{\sigma^2 \max\left\{d, \log\frac{2Kt^2}{\delta}\right\}}{t}} \right\} \tag{29}$$

$$= \left\{ \widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d \,:\, \|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_{i,t}\| \leqslant 4 \max_{\boldsymbol{x} \in \partial\mathcal{B}^{d-1}} \|\boldsymbol{x}\| \sqrt{\frac{\sigma^2 \max\left\{d, \log\frac{2Kt^2}{\delta}\right\}}{t}} =: U_t^\delta \right\}. \tag{30}$$

where (30) follows from the fact that the vector $\widetilde{\boldsymbol{\mu}}' - \widehat{\boldsymbol{\mu}}_{i,t}$ divided by its L2-norm yields a unit vector.

To obtain the result, it is sufficient to prove the following two inequalities under Assumption 3.1:

$$\max_{\widetilde{\boldsymbol{\mu}}_i \in \mathcal{C}_{i,t}^\delta, \widetilde{\boldsymbol{\mu}}_j \in \mathcal{C}_{j,t}^\delta} w(\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\mu}}_j) \leqslant w(\widehat{\boldsymbol{\mu}}_{i,t}, \widehat{\boldsymbol{\mu}}_{j,t}) + 2U_t^\delta, \tag{31}$$

$$\min_{\widetilde{\boldsymbol{\mu}}_i \in \mathcal{C}_{i,t}^\delta, \widetilde{\boldsymbol{\mu}}_j \in \mathcal{C}_{j,t}^\delta} w(\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\mu}}_j) \geqslant w(\widehat{\boldsymbol{\mu}}_{i,t}, \widehat{\boldsymbol{\mu}}_{j,t}) - 2U_t^\delta. \tag{32}$$

We will only prove the first inequality (proof for the second is analogous). Let $(\bar{\boldsymbol{\mu}}_{i,t}, \bar{\boldsymbol{\mu}}_{j,t})$ be a maximizer of $w(\cdot,\cdot)$ in $\mathcal{C}_{i,t}^\delta \times \mathcal{C}_{j,t}^\delta$. Then:

$$\max_{\widetilde{\boldsymbol{\mu}}_i \in \mathcal{C}_{i,t}^\delta, \widetilde{\boldsymbol{\mu}}_j \in \mathcal{C}_{j,t}^\delta} w(\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\mu}}_j) = w(\bar{\boldsymbol{\mu}}_{i,t}, \bar{\boldsymbol{\mu}}_{j,t}) = \|\bar{\boldsymbol{\mu}}_{i,t} - \bar{\boldsymbol{\mu}}_{j,t}\| \tag{33}$$

$$= \|\bar{\boldsymbol{\mu}}_{i,t} - \bar{\boldsymbol{\mu}}_{j,t} - (\widehat{\boldsymbol{\mu}}_{i,t} - \widehat{\boldsymbol{\mu}}_{j,t}) + (\widehat{\boldsymbol{\mu}}_{i,t} - \widehat{\boldsymbol{\mu}}_{j,t})\| \tag{34}$$

$$= \|(\widehat{\boldsymbol{\mu}}_{i,t} - \widehat{\boldsymbol{\mu}}_{j,t}) + ((\bar{\boldsymbol{\mu}}_{i,t} - \widehat{\boldsymbol{\mu}}_{i,t}) - (\bar{\boldsymbol{\mu}}_{j,t} - \widehat{\boldsymbol{\mu}}_{j,t}))\| \tag{35}$$

$$\leqslant \|\widehat{\boldsymbol{\mu}}_{i,t} - \widehat{\boldsymbol{\mu}}_{j,t}\| + \|(\bar{\boldsymbol{\mu}}_{i,t} - \widehat{\boldsymbol{\mu}}_{i,t}) - (\bar{\boldsymbol{\mu}}_{j,t} - \widehat{\boldsymbol{\mu}}_{j,t})\| \tag{36}$$

$$\leqslant \|\widehat{\boldsymbol{\mu}}_{i,t} - \widehat{\boldsymbol{\mu}}_{j,t}\| + \|\bar{\boldsymbol{\mu}}_{i,t} - \widehat{\boldsymbol{\mu}}_{i,t}\| + \|\bar{\boldsymbol{\mu}}_{j,t} - \widehat{\boldsymbol{\mu}}_{j,t}\| \tag{37}$$

$$\leqslant w(\widehat{\boldsymbol{\mu}}_{i,t}, \widehat{\boldsymbol{\mu}}_{j,t}) + 2U_t^\delta, \tag{38}$$

where line (35) is simply a reordering of the terms in the previous line, lines (36) and (37) are obtained by applying the triangle inequality, and line (38) is obtained by combining the first part of the lemma with the fact that $\bar{\boldsymbol{\mu}}_{i,t}$ and $\bar{\boldsymbol{\mu}}_{j,t}$ belong to $\mathcal{C}_{i,t}^\delta$ and $\mathcal{C}_{j,t}^\delta$ respectively. □

Lemma A.2. *If the estimator is the sample mean and under Assumptions 2.1 and 3.1, with probability at least $1-\delta$, Algorithm 1 discards the sub-optimal pair $\{i,j\} \neq \{i^*, j^*\}$ after at most $t_{\{i,j\}}$ phases, where:*

$$t_{\{i,j\}} \coloneqq \frac{c^2\sigma^2}{\Delta_{\{i,j\}}^2} \max\left\{ d, 2\left(\log\left(\frac{2c^2\sigma^2}{\Delta_{\{i,j\}}^2} \sqrt{\frac{2K}{\delta}} \right) + \log\log\left(\frac{2c^2\sigma^2}{\Delta_{\{i,j\}}^2} \sqrt{\frac{2K}{\delta}} \right) + \log 2 \right) \right\}, \tag{39}$$

where $c = 32^2 \left(\max_{\boldsymbol{x} \in \partial\mathcal{B}^{d-1}} \|\boldsymbol{x}\|\right)^2$.

Proof. Consider a sub-optimal pair $\{i,j\} \neq \{i^*, j^*\}$. We have:

$$\Delta_{\{i,j\}} = w(\boldsymbol{\mu}_{i*}, \boldsymbol{\mu}_{j*}) - w(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) \leqslant \max_{\widetilde{\boldsymbol{\mu}}_i^* \in \mathcal{C}_{i*,t}^\delta, \widetilde{\boldsymbol{\mu}}_j^* \in \mathcal{C}_{j*,t}^\delta} w(\widetilde{\boldsymbol{\mu}}_i^*, \widetilde{\boldsymbol{\mu}}_j^*) - \min_{\widetilde{\boldsymbol{\mu}}_i \in \mathcal{C}_{i,t}^\delta, \widetilde{\boldsymbol{\mu}}_j \in \mathcal{C}_{j,t}^\delta} w(\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\mu}}_j) \tag{40}$$

$$\leqslant w(\widehat{\boldsymbol{\mu}}_{i*,t}, \widehat{\boldsymbol{\mu}}_{j*,t}) - w(\widehat{\boldsymbol{\mu}}_{i,t}, \widehat{\boldsymbol{\mu}}_{j,t}) + 4U_t^\delta, \tag{41}$$

where Equation (41) is obtained by applying Equations (31) and (32). From Lemma 3.2, we discard pair $\{i,j\}$ if:

$$w(\widehat{\boldsymbol{\mu}}_{i,t}, \widehat{\boldsymbol{\mu}}_{j,t}) + 2U_t^\delta \leqslant \max_{\{i',j'\}\in\mathcal{S}} w(\widehat{\boldsymbol{\mu}}_{i',t}, \widehat{\boldsymbol{\mu}}_{j',t}) - 2U_t^\delta \tag{42}$$

$$\implies \left(w(\widehat{\boldsymbol{\mu}}_{i*,t}, \widehat{\boldsymbol{\mu}}_{j*,t}) - \Delta_{\{i,j\}} + 4U_t^\delta\right) + 2U_t^\delta \leqslant w(\widehat{\boldsymbol{\mu}}_{i*,t}, \widehat{\boldsymbol{\mu}}_{j*,t}) - 2U_t^\delta \tag{43}$$

$$\implies 8U_t^\delta \leqslant \Delta_{\{i,j\}}. \tag{44}$$

By substituting U_t^δ with its value, we obtain:

$$32 \max_{\boldsymbol{x}\in\partial\mathcal{B}^{d-1}} \|\boldsymbol{x}\| \sqrt{\frac{\sigma^2 \max\left\{d, \log\frac{2Kt^2}{\delta}\right\}}{t}} \leqslant \Delta_{\{i,j\}} \tag{45}$$

which we can rewrite as

$$t \geqslant 32^2 \left(\max_{\boldsymbol{x}\in\partial\mathcal{B}^{d-1}} \|\boldsymbol{x}\|\right)^2 \frac{\sigma^2}{\Delta_{\{i,j\}}^2} \max\left\{d, \log\frac{2Kt^2}{\delta}\right\}. \tag{46}$$

If $d \geqslant \log\frac{2Kt^2}{\delta}$, we simply obtain:

$$t \geqslant 32^2 \left(\max_{\boldsymbol{x}\in\partial\mathcal{B}^{d-1}} \|\boldsymbol{x}\|\right)^2 \frac{\sigma^2 d}{\Delta_{\{i,j\}}^2}. \tag{47}$$

Otherwise, we rewrite the inequality, having denoted with $c^2 = 32^2 \left(\max_{\boldsymbol{x}\in\partial\mathcal{B}^{d-1}} \|\boldsymbol{x}\|\right)^2$ as:

$$\frac{\Delta_{\{i,j\}}^2 t}{2c^2\sigma^2} \geqslant \log\left(\frac{2c^2\sigma^2}{\Delta_{\{i,j\}}^2}\sqrt{\frac{2K}{\delta}}\frac{\Delta_{\{i,j\}}^2 t}{2c^2\sigma^2}\right). \tag{48}$$

The proof follows by applying Lemma A.3:

$$t \leqslant \frac{2c^2\sigma^2}{\Delta_{\{i,j\}}^2}\left(\log\left(\frac{2c^2\sigma^2}{\Delta_{\{i,j\}}^2}\sqrt{\frac{2K}{\delta}}\right) + \log\log\left(\frac{2c^2\sigma^2}{\Delta_{\{i,j\}}^2}\sqrt{\frac{2K}{\delta}}\right) + \log 2\right). \tag{49}$$

We conclude by defining:

$$t_{\{i,j\}} \coloneqq \frac{c^2\sigma^2}{\Delta_{\{i,j\}}^2} \max\left\{d, 2\left(\log\left(\frac{2c^2\sigma^2}{\Delta_{\{i,j\}}^2}\sqrt{\frac{2K}{\delta}}\right) + \log\log\left(\frac{2c^2\sigma^2}{\Delta_{\{i,j\}}^2}\sqrt{\frac{2K}{\delta}}\right) + \log 2\right)\right\}. \tag{50}$$

□

Theorem 3.3. *If the estimator is the sample mean and under Assumptions 2.1 and 3.1, with probability at least $1-\delta$, SED returns the optimal pair $\{i^*, j^*\}$ with a sample complexity bounded by:*

$$\tau \leqslant O\Bigg(\sum_{i\in[\![K]\!]\setminus\{i^*,j^*\}} \frac{\sigma^2}{(\Delta_i^*)^2} \max\Bigg\{d, \log\left(\frac{\sigma^2}{(\Delta_i^*)^2}\sqrt{\frac{K}{\delta}}\right) + \log\log\left(\frac{\sigma^2}{(\Delta_i^*)^2}\sqrt{\frac{K}{\delta}}\right)\Bigg\}\Bigg).$$

Proof. We consider an arm $i \in [\![K]\!]\setminus\{i^*, j^*\}$ at a time. Arm i will be pulled as long as at least one pair of the form $\{i,j\}$ belongs to the set of admissible pairs $\mathcal{S}$. This happens as long as $t \leqslant t_{\{i,j\}}$ for some $j \in [\![K]\!]\setminus\{i\}$. Thus, let us define t_i as the maximum number of times arm i is pulled:

$$t_i \coloneqq \max_{j\in[\![K]\!]\setminus\{i\}} t_{\{i,j\}} = \frac{c^2\sigma^2}{\Delta_i^{*2}} \max\left\{d, 2\left(\log\left(\frac{2c^2\sigma^2}{\Delta_i^{*2}}\sqrt{\frac{2K}{\delta}}\right) + \log\log\left(\frac{2c^2\sigma^2}{\Delta_i^{*2}}\sqrt{\frac{2K}{\delta}}\right) + \log 2\right)\right\}. \tag{51}$$

The sample complexity becomes:

$$\tau \leqslant \sum_{i\in[\![K]\!]\setminus\{i^*,j^*\}} t_i. \tag{52}$$

By applying the Big-O notation, we get the result. □

Theorem 4.1. *There exists a class of Gaussian dissimilarity bandits fulfilling Assumptions 2.1 and 3.1 such that for any fixed-confidence BAI algorithm that fulfills $\mathbb{P}(\{\widehat{I}_\tau, \widehat{J}_\tau\} = \{i^*, j^*\}) \geqslant 1-\delta$, there exists a Gaussian dissimilarity bandit $\underline{\boldsymbol{\nu}}$ such that:*

$$\mathbb{E}_{\underline{\boldsymbol{\nu}}}[\tau] \geqslant \Omega\left(\sum_{i \in [\![K]\!] \setminus \{i^*, j^*\}} \frac{\sigma^2}{(\Delta_i^*)^2} \log\left(\frac{1}{\delta}\right)\right) \tag{10}$$

$$= \Omega\left(HP \cdot \log\left(\frac{1}{\delta}\right)\right). \tag{11}$$

Proof. We consider instances of dissimilarity bandits, where the distributions of the observations $\boldsymbol{\nu}_i$ are Gaussian with diagonal covariance $\sigma^2\mathbf{I}$. Let us consider the base instance $\underline{\boldsymbol{\nu}}$ having expected observation vectors $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)^\top$. The optimal arm pair is $\{i^*, j^*\}$. Let $a \in [\![K]\!]$, we construct an alternative instance $\underline{\boldsymbol{\nu}}'$ in which only the expected observation vector of arm a changes from $\boldsymbol{\mu}_a$ to $\boldsymbol{\mu}'_a$. Let $a \notin \{i^*, j^*\}$ and $\Delta_a^* = \|\boldsymbol{\mu}_{i^*} - \boldsymbol{\mu}_{j^*}\| - \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_{b^*(a)}\|$, where $b^*(a) \in \arg\max_{b \in [\![K]\!]} \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|$ (note that we can assume $b^*(a) \neq a$). Let us define:

$$\boldsymbol{\mu}'_a = \boldsymbol{\mu}_a + 2\Delta_a^* \frac{\boldsymbol{\mu}_a - \boldsymbol{\mu}_{b^*(a)}}{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_{b^*(a)}\|}. \tag{53}$$

We show that in the alternative instance $\underline{\boldsymbol{\nu}}'$ the optimal arm pair is $\{a, b^*(a)\}$. Indeed, now $\{a, b^*(a)\}$ has larger dissimilarity than $\{i^*, j^*\}$:

$$\|\boldsymbol{\mu}'_a - \boldsymbol{\mu}_{b^*(a)}\| = \left\|\left(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{b^*(a)}\right)\left(1 + \frac{2\Delta_a^*}{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_{b^*(a)}\|}\right)\right\| \tag{54}$$

$$= \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_{b^*(a)}\| + 2\Delta_a^* \tag{55}$$

$$= \|\boldsymbol{\mu}_{i^*} - \boldsymbol{\mu}_{j^*}\| + \Delta_a^* > \|\boldsymbol{\mu}_{i^*} - \boldsymbol{\mu}_{j^*}\|, \tag{56}$$

and $b^*(a)$ remains the arm most dissimilar for a, since for every $b \in [\![K]\!]$, we have:

$$\|\boldsymbol{\mu}'_a - \boldsymbol{\mu}_b\| \leqslant \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| + 2\Delta_a^* \tag{57}$$

$$\leqslant \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_{b^*(a)}\| + 2\Delta_a^* \tag{58}$$

$$= \|\boldsymbol{\mu}'_a - \boldsymbol{\mu}_{b^*(a)}\|. \tag{59}$$

Consider the event $\{\{\widehat{I}_\tau, \widehat{J}_\tau\} = \{i^*, j^*\}\}$, any δ-PAC algorithm satisfies $\mathbb{P}_{\underline{\boldsymbol{\nu}}}(\{\widehat{I}_\tau, \widehat{J}_\tau\} = \{i^*, j^*\}) \geqslant 1-\delta$ and $\mathbb{P}_{\underline{\boldsymbol{\nu}}'}(\{\widehat{I}_\tau, \widehat{J}_\tau\} = \{i^*, j^*\}) \leqslant \delta$. Thus, we apply Lemma 1 of Kaufmann et al. (2016) to the stopping time τ, to get:

$$\mathrm{KL}(\boldsymbol{\nu}_a, \boldsymbol{\nu}'_a)\, \mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[N_a(\tau)] \geqslant \log\left(\frac{1}{2.4\delta}\right), \tag{60}$$

where $N_a(\tau)$ is the number of times arm a was pulled before stopping. Let us now compute the KL-divergence between the arm distributions:

$$\mathrm{KL}(\boldsymbol{\nu}_a, \boldsymbol{\nu}'_a) = \frac{1}{2\sigma^2}\left\|\boldsymbol{\mu}_a - \boldsymbol{\mu}'_a\right\|_2^2 = \frac{2(\Delta_a^*)^2}{\sigma^2}\left\|\frac{\boldsymbol{\mu}_a - \boldsymbol{\mu}_{b^*(a)}}{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_{b^*(a)}\|}\right\|_2^2 \leqslant \frac{2(\Delta_a^*)^2}{\sigma^2}\frac{1}{\min_{\boldsymbol{x} \in \partial\mathcal{B}^{d-1}} \|\boldsymbol{x}\|^2}. \tag{61}$$

where $\partial\mathcal{B}^{d-1}$ denotes the surface of the unit sphere. Since the derivation holds for all $a \notin \{i^*, j^*\}$, we can conclude that:

$$\mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[\tau] \geqslant \sum_{a \notin \{i^*, j^*\}} \mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[N_a] \geqslant \sum_{a \notin \{i^*, j^*\}} \frac{\sigma^2}{2(\Delta_a^*)^2} \cdot \min_{\boldsymbol{x} \in \partial\mathcal{B}^{d-1}} \|\boldsymbol{x}\|^2 \cdot \log\left(\frac{1}{2.4\delta}\right). \tag{62}$$

Passing to the Big-Ω notation leads to the result. □

Lemma A.3. *Consider the following inequality:*

$$x \geqslant \log(\alpha x). \tag{63}$$

A solution to the above inequality valid $\forall \alpha \geqslant e$ *is:*

$$x = \log \alpha + \log\log \alpha + \log 2. \tag{64}$$

Proof. By plugging our solution into (63) we obtain:

$$\log \alpha + \log\log \alpha + c \geqslant \log(\alpha(\log \alpha + \log\log \alpha + c)) \tag{65}$$
$$\alpha \log(\alpha) e^c \geqslant \alpha(\log \alpha + \log\log \alpha + c) \tag{66}$$
$$(e^c - 1)\log \alpha \geqslant \log\log \alpha + c \tag{67}$$
$$\alpha^{e^c - 1} \geqslant e^c \log \alpha. \tag{68}$$

Choosing $c = \log 2$ trivially satisfies the inequality. It can be shown that the inequality is satisfied for all $c \geqslant \log \frac{e}{e-1}$. □

B Additional Experiments

In the following, we present some information to allow the reproducibility of the presented results and some additional experiments to have a complete picture of the capabilities of the SED algorithm in different scenarios.

B.1 Additional Information for Reproducibility

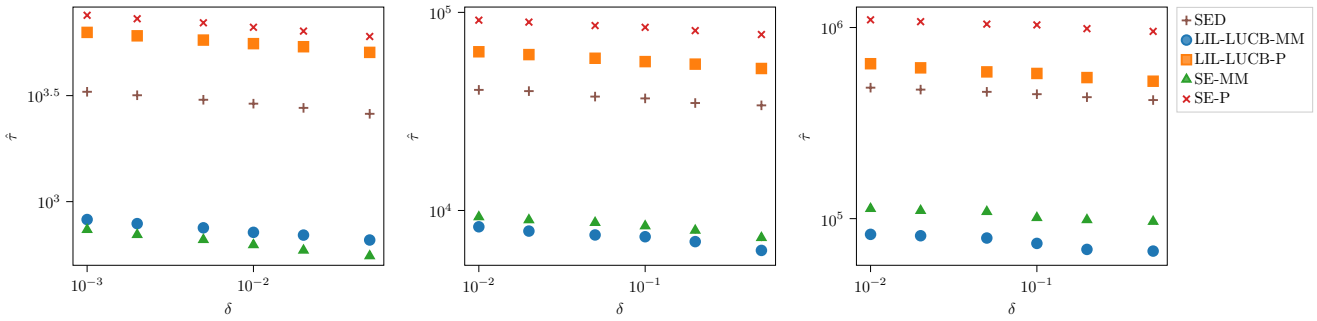
In this section, we provide additional information for the full reproducibility of the experiments provided in the main paper.

The code[5] has been run on an Intel(R) Core(TM) $i7-8750H$ CPU with 16 GiB of system memory. The operating system was Ubuntu 18.04 LTS, and the experiments were run on Python 3.9.12. The libraries used in the experiments, with the corresponding versions, were:

- `matplotlib==3.6.2`
- `tikzplotlib==0.10.1`
- `numpy==1.22.3`
- `scipy==1.8.1`

B.2 Noise Variance Experiments

The experiments provided in this section use a similar experimental setting as the one presented in Env. #1, #2 and #3 described in the main paper. The only differences are about the confidence set, which is $\delta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$, and the number of runs that has been fixed to 100. Even with fewer experiments, the confidence intervals are not visible in the pictures. In these experiments, we analyze the behavior of the SED and the baseline used in the paper as the noise variance used for generating the sample $\sigma^2 \in \{0.01, 0.1, 1\}$.



(a) $\sigma^2 = 0.01$ (b) $\sigma^2 = 0.1$ (c) $\sigma^2 = 1$

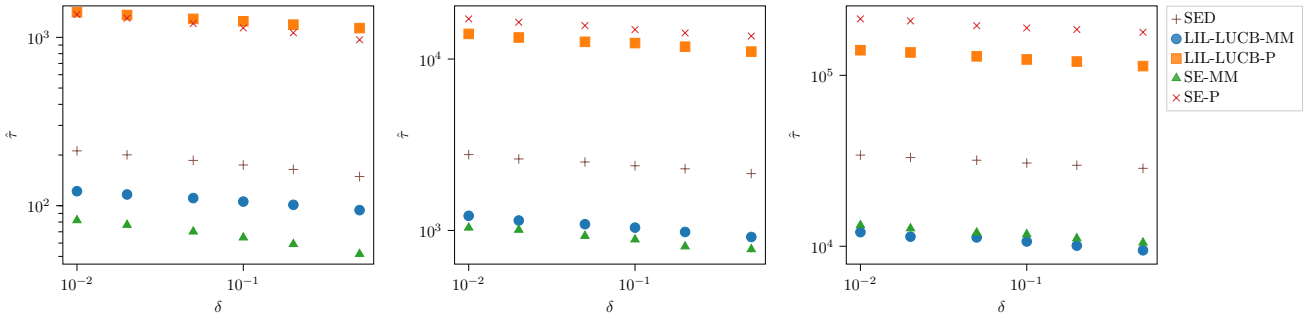
Figure 3: Results for the Env. #1.

Results The results are presented in Figures 3, 4 and 5. We used logarithmic scales for both x and y axes for ease of visualization. As expected, as the variance of the noise increases, the number of samples required for the identification of the optimal pair increases for all the methods. However, the relative ordering of the analyzed methods remains constant over the different scenarios (Except for the two LIL-LUCB variants in Env. #2) and for different values of the noise variance σ^2. The percentage of improvement over the $-P$ approach remains almost unchanged over the different scenarios.

B.3 Heterogeneous Noise Variance Experiment

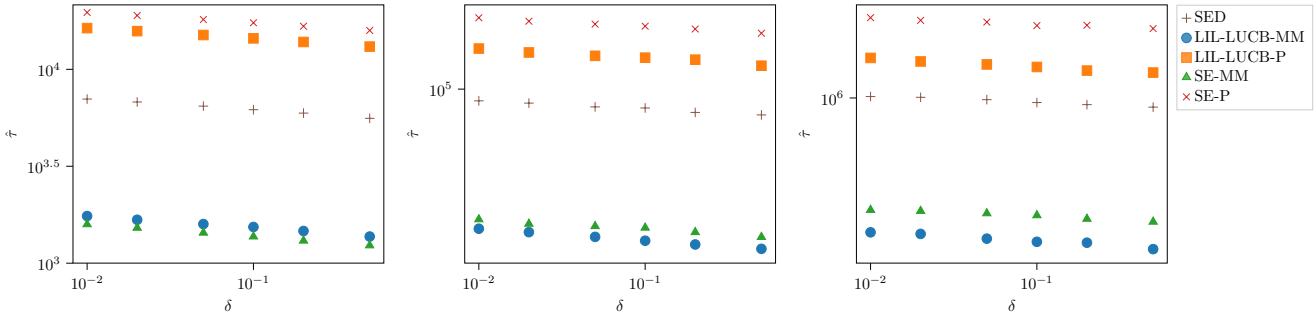
In this experiment we allow the noise variance of the observation to be heterogeneous over the different components. In particular, we draw a value σ_h uniformly over the interval $[0.01, 0.1]$ for each one of the components x_h

[5]Full code is available at `https://github.com/paolob2/sed`.



(a) $\sigma^2 = 0.01$ (b) $\sigma^2 = 0.1$ (c) $\sigma^2 = 1$

Figure 4: Results for the Env. #2.



(a) $\sigma^2 = 0.01$ (b) $\sigma^2 = 0.1$ (c) $\sigma^2 = 1$

Figure 5: Results for the Env. #3.

and we use σ_h^2 as the variance for the realization of the observation for that specific component. The remaining parameters of the experiments are the same as the ones provided in Section 7.2. We average our results over 100 independent runs.

Results The results are presented in Figure 6. They are in line with the one provided in the main paper, where the use of the SED algorithm provides an improvement of around 2 order of magnitudes in terms of stopping time $\hat{\tau}$ over the $-P$ approach. This behavior suggests that the dynamics of the phenomenon are driven by the maximum noise variance since the identification even in this case has the same empirical complexity.

B.4 Observation Vector Dimension Experiment

The following experiments are executed in the same d-dimensional environment introduced in the main paper, with the only difference being the value of d, which is chosen in $d \in \{4, 16, 64\}$ (experiments with $d = 16$ are the same as the one provided in the main paper and are reported for the sake of completeness). We provide the averaged results over 100 runs.

Results The results are shown in Figure 7. In all scenarios, the SED algorithm outperforms the *BAI on Pairs* methods by at least one order of magnitude. Even in this case, the improvement over the $-P$ approach remains almost constant (of a multiplicative factor) as d increases. This reflects the dependence on d of the regret bound in Theorem 3.3 in which the increase in terms of sample complexity has only a $\log(d)$ dependence. Overall, these results strengthen the idea that exploiting the structure, as SED does, in the dissimilarity bandit setting is providing a significant improvement over currently available options.

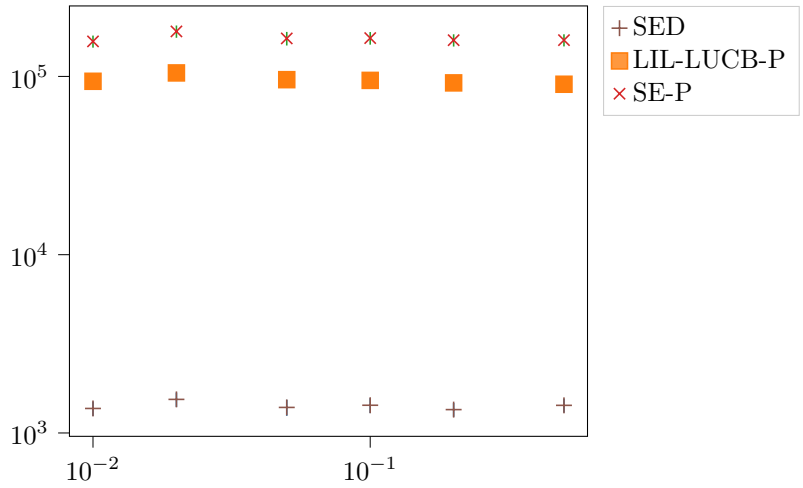
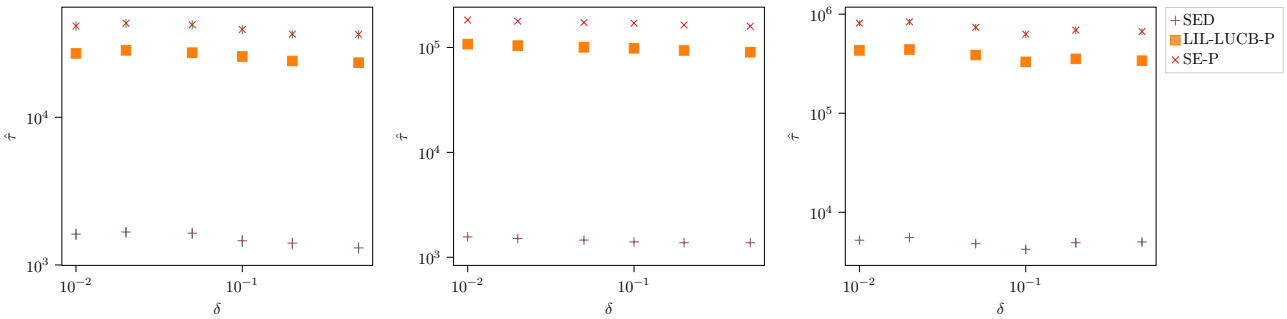


Figure 6: Results for heterogeneous variance in the 16-dimensional case.



(a) $d = 4$ (b) $d = 16$ (c) $d = 64$

Figure 7: Results for the d-dimensional scenario.