Improved Learning Complexity in Combinatorial Pure Exploration Bandits

Victor Gabillon
Queensland University of Technology (QUT)

Alessandro Lazaric
Inria Lille

Mohammad Ghavamzadeh
Adobe Research & Inria Lille

Ronald Ortner
Montanuniversität Leoben

Peter Bartlett
University of California, Berkeley & QUT

Abstract

We study the problem of combinatorial pure exploration in the stochastic multi-armed bandit problem. We first construct a new measure of complexity that provably characterizes the learning performance of the algorithms we propose for the fixed confidence and the fixed budget setting. We show that this complexity is never higher than the one in existing work and illustrate a number of configurations in which it can be significantly smaller. While in general this improvement comes at the cost of increased computational complexity, we provide a series of examples, including a planning problem, where this extra cost is not significant.

1 Introduction

In the problem of best arm identification in the stochastic multi-armed bandit (MAB) setting (e.g., Even-Dar et al. [2006], Bubeck et al. [2009], Audibert et al. [2010]), a learner has to identify the best arm/decision in a given decision space. At each step, the learner selects an action and receives a sample drawn from its corresponding reward distribution. Unlike in standard MAB, where the goal is to maximize the cumulative sum of rewards (e.g., Robbins [1952], Auer et al. [2002]), here the performance is evaluated based on the value of the arm(s) returned at the end.

In the original form of the problem, the decision set is composed of a finite number of arms/actions and the task is to identify the one with the highest expected value. This problem has been studied in two different settings. In the *fixed confidence* setting, the learner aims to minimize the number of pulls that allow her to identify the best arm with the desired confidence. In the *fixed budget* setting, the total number of pulls is fixed and the objective is to return the best arm with the highest confidence. In recent years, more complex forms of this problem have been studied. In [Kalyanakrishnan and Stone, 2010, Kalyanakrishnan et al., 2012, Gabillon et al., 2012, Kaufmann and Kalyanakrishnan, 2013], the objective is to recommend the set of m best arms. Gabillon et al. [2011], Wang et al. [2013] studied a scenario in which the best arm must be identified within each of m independent parallel bandit problems. Soare et al. [2014] considered the case in which the rewards of the arms depend linearly on an unknown parameter. Motivated by applications in project management and surveillance over a network of hospitals, Ryzhov and Powell [2011] moved to combinatorial decision sets and studied the scenario in which at each step the learner samples an edge of the graph and the goal is to find the path with the highest reward (i.e., the sum of the rewards of its edges). They assumed a Bayesian prior over the rewards of the arms and provided asymptotic results on the probability of error. Chen et al. [2014] studied the same setting and proposed two novel algorithms for the fixed confidence and the fixed budget setting, called CLUCB and CSAR. They proved an upper on their performance that was complemented by a general lower bound on the problem setting. Finally, Wu et al. [2015] studied the combinatorial case in which at each step the learner samples a path of the graph and the goal is to find the edge with the highest value. Finally, we note that the case of combinatorial actions/decisions has also been studied in the cumulative regret setting [Cesa-Bianchi and Lugosi, 2012, Chen et al., 2013, Kveton et al., 2015].

In this paper, we follow the setting of Chen et al. [2014] with the objective of designing algorithms with improved *learning complexity* relating the number of samples to the probability of error. That is, in the fixed confidence setting, the learning complexity is the required number of samples to achieve the desired confidence, while in the fixed budget setting it is the probability of error for a given budget

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

of arm pulls available. We first introduce a new measure of complexity in Sect. 3. In Sect. 4, we propose algorithms for the fixed confidence and the fixed budget setting whose learning complexity depends on this new measure. Then in Sect. 5, we show that as our complexity measure is never larger than the one of Chen et al. [2014], this leads to improved learning complexity bounds. Finally in Sect. 6, we discuss the computational complexity of our algorithms and show that although they are computationally more expensive than those of Chen et al. [2014], this extra cost is not significant in several practical scenarios.

2 Problem Formulation

We consider a set $\mathcal{K}$ of $K = |\mathcal{K}|$ arms, where each arm $i \in \mathcal{K}$ is characterized by a distribution $\nu_i \in [0,1]$ with expected value μ_i.[1] The (combinatorial) decision space $\mathcal{C} \subseteq 2^K$ contains decision sets (sets of arms) $U \subseteq \mathcal{K}$, and the *value* of a decision set $U \in \mathcal{C}$ is defined as $\mu_U = \sum_{i \in U} \mu_i$. In the following, we use upper-case letters U to Z to refer to decision sets. Without loss of generality we assume that for each arm $i \in \mathcal{K}$, there exists at least one decision set $U \in \mathcal{C}$ such that $i \in U$ and at least one decision set $V \in \mathcal{C}$ such that $i \notin V$. The gap between two decision sets is denoted by $\Delta_{U,V} = \mu_U - \mu_V$, and $U^* = \arg\max_{U \in \mathcal{C}} \mu_U$ is the best decision set with value $\mu^* = \mu_{U^*}$, which is assumed to be unique. We denote by $U \oplus V = (U \backslash V) \cup (V \backslash U)$ the exclusive disjunction between sets U and V, i.e., the set of arms either in U or in V, but not in both. Finally, the *symmetric* and *asymmetric distances* between two decision sets are defined as $\overline{d}_{U,V} = |U \oplus V|$ and $d_{U,V} = |U \backslash V|$, respectively.

Following Chen et al. [2014], we characterize the decision space $\mathcal{C}$ by a set of *patches* that can transform any decision set $U \in \mathcal{C}$ to any other decision set $V \in \mathcal{C}$, without leaving the decision space $\mathcal{C}$.

Definition 1. *An **exchange set** b is an ordered pair of disjoint sets $b = (b_+, b_-)$ such that $b_+, b_- \subseteq \mathcal{K}$ and $b_+ \cap b_- = \varnothing$. For any set U and any exchange set $b = (b_+, b_-)$, we define $U \pm b = (U \backslash b_-) \cup b_+$ and $U \mp b = (U \backslash b_+) \cup b_-$.*

Definition 2. *The set $\mathcal{B}$ is an **exchange class** for the decision space $\mathcal{C}$ if for any pair of decision sets $U \neq V \in \mathcal{C}$ and any arm $k \in U \backslash V$, there exists an exchange set $b = (b_+, b_-) \in \mathcal{B}$ that satisfies the five constraints: **(a)** $k \in b_-$, **(b)** $b_+ \subseteq V \backslash U$, **(c)** $b_- \subseteq U \backslash V$, **(d)** $(U \pm b) \in \mathcal{C}$, **(e)** $(V \mp b) \in \mathcal{C}$. The width of an exchange class is defined as width$(\mathcal{B}) = \max_{(b_+, b_-) \in \mathcal{B}} |b_+| + |b_-|$.*

Definition 3. *The decision set $U \in \mathcal{C}$ is **independent** of the decision set $V \in \mathcal{C}, U \neq V$, denoted by $U \perp V$, iff $b = (V \backslash U, U \backslash V)$ is the only non-empty exchange set that satisfies the constraints **(b)–(e)** of Definition 2 for the pair of decision sets (U, V). It is easy to see that independence is symmetric, i.e., $U \perp V$ iff $V \perp U$.*

[1] Actually, our results hold generally for bounded/sub-Gaussian distributions.

The distributions $\{\nu_i\}_{i=1}^K$ are unknown to the learner. At each round t, the learner pulls an arm $I(t)$ and observes a sample drawn from $\nu_{I(t)}$, independent from the past. The learner estimates the mean of each arm i by averaging the samples drawn from ν_i over time, i.e, $\widehat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{T_i(t)} X_i(s)$, where $T_i(t)$ is the number of times that i has been pulled by the end of round t and $X_i(s)$ is the s-th sample observed from ν_i. We denote by $\widehat{\mu}_U(t) = \sum_{i \in U} \widehat{\mu}_i(t)$ the empirical value of a decision set U, and by $\widehat{U}^*(t) = \arg\max_{U \in \mathcal{C}} \widehat{\mu}_U(t)$ the best empirical decision set at round t.

In this paper, we consider both the *fixed budget* and the *fixed confidence* setting defined as follows.

In the **fixed budget** setting, the objective is return the best decision set with the largest possible confidence using a fixed budget of n arm pulls. More formally, given a budget n, the performance of an algorithm is measured by the probability $\widetilde{\delta}$ of not identifying the best decision set, i.e., $\widetilde{\delta} = \mathbb{P}\big[\widehat{U}^*(n) \neq U^*\big]$. The smaller $\widetilde{\delta}$, the better the algorithm is.

In the **fixed confidence** setting, the goal is to return the optimal decision set with fixed confidence after the smallest possible number of arm pulls. Given a confidence level δ, if we denote by $\widetilde{n}$ the time when the algorithm stops, we want to have $\mathbb{P}\big[\widehat{U}^*(\widetilde{n}) \neq U^*\big] \leq \delta$. The performance of the algorithm is thus measured by the number of rounds $\widetilde{n}$, either in expectation or in high probability.

3 Definition of Learning Complexity

In this section, we introduce our novel complexity measure for combinatorial pure exploration problems. While in Section 4 we derive algorithms whose performance is actually characterized by this new measure of complexity, in the following we introduce it in a constructive way to provide a more solid intuition about its properties. In Section 7, we discuss its relationship to existing lower bounds and possible improvements.

Since the objective in combinatorial pure exploration is to identify the optimal set U^* in $\mathcal{C}$, we first focus on characterizing the complexity of discriminating between any two decision sets $U, V \in \mathcal{C}$, i.e., determining whether $\mu_V > \mu_U$ or $\mu_V \leq \mu_U$. As usual, we expect that the smaller the gap $\Delta_{V,U}$, the harder it is to identify the better set. However in our setting, resources cannot be directly allocated to the sets, but rather need to be allocated to the arms in $U \cup V$ until the estimates of μ_U and μ_V are accurate enough to discriminate U from V. In order to simplify the discussion, we focus on how often we have to pull arms $i \in U \cup V$ to identify the better set *with confidence* $1 - \delta$.[2] We consider an algorithm that sequentially selects arms in $U \cup V$ and

[2] As shown in Section 4.1, the arguments used in constructing the arm complexity H_i are still valid in the fixed budget setting.

at the end of each step t constructs the empirical estimate $\widehat{\mu}_i(t)$ for each arm i using the $T_i(t)$ samples of arm i that have been observed so far. By a direct application of Hoeffding's inequality, we may construct confidence intervals

$$|\widehat{\mu}_i(t) - \mu_i| \leq \beta_i(t) = \sqrt{\frac{\log \frac{4K't^2}{\delta}}{2T_i(t)}}, \qquad (1)$$

which hold with probability at least $1-\delta$ for all $K' = \overline{d}_{U,V}$ arms at any time step $t > 0$. At the end of step t, we construct the empirical estimate of the gap between U and V as

$$\widehat{\Delta}_{V,U}(t) = \sum_{i\in V}\widehat{\mu}_i(t) - \sum_{i\in U}\widehat{\mu}_i(t) = \sum_{i\in V\setminus U}\widehat{\mu}_i(t) - \sum_{j\in U\setminus V}\widehat{\mu}_j(t),$$

which shows that only arms in $U \oplus V$ actually play a role in discriminating between U and V. As a result, we consider a simple extension of the Hoeffding Races algorithm [Maron and Moore, 1993], which selects arms in $U \oplus V$ using a round-robin strategy (in an arbitrary order) and stops at the first step t when the lower-bound on the gap is positive, i.e.,

$$\widehat{\Delta}_{V,U}(t) - \sum_{i\in U\oplus V}\beta_i(t) > 0. \qquad (2)$$

The sample complexity of such an algorithm is bounded in the following lemma.

Lemma 1. *Let $U, V \in \mathcal{C}$ such that $\mu_V > \mu_U$ and let*

$$H_{U,V} = \overline{d}^2_{U,V}/\Delta^2_{U,V}.$$

When the round-robin algorithm with the termination condition (2) stops after t steps, then for any arm $i \in U \oplus V$, we have $T_i(t) \leq 2H_{U,V}\log\left(\frac{4K't^2}{\delta}\right) + 1$ and V is returned as the better set with probability at least $1-\delta$.

Lemma 1 is obtained using classical techniques. The full proof is reported in Appendix A. Lemma 1 provides an upper bound on the number of times each arm in the disjunction $U \oplus V$ should be pulled before learning that V is better than U with sufficiently high confidence. In particular, Lemma 1 shows that beyond the inverse dependency on the gap $\Delta_{U,V}$, the upper bound also depends on the number of arms in the disjunction $U \oplus V$. The number of arms $\overline{d}_{U,V}$ can be interpreted as a variance term, as the confidence interval associated to a set is proportional to $\overline{d}_{U,V}$. As a result, given a fixed gap Δ, it is easier to discriminate sets that differ by only few arms. This property implies that when trying to discard a suboptimal set U from $\mathcal{C}$ (i.e., find that $U \neq U^*$ with high confidence), it may be easier to compare U to a set $V \neq U^*$ with $\mu_V > \mu_U$ and smaller complexity $H_{U,V}$. Thus, we introduce the following definition.

Definition 4. *The complement of any decision set $U \neq U^*$ is*

$$\mathsf{C}_U = \operatorname*{arg\,min}_{V\in\mathcal{C}:\mu_V>\mu_U} H_{U,V}, \qquad (3)$$

where ties are broken in favor of V with smaller $\overline{d}_{U,V}$.

If $H_{U,V}$ characterizes the complexity of discriminating between U and V, C_U is the set that is the most effective in *revealing* that U is actually suboptimal. The complement C_U has also an additional important property.

Lemma 2. *$U \perp \mathsf{C}_U$ holds for all $U \in \mathcal{C}$ with $U \neq U^*$.*

This lemma (proof in App. E), shows that for any suboptimal set U is independent from its complement C_U.

Lemma 1 suggests that in order to discard a suboptimal set U, the most effective strategy is to pull all the arms in $U \oplus \mathsf{C}_U$ a number of times proportional to H_{U,C_U}. Thus, we define the complexity of an arm as the largest complexity for discarding a set U with i in $U \oplus \mathsf{C}_U$.

Definition 5. *The complexity of an arm $i \in \mathcal{K}$ is*[3]

$$H_i = \max_{U\in\mathcal{C}:i\in U\oplus\mathsf{C}_U} H_{U,\mathsf{C}_U}. \qquad (4)$$

As a direct consequence of Lemma 1 and Definition 5, we note that an algorithm pulling each arm proportionally to H_i and stopping when all sets but one are discarded returns an empirical best set $\widehat{U}^* = \arg\max_{U\in\mathcal{C}}\widehat{\mu}_U$ that is optimal with probability at least $1-\delta$. Consequently, we define the global complexity H as the sum of the complexities of the individual arms in $\mathcal{K}$.

Definition 6. *The global complexity H is defined as*

$$H = \sum_{i\in\mathcal{K}} H_i. \qquad (5)$$

For notational convenience, we also introduce the notion of *simplicity* of a pair of decision sets (U, V) as

$$G_{U,V} = \Delta_{U,V}/\overline{d}_{U,V}.$$

Unlike $H_{U,V}$, $G_{U,V}$ is an asymmetric quantity, i.e., $G_{U,V} = -G_{V,U}$. We also define the simplicity of an arm $i \in \mathcal{K}$ as $G_i = \min_{U\in\mathcal{C}:i\in U\oplus\mathsf{C}_U} G_{\mathsf{C}_U,U}$. Note that simplicity is a positive quantity, i.e., $G_{\mathsf{C}_U,U} > 0$, since $\mu_{\mathsf{C}_U} > \mu_U$. Also note that $H_{U,V} = G^{-2}_{U,V}$ and $H_i = G_i^{-2}$.

4 Learning Algorithms

In this section, we introduce novel learning algorithms for the fixed budget and the fixed confidence setting. Both algorithms are designed to discard an arm i whenever sufficient information is gathered to decide whether or not it belongs to U^*. While for existing algorithms, this requires that the arms in U^* are pulled sufficiently often, our algorithms compare a decision set U not always to U^*, but rather to C_U. Thus, they focus on pulling arms in both U and C_U sufficiently often. This is achieved by computing

[3] See Appendix B for a proof that H_i is well defined.

Parameters: number of rounds n, set of arms $\mathcal{K}$, decision set $\mathcal{C}$, and cumulative pulls scheme $n_0, n_1, \ldots, n_K$.
Let $\mathcal{K}_1 = \mathcal{K}$, $k = 1$.
while $|\mathcal{K}_k| \geq 1$ **do**
  Pull each arm $i \in \mathcal{K}_k$ for $n_k - n_{k-1}$ rounds.
  Compute $\widehat{U}^*(k) = \arg\max_{U \in \mathcal{C}} \widehat{\mu}_U(k)$.
  Find $j_k = \max_{i \in \mathcal{K}_k} \widehat{G}_i(k)$.
  Deactivate arm j_k, i.e., set $\mathcal{K}_{k+1} = \mathcal{K}_k \backslash j_k$.
  $k \longleftarrow k + 1$
end while
Return $J_n = \arg\max_{U \in \mathcal{C}} \widehat{\mu}_U(n)$

Figure 1: The fixed budget algorithm.

empirical estimates of the complexity measure H_i and progressively discarding arms with low complexity. The resulting algorithms enjoy performance guarantees on probability of error and on sample complexity, where the bounds exhibit an explicit dependency on H. In Section 5, we show that this leads to a potential significant gain w.r.t. the algorithms of Chen et al. [2014]. The computational complexity of our algorithms is discussed in Section 6.

4.1 Fixed Budget

Figure 1 shows our fixed budget algorithm. Apart from the introduction of the new notion of complexity, H_i, the algorithm builds upon a rather standard rejection strategy shared by many existing algorithms such as Successive Rejects (SR) [Audibert et al., 2010], SAR [Wang et al., 2013], and CSAR [Chen et al., 2014], which is specifically designed for combinatorial problems. The algorithm runs over K phases. At each phase k, it maintains a set of active arms $\mathcal{K}_k$ that are all pulled uniformly until they reach n_k samples. At the end of a phase, we compute the empirical means $\widehat{\mu}_i(k)$, the empirical gap for any pair of sets U and V, as $\widehat{\Delta}_{V,U}(k) = \widehat{\mu}_V(k) - \widehat{\mu}_U(k)$, and the estimated optimal decision $\widehat{U}^*(k) = \arg\max_{U \in \mathcal{C}} \widehat{\mu}_U(k)$. Using these estimates, we also build empirical versions of the terms introduced in Section 3, such as the estimated simplicity between two sets U and V as $\widehat{G}_{V,U}(k) = \widehat{\Delta}_{V,U}(k)/\overline{d}_{V,U}$, which in turn implies the following definitions for the empirical complement of a decision set $U \neq \widehat{U}^*$,

$$\widehat{\mathsf{C}}_U(k) = \underset{V \in \mathcal{C}: \widehat{\mu}_V(k) > \widehat{\mu}_U(k)}{\arg\max} \widehat{G}_{V,U}(k), \tag{6}$$

and the estimated simplicity of an arm $i \in \mathcal{K}$,

$$\widehat{G}_i(k) = \min_{U \in \mathcal{C}: i \in U \oplus \widehat{\mathsf{C}}_U(k)} \widehat{G}_{\widehat{\mathsf{C}}_U(k),U}(k). \tag{7}$$

In (6), ties are broken in favor of V with the smaller distance $\overline{d}_{V,U}$. At the end of each phase k, the *easiest* arm $j_k = \arg\max_{i \in \mathcal{K}_k} \widehat{G}_i(k)$, i.e., the arm with largest estimated simplicity in $\mathcal{K}_k$, is removed from the active set. Note that j_k is the arm for which it is easiest to determine whether it belongs to U^* or not, and thus, if $j_k \in \widehat{U}^*(k)$, then j_k is *accepted* and will be a part of the final recommended solution J_n, otherwise it is *rejected*. In either case, it is not included in $\mathcal{K}_{k+1}$ and is not pulled anymore. [4] We use $n_k = \left\lceil \frac{n-K}{\overline{\log}(K)(K+1-k)} \right\rceil$, $k \in \mathcal{K}$, with $n_0 = 0$ and $\overline{\log}(K) = \sum_{i=1}^{K} 1/i$. It is easy to verify that with this scheme the algorithm never exceeds the budget n. In fact, since at each of the K phases one arm is deactivated, the total budget used is $n_{FB} = \sum_{k=1}^{K} n_k \leq K + \frac{n-K}{\overline{\log}(K)} \left(\sum_{k=1}^{K} \frac{1}{K+1-k} \right) = n$. We prove the following performance guarantee for the algorithm.

Theorem 1. *The probability of error of the fixed budget algorithm in Figure 1 is*

$$\mathbb{P}\left[\widehat{U}^*(n) \neq U^*\right] \leq 2K^2 \exp\left(\frac{n-K}{32\overline{\log}(K)\overline{H}}\right),$$

where $\overline{H} = \max_{i \in \mathcal{K}} i H_{\pi(i)}$ *and* π *is a permutation of* $\mathcal{K}$ *such that* $H_{\pi(1)} \geq H_{\pi(2)} \geq \ldots \geq H_{\pi(K)}$. *As noted in Audibert et al. [2010], it holds that* $\overline{H} \leq H \leq \overline{H}\overline{\log}(K)$.

The full proof that –like the algorithm– borrows ideas from Audibert et al. [2010], Wang et al. [2013], and Chen et al. [2014] can be found in Appendix G. Here we only provide a proof sketch. The proof proceeds by induction on the phases of the algorithm. The two induction hypotheses essentially claim that if an arm $i \notin \mathcal{K}_k$ has been deactivated during phase $l \in \{1, \ldots, k-1\}$, the number n_l of samples obtained for arm i is proportional to its complexity $H_i = 1/G_i^2$, which is crucial for the correct functioning of the method. It first means that the deactivated arms have been pulled sufficiently often in order to determine whether they belong to the optimal set U^*. Moreover, although j_k is selected among the active arms in $\mathcal{K}_k$ on the basis of the estimated simplicity, the computation of $\widehat{G}_i$ requires comparing each set U containing i to its (estimated) complement $\widehat{\mathsf{C}}_U$ (see Definition 5). Since the arms in C_U may no longer be active, we need to guarantee that when they are deactivated, their values are estimated sufficiently precise, so that $\widehat{\mathsf{C}}_U$, and as a result $\widehat{G}_i$, are accurate.

4.2 Fixed Confidence

Figure 2 shows our fixed confidence algorithm. At each step t, the algorithm first uses the samples up to step $t-1$ to compute an upper bound on the simplicity $G_{U,V}(t)$ of any pair of decision sets (U, V). To do so, we first define upper and lower bounds for the mean of an arm i as $\widehat{\mu}_i^+(t) = \widehat{\mu}_i(t-1) + \beta_i(t-1)$ and $\widehat{\mu}_i^-(t) = \widehat{\mu}_i(t-1) - \beta_i(t-1)$, where $\beta_i(t-1)$ is the confidence interval for arm i at time t defined in (1). For any pair of decision sets (U, V), we then compute an upper bound on their gap as $\widehat{\Delta}_{U,V}^+(t) = \sum_{i \in U \backslash V} \widehat{\mu}_i^+(t) - \sum_{j \in V \backslash U} \widehat{\mu}_j^-(t)$ and on their simplicity as $\widehat{G}_{U,V}^+(t) = \widehat{\Delta}_{U,V}^+(t)/\overline{d}_{U,V}$.

[4] Note that the empty set can be considered a decision in $\mathcal{C}$, which explains why K and not $K-1$ phases are necessary. A default value then need to be associated with the empty set decision.

Parameters: confidence δ, set of arms $\mathcal{K}$, and decision set $\mathcal{C}$.
Initialize: Pull each arm i once
Set $\mathcal{U}_{K+1} = \{U : \forall V \in \mathcal{C}, \widehat{\Delta}^+_{U,V}(K+1) > 0\}$.

while $|\mathcal{U}_t| > 1$ **do**
Set threshold $\mathcal{T}_{U,V}(t) = \overline{d}_{U,V} \max_{W \in \mathcal{C}} \widehat{G}^+_{W,U}(t)/2$
Set $\mathcal{U}'_t = \{U : \forall V \in \mathcal{C}, \widehat{\Delta}^+_{U,V}(t) > -\mathcal{T}_{U,V}(t)\}$
Let $(U_t, V_t) = \operatorname{arg\,max}_{U \in \mathcal{U}'_t, V \in \mathcal{C}, U \neq V} \widehat{G}^+_{V,U}(t)$
Let $(W_t, Z_t) = \operatorname{arg\,max}_{(W,Z) \in \{(U_t,V_t),(V_t,U_t)\}} \sum_{i \in W \setminus Z} \beta_i(t-1)$
Sample arm $I(t) = \arg\max_{i \in W_t \setminus Z_t} \beta_i(t-1)$.
Update $\mathcal{U}_{t+1} = \{U : \forall V \in \mathcal{C}, \widehat{\Delta}^+_{U,V}(t+1) > 0\}$
$t \longleftarrow t+1$
end while
Return the unique decision set in $\mathcal{U}_t$, $\widehat{U}^*(t)$.

Figure 2: The fixed confidence algorithm.

At each step t, the set $\mathcal{U}_t$ is constructed as the set of decision sets U, whose upper bound on the gap is positive w.r.t. any other set $V \in \mathcal{C}$. This corresponds to all sets that are still potential candidates for the best set U^* (i.e., there is not enough confidence to discard them). Then, the most uncertain arm belonging to the simplest pair (U_t, V_t) is selected, and this is repeated until only one set in $\mathcal{U}_t$ is left, which is then returned as $\widehat{U}^*(t)$. While it would be natural to select the sets (U_t, V_t) among those still "active" in $\mathcal{U}_t$, this would not guarantee a proper behavior for the algorithm. Similarly to the fixed budget case, the largest simplicity for a set U is associated to its complement $V = \mathsf{C}_U$, and thus, in order to guarantee that the upper bound on the estimated simplicity, $\widehat{G}^+_{U,V}$, is accurate, we need to guarantee that all the arms in V have been pulled at least a number of times proportional to their complexity. This is achieved by introducing an additional set $\mathcal{U}'_t$. While in constructing $\mathcal{U}_t$, a set U is dropped when it is dominated with high confidence, i.e., the upper bound on its gap $\widehat{\Delta}^+_{U,V}(t)$ is negative for at least one set V, $\mathcal{U}'_t$ is more conservative and requires the gap to be negative by "enough" margin before actually discarding a set. That is, we introduce the threshold $\mathcal{T}_{U,V}(t) > 0$, and let a set U be discarded from $\mathcal{U}'_t$ only if there is a set V such that $\widehat{\Delta}^+_{U,V}(t) < -\mathcal{T}_{U,V}(t)$. This allows us to guarantee that all the arms that could be involved in identifying a suboptimal set are pulled often enough. In fact, after computing $\mathcal{U}'_t$, the algorithm identifies the pair of decision sets (U_t, V_t) with the highest upper bound on simplicity in $\mathcal{U}'_t$ and selects among (U_t, V_t) the decision W with the largest sum of uncertainty terms $\beta_i(t-1)$ for $i \in W$. Then the algorithm pulls the arm with the largest uncertainty in $W \cap (U_t \oplus V_t)$.

We now state the sample complexity of the algorithm.

Theorem 2. *The algorithm in Figure 2 stops after* $\tilde{n} \leq O\big(H \log(HK/\delta)\big)$ *steps and returns the optimal decision set* U^* *with probability at least* $1-\delta$.

We report the proof in Appendix H. For all t, we define $\widehat{G}^+(t) = \max_{U \in \mathcal{U}'_t, V \in \mathcal{C}} \widehat{G}^+_{V,U}(t)$. The main idea is to show that $\widehat{G}^+(t)$ is upper bounded by $\beta_{I(t)}$ (Lemma 10) and lower bounded by $G_{I(t)}$ (Lemma 8), thus obtaining that $G_{I(t)} \leq \widehat{G}^+(t) \leq \beta_{I(t)}$. Given the definition of β_i in (1), we recover an upper bound on the number of pulls $T_i(t)$ for each arm, and thus, bound the overall sample complexity.

5 Comparison of Learning Complexities

In this section we show that the interest in designing algorithms whose performance is characterized by the complexity measure of Definition 5 resides on the fact that this represents a significant improvement w.r.t. previous pure exploration combinatorial algorithms. We first show that the complexity H_i is never higher than the complexity measure $H_i^{\odot}$ introduced by [Chen et al., 2014] for the performance analysis of the algorithms CLUCB and CSAR. Then we provide illustrative examples showing that our new complexity measure can be significantly smaller.

We recall the definition of the complexity measure of Chen et al. [2014]. For any arm $i \in \mathcal{K}$, the gap is defined as

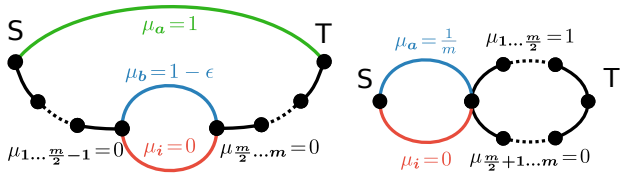
$$\Delta_i^{\odot} = \begin{cases} \mu^* - \max_{U \in \mathcal{C}: i \in U} \mu_U & \text{if } i \notin U^*, \\ \mu^* - \max_{U \in \mathcal{C}: i \notin U} \mu_U & \text{if } i \in U^*. \end{cases}$$

The width of $\mathcal{C}$ is the width of the smallest exchange class for $\mathcal{C}$, that is $\text{width}(\mathcal{C}) = \min_{\mathcal{B} \in \text{exchange}(\mathcal{C})} \text{width}(\mathcal{B})$ and the resulting complexity of arm $i \in \mathcal{K}$ is $H_i^{\odot} = \text{width}(\mathcal{C})^2 / \left(\Delta_i^{\odot}\right)^2$, leading to the global complexity $H^{\odot} = \sum_{i \in \mathcal{K}} H_i^{\odot}$.[5] The following theorem shows that for any arm $i \in \mathcal{K}$ our complexity measure H_i is never higher than the measure $H_i^{\odot}$ of Chen et al. [2014].

Theorem 3. *For all* $i \in \mathcal{K}$, $H_i^{\odot} \geq H_i$.

We first provide some intuition about the statement of the theorem. The complexity H_i of an arm i is defined as the maximum over the complexities of the decision sets U for which $i \in U \oplus \mathsf{C}_U$. On the other hand, $H_i^{\odot}$ only considers the maximum over the sets U with $i \in U$. We therefore need to guarantee that the extra terms in the maximum of our definition do not lead to a larger value compared to $H_i^{\odot}$. Consider for instance the specific case $\mathcal{C} = \{U, V, U^*\}$ with $V = \mathsf{C}_U$ and V being the only decision set containing i. Then $U^* = \mathsf{C}_V$, $H_i^{\odot} = H_{V,\mathsf{C}_V}$, and $H_i = \max\{H_{V,\mathsf{C}_V}, H_{U,V}\}$, so that if $H_{U,V} > H_{V,\mathsf{C}_V}$ then $H_i > H_i^{\odot}$. Fortunately, we can show in Appendix C and D that generally $H_{V,\mathsf{C}_V} \geq H_{U,V}$, so that the additional terms in the maximum do not increase the value of H_i. More

[5] Notice that in the original paper, the complexity of an arm is defined as $\left(1/\Delta_i^{\odot}\right)^2$, but looking at the statements of the theorems, the complexity of i is always multiplied by the square of the width of the decision space.



(a) Graph Example 1 (b) Graph Example 2

Figure 3: Examples of decision spaces where H is significantly smaller than $H^{\odot}$. Each arm is identified with an edge of a graph, and $\mathcal{C}$ corresponds to the possible paths (without loops) from the source S to the target T.

generally, we prove that for all $i \notin U^*$, the maximum in the complexity H_i is attained by a set U such that $i \in U \backslash \mathsf{C}_U$.

Now we proceed with a proof sketch. Thus, consider an arm $i \notin U^*$ and let

$$U_i^{\odot} = \operatorname*{arg\,max}_{U \in \mathcal{C}: i \in U} \mu_U, \quad U_i = \operatorname*{arg\,max}_{U \in \mathcal{C}: i \in U \oplus \mathsf{C}_U} H_{U,\mathsf{C}_U}.$$

Notice that $U_i^{\odot}$ is the decision set that implicitly defines the complexity of arm i according to the definition of $\Delta_i^{\odot}$. As mentioned before, $i \in U_i$. Let $V_i = \mathsf{C}^*_{U_i}$, where C^* is a variation of C such that if we define an exchange set b with $b_+ = V_i \backslash U_i$ and $b_- = U_i \backslash V_i$, we have that V_i is indeed "between" U and U^*, i.e. requiring $U^* \mp b \in \mathcal{C}$ and $i \in b_-$. As a result, we have that $\Delta_{U^*,U_i^{\odot}} = \min_{U \in \mathcal{C}: i \in U} \Delta_{U^*,U} \leq \Delta_{U^*,U^* \mp b} = \Delta_{\mathsf{C}^*_{U_i},U_i}$, since i belongs to b_-. Furthermore, recalling Definition 3 of independent sets, we notice that an equivalent interpretation of the width of $\mathcal{C}$ is to consider it as the maximal distance $\overline{d}_{U,V}$ between any two independent sets U, V (i.e., $U \perp V$). In fact, $\overline{d}_{U,V}$ counts the number of arms in the disjunction $U \oplus V$. However, in the case of independent sets, this coincides with an exchange set b with $b_+ = V \backslash U$ and $b_- = U \backslash V$ such that $\overline{d}_{U,V} = |b_+| + |b_-|$. Since by Lemma 2, U_i is independent from its complement V_i, we have $\text{width}(\mathcal{C}) \geq \overline{d}_{V_i,U_i}$. Summarizing, we obtain the claimed

$$H_i^{\odot} = \text{width}(\mathcal{C})^2 / \Delta^2_{U^*,U_i^{\odot}} \geq \overline{d}^2_{V_i,U_i} / \Delta^2_{V_i,U_i} = H_i.$$

The most interesting aspect of this result is that the potential improvement of H_i over $H_i^{\odot}$ may be achieved on both terms characterizing the complexity, that is, the distance $\overline{d}_{\mathsf{C}_U,U}$ (which can be smaller than the width of $\mathcal{C}$) and the gap $\Delta_{\mathsf{C}_{U_i},U_i}$ (which can be larger than the gap between U_i and U^*). This is demonstrated in the two following illustrative examples, in which H_i is indeed much smaller than $H_i^{\odot}$.

Example 1: Comparing U to C_U instead of U^*. The definition of $\Delta_i^{\odot}$ always depends on the comparison of sets U containing i to the optimal decision set U^*. The following example demonstrates that comparing decision sets to their complement can considerably reduce the overall complexity of an arm i. Consider the shortest path problem[6] with $\mathcal{K} = \{1, \dots, m, a, b, i\}$ illustrated in Figure 3(a). The optimal path between source node S and exit node T is the green path $U^* = \{a\}$ with $\mu^* = 1$. We first focus on the complexity of the red arm i. This arm only belongs to the decision set $U = \{i, 1, \dots, m\}$ (i.e., the black and red path). The complexity of discriminating U from U^* is $H_{U^*,U} = \frac{\overline{d}^2_{U^*,U}}{\Delta^2_{U^*,U}} = \frac{(m+2)^2}{1^2} = (m+2)^2$. Notice that $H_{U^*,U}$ coincides with $H_i^{\odot}$ since the largest exchange set in this problem is indeed the one transforming U into U^*, and thus $H_i^{\odot} = \text{width}(\mathcal{C})^2 / \left(\Delta_i^{\odot}\right)^2 = (m+2)^2$. On the other hand, the complexity of discriminating U from the set $V = \{b, 1, \dots, m\}$, which differs from U only by the exchange set $(\{i\}, \{b\})$, corresponds to $H_{V,U} = \frac{\overline{d}^2_{V,U}}{\Delta^2_{V,U}} = \frac{2^2}{(1-\epsilon)^2}$. As a result, as soon as $m > 2\epsilon/(1-\epsilon)$, we have that $H_i = H_{V,U} < H_{U^*,U} = H_i^{\odot}$. In particular, $H_i^{\odot} = \frac{(1-\epsilon)^2(m+2)^2}{4} H_i$. Since we can take ϵ arbitrarily small while m is of order of K, we have $H_i^{\odot} = O(K^2) H_i$, implying that complexity H_i can be K^2 times smaller than the complexity proposed by Chen et al. [2014]. While this shows already the potential of the complexity measure H_i, it is limited to one single arm, and it does not immediately show that the overall complexity of finding the optimal set is significantly reduced. However, it is enough to slightly modify the previous example by adding p copies of arm i, thus leading to a total number of $K = m + 2 + p$ arms. Choosing $\epsilon = 1/2$ we have $H = \sum_{j \in \mathcal{K}} H_j = \sum_{j=1}^{m+2} H_j + pH_i = 4(m+2)^3 + 16p$ and $H^{\odot} = \sum_{j \in \mathcal{K}} H_j^{\odot} = \sum_{j=1}^{m+2} H_j^{\odot} + pH_i^{\odot} = 4(m+2)^3 + p(m+2)^2$. Then choosing $p = O\left(m^3\right)$, we have $K = O\left(m^3\right)$, $H = O\left(m^3\right)$, and $H^{\odot} = O\left(m^5\right) = HO\left(K^{2/3}\right)$. This shows that not only the per-arm complexity H_i can be significantly smaller but that this may have a major impact in the overall complexity of the combinatorial pure exploration problem.

Example 2: Width of the graph vs individual pair distance. Unlike in the definition of $H_i^{\odot}$, where the distance between sets only appears in form of the width of the global decision set $\mathcal{C}$, H_i takes into consideration the specific distances $\overline{d}_{U,\mathsf{C}_U}$ for each U with $i \in U$. Since $\text{width}(\mathcal{C})$ may be larger than $\overline{d}_{U,\mathsf{C}_U}$, we expect H_i to better adapt to the "local" geometry of $\mathcal{C}$. We illustrate this intuition in the example shown in Figure 3(b). This is a shortest path problem[7] in a graph between a source node S and an exit node T, where $\mathcal{K} = \{1, \dots, m, a, i\}$ and the optimal path is $U^* = \{a, 1, \dots, m/2\}$ (i.e., the blue edge followed by the top path). We focus on the complexity of the red arm i. Let $V = \{i, 1, \dots, m/2\}$ (i.e., the red edge followed by the top path) be the best path containing i. Then $\Delta_i^{\odot} = \Delta_{U^*,V} = \frac{1}{m}$. The width of the decision space is

[6] Actually, the goal is to maximize the rewards over the edges in a path.

[7] Again, we actually want to maximize rewards over the path.

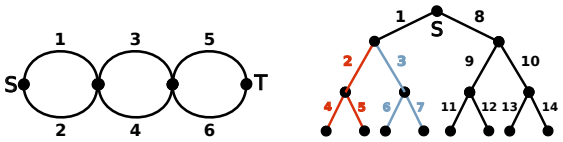
m, since the largest exchange set needs to remove all the (black) arms in the top (bottom) path and add all the (black) arms in the bottom (top) path (e.g., consider the exchange sets needed to move from U^* to $\{a, m/2+1, \dots, m\}$). Hence, $H_i^{\odot} = \frac{\text{width}(\mathcal{C})^2}{(\Delta^{\odot}_{U^*,V})^2} = \frac{m^2}{(1/m)^2} = m^4$, while $H_i = \frac{\overline{d}^2_{U^*,V}}{\Delta^2 U^*,V} = \frac{2^2}{(1/m)^2} = 4m^2 \leq H_i^{\odot}$, showing that the complexity of arm i can be $\frac{1}{4}m^2$ times smaller than $H_i^{\odot}$. As in the previous example, we are also interested in comparing the global complexities H and $H^{\odot}$. In this case, we can show that $H = O\left(m^2\right)$ since the complexity of all black arms is 1. On the other hand, $H^{\odot} = O\left(m^4\right)$. As m increases, $m \approx K$, we have that $H^{\odot} = O\left(K^2\right) H$, which suggests that overall complexity H can be K^2 times smaller than $H^{\odot}$.

6 Computational Complexity

In this section, we discuss the computational complexity of the algorithms presented in Section 4 and compare it to the complexity of the algorithms of Chen et al. [2014], also taking into consideration the respective learning complexity discussed in Section 5.

In [Chen et al., 2014] the complexity is dominated by an oracle solving the combinatorial optimization problem $U^* = \arg\max_{U \in \mathcal{C}} \mu_U$. While this task is NP-hard in general, in some particular instances such as maximum matching or maximum weight spanning tree, the computational complexity of finding the best estimated decision in $\mathcal{C}$ is polynomial in K. On the other hand, in both our algorithms, the computational complexity is dominated by the computation of the learning complexity of the arms. In fact, as shown in equation (5), computing H_i for all arms $i \in \mathcal{K}$ (the same for G_i and their empirical counterparts) requires evaluating the complexity of any pair of sets U, V in $\mathcal{C}$. As a result, in the worst case the computational complexity for both algorithms is $O\left(|\mathcal{C}|^2\right)$, which in some cases can be exponential in the number of arms (e.g., maximum matching or maximum weight spanning tree). While in general this may be the unavoidable price to pay for improving the learning complexity, in the following we show that, **1)** in some problems where we do not improve the learning complexity, the computational complexity is indeed not worse than for Chen et al. [2014], **2)** there exists a class of planning problems where we obtain a better learning complexity with only limited extra computational cost.

Taking advantage of independence, the multi-bandit example. Lemma 2 allows to move from looping over $\mathcal{C}$ to looping over the exchange sets in the exchange class $\mathcal{B}$. In particular, we can easily construct the exchange class obtained by considering all the exchange sets defined by the disjunction of all pairs of independent decisions. In some problems, this observation may lead to a much more efficient implementation of our algorithms. For instance, let us consider a multi-bandit problem [Gabillon et al., 2011] with M bandits, each composed of K/M arms (see illustration in the "sausage graph" in Fig. 4(a) for the three-bandits, two-arm case). In this problem, the learning complexity cannot be improved w.r.t. Chen et al. [2014]. Nonetheless, we can exploit the structure of the problem to match their computational complexity. We first notice that any two independent sets U and V always differ by only one arm. In fact, if they differ by more than one arm, then there always exists more than one way to bring U closer to V by a one-arm transformation and at the same time stay in $\mathcal{C}$. As a result, when computing H_i we can directly compare i with all the arms in the same bandit. For instance, in Fig. 4(a), computing H_1 would normally require looping over each decision set U including arm 1 and considering all the other decisions to identify its complement. Let us consider $U = \{1, 3, 5\}$, then its complement is $\mathsf{C}_U = \{2, 3, 5\}$, which only differs by arm 2. This is the same for all U containing 1 and thus, when computing H_1 we can simply compare it to arm 2 without actually looping over all the decision sets. As a result, computing the learning complexity reduces to comparing arms within the same bandit, thus leading to M independent problems and a complexity of $O(K)$.



(a) 2 arm, 3 bandit problem (b) Planning problem

Figure 4: Examples of learning problems in which our algorithms perform well.

***K* $\approx$ $|\mathcal{C}|$, the tree-planning case:** Whenever $|\mathcal{C}|$ is of the same order as K, then the computational complexity is tractable. This means that, for instance, in the illustrative examples discussed in Sect. 5, we not only enjoy a significantly smaller learning complexity (e.g., with a reduction of order $O(K^2)$ but we also match the computational complexity of the methods proposed by Chen et al. [2014]. An even more interesting case is the problem of planning. In this case, $\mathcal{C}$ describes a tree structure of depth m and each node has the same branching factor a. An arm i is an edge of the tree with associated weight μ_i, and a decision set U is a path from the root to a leaf. The objective is to find a decision set (i.e., a path) that maximizes the sum of its weights. This setting corresponds to the open-loop planning problem of maximizing the expected sum of rewards over m consecutive actions (chosen in a set of a actions) from a starting state when the state dynamics is deterministic and the reward distributions are unknown. This type of problem has been previously studied with discounted rewards [Bubeck and Munos, 2010, Munos, 2014].

In this problem, $2|\mathcal{C}| \approx K$ as the number of decisions (paths) is equal to the number of leaves in the tree. While this already shows that our computational complexity is

comparable to previous methods, in the following we push the comparison even further. In the fixed budget setting, the CSAR algorithm [Chen et al., 2014] needs to query a shortest path oracle for each edge i to determine the paths with largest value including and not including arm i. This procedure, if implemented naively, leads to an overall complexity of order $\widetilde{O}\left(K^2\right)$ (for each arm the computation requires $O(K\log(K))$ operations using a simplified version of Dijkstra), which matches our $O\left(|\mathcal{C}|^2\right)$ complexity. We conjecture that this computational complexity can be actually reduced to Km for both our and their algorithm. We focus on our algorithms and highlight a technique that can be used to reduce the computational complexity in general. First, using the Dijkstra algorithm gives the best path from any node to the corresponding leaves. Our algorithms will then compute H_{U,C_U} for all U. However, the set of decisions (paths) $V \neq U$ can be clustered into m groups of sets depending on the first node where they differ from U among the m possible ones. Since the distance $d_{U,V}$ is constant within these clusters, identifying the complement $\arg\min H_{U,V}$ within each cluster corresponds to finding the set V with the largest value, which has already been computed by the Dijkstra algorithm. So for each U we need to consider m clusters, which would permit to reduce the overall complexity to Km. This complexity is of the same order as for the CLUCB fixed confidence algorithm [Chen et al., 2014] where just one call to the Dijkstra algorithm is needed.

Not only in this setting our algorithms can be implemented efficiently, but we can also show an example where the learning complexity is significantly improved over [Chen et al., 2014]. Let us consider the tree in Figure 4(b). If $\mu_i = .9$ for all edges except the leaf edges with odd numbers $5, 7, 11, 13$ for which $\mu_i = 0$ and $\mu_4 = 1$. In this case, arms $5, 7, 11, 13$ belong to only one decision set U each and thus computing their complexity coincides with finding the complement C_U and computing H_{U,C_U}. Since almost all paths have the same value, the C_U is chosen as the set V minimizing $\overline{d}_{U,V}$, which is simply the path differing from U for only the last edge, i.e., $\overline{d}_{U,V} = 1$. Comparing to $H_i^{\odot}$, which has width$(\mathcal{C}) = m$, for all such arms we have $mH_i = O\left(H_i^{\odot}\right)$. Since the proportion of this type of arms grows with the branching factor a, this improvement can reduce the global complexity H by a factor m.

7 Discussion

We have seen in Section 5 that using the complexity measure H one can obtain improved results on the learning complexity. This naturally raises the question whether the obtained upper bounds for our algorithms are optimal. The core of the definition of H_i is indeed the complexity $H_{U,V}$ of discriminating between any two decision sets U and V. While in Section 3 we gave a constructive definition, Chen et al. [2014] provide a lower bound on the "cumulative" number of samples for each exchange set. In particular, they show that for any arm $j \in \mathcal{K}$, there exists an exchange set $b = (b_+, b_-)$ such that $j \in b_+ \cup b_-$ and

$$\mathbb{E}\left[\sum_{i \in b_+ \cup b_-} T_i(t)\right] \geq \frac{(|b_+| + |b_-|)^2}{(\Delta_j^{\odot})^2} \log(1/4\delta), \quad (8)$$

where t is the stopping time at which the optimal set is returned w.p. $1-\delta$. As a result, a proper lower bound in the fixed confidence setting is derived from the optimization problem $\min_{\mathbb{E}[T_i(t)],\mathcal{B}} \sum_{i\in\mathcal{K}} \mathbb{E}\big[T_i(t)\big] \log(1/4\delta)$, subject to the constraints of the form in (8), for all exchange sets $b \in \mathcal{B}$ and arms $j \in b_+ \cup b_-$. While it is difficult to have a clear understanding of the resulting overall sample complexity for the lower bound, we can greatly simplify it by considering the simple case of $\mathcal{C} = \{U, U^*\}$, for which $b = (U^* \backslash U, U \backslash U^*)$, and an algorithm that pulls all the arms in $b_+ \cup b_-$ uniformly. Then, for each arm $i \in U \oplus U^*$, the lower bound of equation (8) becomes

$$\mathbb{E}\big[T_i(t)\big] \geq \frac{|b_+| + |b_-|}{\Delta^2_{U^*,U}} \log(1/4\delta), \quad (9)$$

which strongly resembles our definition of H_{U,U^*}. We first notice that the major gap is related to the fact that in H_{U,U^*} the numerator has the distance squared. We conjecture that this gap could be filled by using more accurate deviation inequalities in bounding $|\mu_U - \widehat{\mu}_U|$ so that the confidence bound has the sum inside the square root. On the other hand, the major gap resides on the fact that equation (9) considers a simple uniform allocation over arms in the disjunction, while the full lower bound in equation (8) allows for more sophisticated allocation strategies and considers the interplay between the constraints imposed by the exchange sets in $\mathcal{B}$. Quantifying the resulting gap and determining whether it can be actually filled remains an open question. A first step may be to ask whether the complexity could be defined using the asymmetric distance $d_{U,V}$ between two sets U, V, instead of $\overline{d}_{U,V}$. Notice that, as this distance plays a similar role here as the variance σ_i^2 of an arm i in the standard single-bandit best arm identification problem, it is already an open question to see whether in the fixed budget setting the complexity of an arm i is σ_i^2/Δ_i^2 instead of the standard $\left(\sigma_i^2 + \sigma_{i^*}^2\right)/\Delta_i^2$, where i^* is the best arm.

Acknowledgements

We gratefully acknowledge the support of the Australian Research Council through an Australian Laureate Fellowship (FL110100281) and through the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). A. Lazaric was supported by CRIStAL (Centre de Recherche en Informatique et Automatique de Lille) and the French National Research Agency (ANR) under project ExTra-Learn n.ANR-14-CE24-0010-01. This research was funded by the Austrian Science Fund (FWF): P 26219-N15. We want to thank Sébastien Bubeck for his help.

References

A. Antos, V. Grover, and Cs. Szepesvári. Active Learning in Heteroscedastic Noise. *Theoretical Computer Science*, 411(29-30):2712–2728, 2010.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best Arm Identification in Multi-Armed Bandits. In *Proceedings of the Twenty-Third Conference on Learning Theory*, pages 41–53, 2010.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47:235–256, 2002.

S. Bubeck and R. Munos. Open Loop Optimistic Planning. In *Proceedings of the Twenty-Third Conference on Learning Theory*, 2010.

S. Bubeck, R. Munos, and G. Stoltz. Pure Exploration in Multi-Armed Bandit Problems. In *Proceedings of the Twentieth International Conference on Algorithmic Learning Theory*, pages 23–37, 2009.

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

S. Chen, T. Lin, I. King, M. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems 27*, pages 379–387, 2014.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

V. Gabillon, M. Ghavamzadeh, A. Lazaric, and S. Bubeck. Multi-Bandit Best Arm Identification. In *Proceedings of the Advances in Neural Information Processing Systems 25*, pages 2222–2230, 2011.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Proceedings of the Advances in Neural Information Processing Systems 26*, pages 3221–3229, 2012.

S. Kalyanakrishnan and P. Stone. Efficient Selection of Multiple Bandit Arms: Theory and Practice. In *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, pages 511–518, 2010.

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

É. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Proceedings of the Twenty-Sixth Conference on Learning Theory*, pages 228–251, 2013.

B. Kveton, Z. Wen, A. Ashkan, and Cs. Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.

O. Maron and A. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In *Proceedings of the Advances in Neural Information Processing Systems 7*, 1993.

R. Munos. From bandits to Monte-Carlo tree search: The optimistic principle applied to optimization and planning. *Foundation and Trends in Machine Learning*, 2014.

H. Robbins. Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

I. Ryzhov and W. Powell. Information collection on a graph. *Operations Research*, 59(1):188–201, 2011.

M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pages 828–836, 2014.

T. Wang, N. Viswanathan, and S. Bubeck. Multiple Identifications in Multi-Armed Bandits. In *Proceedings of the Thirtiethth International Conference on Machine Learning*, volume 28, pages 258–265, 2013.

Y. Wu, A. Gyorgy, and Cs. Szepesvari. On identifying good options under combinatorially structured feedback in finite noisy environments. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1283–1291, 2015.

A Proof of Lemma 1

Proof. Let $\bar{t}$ be the first time when the condition in equation (2) is met, then at time $\bar{t}-1$ we have that

$$\widehat{\Delta}_{V,U}(\bar{t}-1) - \sum_{i \in U \oplus V} \beta_i(\bar{t}-1) \leq 0,$$

which implies that

$$\Delta_{V,U} - 2 \sum_{i \in U \oplus V} \beta_i(\bar{t}-1) \leq 0,$$

and thus

$$\sum_{i \in U \oplus V} \sqrt{\frac{\log \frac{4K'(\bar{t}-1)^2}{\delta}}{2T_i(\bar{t}-1)}} \geq \frac{\Delta_{V,U}}{2}.$$

Since the algorithm is selecting arms in $U \oplus V$ using a round-robin strategy, at any time $T_i(t) = T_j(t) \pm 1$ for any pair of arms i, j. Thus, let j be the least pulled arm at round $\bar{t}-1$ (i.e., $j \in \arg\min T_i(\bar{t}-1)$), then the previous inequality can be written as

$$\overline{d}_{U,V} \sqrt{\frac{\log \frac{4K'(\bar{t}-1)^2}{\delta}}{2T_j(\bar{t}-1)}} \geq \frac{\Delta_{V,U}}{2},$$

which leads to the statement. □

B The Complexity in Equation (4) is Well-defined

In order to prove that our complexity measure in equation (4) is well-defined, we have to show that the $\max$ operator actually returns a value, that is, that there exists at least one element in the argument of the $\max$ operator. This is done in the following proposition.

Proposition 1. *Let Q_i be the set of the decision sets in the argument of the* $\max$ *operator in equation* (4)*, i.e.,*

$$Q_i = \{U \in \mathcal{C} : i \in U \oplus \mathsf{C}_U\}. \tag{10}$$

Then $Q_i \neq \varnothing$ for all i in $\mathcal{K}$.

Proof. We distinguish the following two cases:

Case 1) $i \notin U^*$

According to the assumption we made in Section 2, there exists a decision set $V \in \mathcal{C}$ such that $i \in V$. Note that i does not necessarily belong to $V \oplus \mathsf{C}_V$. We construct a sequence of decision sets $\mathcal{X} = \{X_1, \ldots, X_p\}$ such that $X_1 = V$, for all $j \in \{1, \ldots, p-1\}$, $i \in X_j$ and $X_{j+1} = \mathsf{C}_{X_j}$, and $i \notin X_p$.[8] As a result, setting $U = X_{p-1}$ and $\mathsf{C}_U = X_p$, we have that $i \in U \oplus \mathsf{C}_U$, and thus, $Q_i \neq \varnothing$.

Case 2) $i \in U^*$

Let $V = \underset{U \in \mathcal{C} \,:\, i \notin U}{\arg\max}\, \mu_U$. Then C_V exists, because $i \in U^*$. Moreover $i \in \mathsf{C}_V$, as otherwise we obtain the contradiction $\mu_{\mathsf{C}_V} > \mu_V = \underset{U \in \mathcal{C} \,:\, i \notin U}{\max} \mu_U \geq \mu_{\mathsf{C}_V}$. Therefore, $i \in \mathsf{C}_V \backslash V$, so that $V \in Q_i$ and $Q_i \neq \varnothing$. □

[8] Note that $\mathcal{X}$ is not only finite but also contains a decision set X_p, such that $i \notin X_p$. The first claim comes from the definition of complement that gives us $\mu_{X_{j+1}} > \mu_{X_j},\ \forall j \in \{1, \ldots, p-1\}$, and the fact that $\mathcal{C}$ is finite. For the second claim note that as $i \notin U^*$, $\mathcal{X}$ has at least two elements.

C Useful Properties of the Complexity $H_{U,V}$

In this appendix, we prove several useful properties of the complexity $H_{U,V}$ later, particularly in the proof that our complexity measure is not higher than the measure used by Chen et al. [2014].

The first proposition shows that the distance $\overline{d}_{U,V}$ between two decision sets U and V follows a triangle inequality.

Proposition 2. *For any three distinct decision sets $U, V, W \in \mathcal{C}$, we have, $U \oplus V \subseteq (U \oplus W) \cup (W \oplus V)$ and $\overline{d}_{U,V} \leq \overline{d}_{U,W} + \overline{d}_{W,V}$. Moreover, if $(U \oplus W) \cap (W \oplus V) = \varnothing$ then $U \oplus V = (U \oplus W) \cup (W \oplus V)$ and $\overline{d}_{U,V} = \overline{d}_{U,W} + \overline{d}_{W,V}$.*

Proof. To prove the statements for $U \oplus V$, we first prove that

$$\begin{aligned} U\backslash V &= \big((U\backslash V)\backslash W\big) \cup \big((U\backslash V) \cap W\big) \\ &= \Big((U\backslash W)\backslash\big((U\backslash W) \cap (V\backslash W)\big)\Big) \cup \Big((W\backslash V)\backslash\big((W\backslash V) \cap (W\backslash U)\big)\Big) \qquad (11) \\ &\subseteq (U\backslash W) \cup (W\backslash V). \qquad (12) \end{aligned}$$

Similar to (12), we may prove

$$V\backslash U \subseteq (V\backslash W) \cup (W\backslash U). \qquad (13)$$

Taking the union of (12) and (13) gives

$$(U\backslash V) \cup (V\backslash U) \subseteq (U\backslash W) \cup (W\backslash V) \cup (V\backslash W) \cup (W\backslash U),$$

and the first claim of the proposition

$$U \oplus V \subseteq (U \oplus W) \cup (W \oplus V). \qquad (14)$$

follows by definition of $\oplus$. The second claim is straightforward from (14) and the definition of symmetric distance $\overline{d}$, i.e.,

$$\overline{d}_{U,V} \leq \overline{d}_{U,W} + \overline{d}_{W,V}.$$

To prove the second part of the proposition, we start from (11), that is,

$$U\backslash V = \Big((U\backslash W)\backslash \underbrace{\big(\overbrace{(U\backslash W)}^{X} \cap \overbrace{(V\backslash W)}^{Y}\big)}_{A}\Big) \cup \Big((W\backslash V)\backslash \underbrace{\big(\overbrace{(W\backslash V)}^{S} \cap \overbrace{(W\backslash U)}^{Z}\big)}_{B}\Big),$$

from which by our assumption $(U \oplus W) \cap (W \oplus V) = \varnothing$ we obtain

$$A \cup B = (X \cup S) \cap \underbrace{\overbrace{(X \cup Z)}^{U \oplus W} \cap \overbrace{(Y \cup S)}^{W \oplus V}}_{\varnothing} \cap (Y \cup Z) = \varnothing \implies A = \varnothing \text{ and } B = \varnothing. \qquad (15)$$

From (15), (C) may be written as

$$U\backslash V = (U\backslash W) \cup (W\backslash V). \qquad (16)$$

Similarly, one can show that

$$V\backslash U = (V\backslash W) \cup (W\backslash U). \qquad (17)$$

Taking union from both sides of (16) and (17), we obtain

$$U \oplus V = (U \oplus W) \cup (W \oplus V),$$

and as a result $\overline{d}_{U,V} = \overline{d}_{U,W} + \overline{d}_{W,V}$, which completes the proof of the second part of the proposition. □

The next proposition proves useful properties for the complexity of two decision sets.

Proposition 3. *For any three decision sets $U, V, W \in \mathcal{C}$ with $\mu_U < \mu_V < \mu_W$, we have*

$$H_{U,W} \leq \max\left(H_{U,V}, H_{V,W}\right). \qquad (18)$$

Furthermore, if $(U \oplus V) \cap (V \oplus W) = \varnothing$, then

$$H_{U,W} \geq \min\left(H_{U,V}, H_{V,W}\right), \qquad (19)$$

and finally, the above two inequalities are strict if $H_{U,V} \neq H_{V,W}$.

Proof. We write

$$\Delta_{W,U} = \mu_W - \mu_U = \mu_W - \mu_V + \mu_V - \mu_U = \Delta_{W,V} + \Delta_{V,U}. \tag{20}$$

By assumption, we have $\mu_W > \mu_V$ and $\mu_V > \mu_U$, and thus, $\Delta_{W,V} > 0$ and $\Delta_{V,U} > 0$. As a result, we may write

$$\frac{\overline{d}_{U,W}}{\Delta_{W,U}} \overset{(a)}{\leq} \frac{\overline{d}_{U,V} + \overline{d}_{V,W}}{\Delta_{W,V} + \Delta_{V,U}} \overset{(b)}{\leq} \max\left(\frac{\overline{d}_{U,V}}{\Delta_{V,U}}, \frac{\overline{d}_{V,W}}{\Delta_{W,V}}\right), \tag{21}$$

where **(a)** follows from Proposition 2 and (20), and **(b)** follows from the fact that for any positive $a, b, c, d \geq 0$, it holds that $\frac{a+b}{c+d} \leq \max\left(\frac{a}{c}, \frac{b}{d}\right)$.[9] From (21), we get

$$\frac{\overline{d}^2_{U,W}}{\Delta^2_{W,U}} \leq \max\left(\frac{\overline{d}^2_{U,V}}{\Delta^2_{V,U}}, \frac{\overline{d}^2_{V,W}}{\Delta^2_{W,V}}\right),$$

which gives us (18).

The second statement is similarly proved as

$$\frac{\overline{d}_{U,W}}{\Delta_{W,U}} \overset{(a)}{=} \frac{\overline{d}_{U,V} + \overline{d}_{V,W}}{\Delta_{W,V} + \Delta_{V,U}} \overset{(b)}{\geq} \min\left(\frac{\overline{d}_{U,V}}{\Delta_{V,U}}, \frac{\overline{d}_{V,W}}{\Delta_{W,V}}\right), \tag{22}$$

where **(a)** is true under the assumption that $(V \oplus U) \cap (V \oplus W) = \varnothing$ from Proposition 2, and **(b)** follows from the fact that for any positive $a, b, c, d \geq 0$, it holds that $\min\left(\frac{a}{c}, \frac{b}{d}\right) \leq \frac{a+b}{c+d}$.[10] From (22) it follows that

$$\frac{\overline{d}^2_{U,W}}{\Delta^2_{W,U}} \geq \min\left(\frac{\overline{d}^2_{U,V}}{\Delta^2_{V,U}}, \frac{\overline{d}^2_{V,W}}{\Delta^2_{W,V}}\right),$$

which gives us (19).

The proof of the very last statement, the strict inequalities, comes directly from the fact that when $\frac{a}{c} \neq \frac{b}{d}$, the two inequalities at step **(b)** of (21) and (22) become strict. □

In the last proposition of this section, we show that the complexity of discriminating between a decision set $U \neq U^*$ and its complement $V = C_U$ is less than the complexity of discriminating between $V = C_U$ and its complement $W = C_V = C_{C_U}$, provided that the complement of U is not the best decision set U^*, i.e., $V = C_U \neq U^*$.

Proposition 4. *For any decision set $U \neq U^*$ with $V = \mathsf{C}_U \neq U^*$ and $W = \mathsf{C}_V = C_{C_U}$, it holds that $H_{U,V} \leq H_{V,W}$.*

Proof. We prove the statement by contradiction. Let us assume that $H_{U,V} > H_{V,W}$. Since $V \neq U^*$ and by definition of complement, we have $\mu_U < \mu_V < \mu_W$. As a result, $H_{U,W} \leq \max(H_{U,V}, H_{V,W})$ from Proposition 3. Note that this inequality is strict, whenever $H_{U,V} \neq H_{V,W}$, again according to Proposition 3, and since we assumed that $H_{U,V} > H_{V,W}$, we have $H_{U,W} < H_{U,V}$. This gives us

$$H_{U,V} = H_{U,\mathsf{C}_U} = \min_{Z \in \mathcal{C}: \mu_Z > \mu_U} H_{U,Z} \leq H_{U,W} < H_{U,V},$$

which leads to the contradiction that $H_{U,V} < H_{U,V}$. □

D Equivalence of the Different Notions of Arm Complexity

In this section, we give two alternative notions of complexity of an arm that are equivalent to the original definition H_i of equation 4. In the analysis of the algorithms (see Appendices G and H) we will use the definition of the complexity that is the most handy. The equivalence proof requires the results of Appendix C, especially Proposition 4. We start with the definition of the alternative complexity notions and two intermediate results that will be needed for the equivalence proof given at the end of this section.

[9] Here is the proof: Assume without loss of generality that $\frac{a}{c} \leq \frac{b}{d}$. Then $\frac{a+b}{c+d} \leq \frac{bc/d+b}{c+d} = \frac{b(c/d+1)}{d(c/d+1)} = \frac{b}{d} = \max\left(\frac{a}{c}, \frac{b}{d}\right)$.

[10] The proof is analogous to the previous footnote: Assume without loss of generality that $\frac{a}{c} \leq \frac{b}{d}$. Then $\frac{a+b}{c+d} \geq \frac{a+da/c}{c+d} = \frac{a(d/c+1)}{c(d/c+1)} = \frac{a}{c} = \min\left(\frac{a}{c}, \frac{b}{d}\right)$.

Definition 7. *Our two notions of complexity for an arm $i \in \mathcal{K}$, $\mathcal{H}^1$ and $\mathcal{H}^2$, are defined as*

$$\mathcal{H}_i^1 = \begin{cases} \max_{U \in \mathcal{C}: i \in U \setminus \mathsf{C}_U} H_{U,\mathsf{C}_U} & \text{if } i \notin U^*, \\ \max_{U \in \mathcal{C}: i \in \mathsf{C}_U \setminus U} H_{U,\mathsf{C}_U} & \text{if } i \in U^*. \end{cases} \qquad \mathcal{H}_i^2 = \begin{cases} \max_{U \in \mathcal{C}: i \in U} H_{U,\mathsf{C}_U} & \text{if } i \notin U^*, \\ \max_{U \in \mathcal{C}: i \notin U} H_{U,\mathsf{C}_U} & \text{if } i \in U^*. \end{cases}$$

The following proposition plays an important role in proving the equivalence $\mathcal{H}_i^1 = H_i, \forall i \in \mathcal{K}$. It shows that if $i \in (U \oplus \mathsf{C}_U)$, then $\mathcal{H}_i^1 \geq H_{U,\mathsf{C}_U}$.

Proposition 5. *For any decision set $U \in \mathcal{C}$ such that $U \neq U^*$, and any arm $i \in (U \oplus \mathsf{C}_U)$, we have $\mathcal{H}_i^1 \geq H_{U,\mathsf{C}_U}$.*

Proof. We consider the following two cases for a fixed arm $i \in (U \oplus \mathsf{C}_U)$:

Case 1) $i \in U^*$

If $i \in \mathsf{C}_U \setminus U$, the result follows directly from the definition of $\mathcal{H}_i^1$. If $i \in U \setminus \mathsf{C}_U$, similar to the proof of Proposition 1 in Appendix B, we construct a sequence of decision sets $\{X_1, \ldots, X_p\}$ such that $X_1 = U$, for all $j \in \{2, \ldots, p-1\}$, $i \notin X_j$ and $X_{j+1} = \mathsf{C}_{X_j}$, and $i \in X_p$. Note that $\{X_2, \ldots, X_p\}$ is a sequence of decision sets and their complements that do not contain arm i until a set X_p is generated that contains i.[11] From Proposition 4, we have that $H_{X_j,X_{j+1}} \geq H_{X_{j-1},X_j}, \forall j \in \{2, \ldots, p-1\}$. Now starting from the definition of $\mathcal{H}_i^1$, we may write

$$\mathcal{H}_i^1 = \max_{Z: i \in \mathsf{C}_Z \setminus Z} H_{Z,\mathsf{C}_Z} \overset{\text{(a)}}{\geq} H_{X_{p-1},X_p} \geq H_{X_1,X_2} = H_{U,\mathsf{C}_U},$$

which proves the claim of the proposition. Note that **(a)** comes from the fact that $i \in \mathsf{C}_{X_{p-1}} \setminus X_{p-1}$ by definition of the sequence.

Case 2) $i \notin U^*$

If $i \in U \setminus \mathsf{C}_U$, the result follows directly from the definition of $\mathcal{H}_i^1$. When $i \in \mathsf{C}_U \setminus U$, we construct a sequence of decision sets $\{X_1, \ldots, X_p\}$ such that $X_1 = U$, for all $j \in \{2, \ldots, p-1\}$, $i \in X_j$ and $X_{j+1} = \mathsf{C}_{X_j}$, and $i \notin X_p$. This is a sequence of decision sets and their complements that contain arm i until a set X_p is generated that does not contain i. As a result $i \in X_{p-1} \setminus \mathsf{C}_{X_{p-1}}$. From Proposition 4, we have that $H_{X_j,X_{j+1}} \geq H_{X_{j-1},X_j}, \forall j \in \{2, \ldots, p-1\}$. Now starting from the definition of $\mathcal{H}_i^1$, we may write

$$\mathcal{H}_i^1 = \max_{Z: i \in Z \setminus \mathsf{C}_Z} H_{Z,\mathsf{C}_Z} \geq H_{X_{p-1},X_p} \geq H_{X_1,X_2} = H_{U,\mathsf{C}_U}, \tag{23}$$

which proves the claim of the proposition. □

Proposition 6. *For any decision set $U \in \mathcal{C}$ such that $U \neq U^*$, and any arm $i \in (U \oplus U^*)$, we have $\mathcal{H}_i^1 \geq H_{U,\mathsf{C}_U}$.*

Proof. Let us construct a sequence of decision sets $\{X_1, \ldots, X_p\}$ such that $X_1 = U$, for all $j \in \{2, \ldots, p-1\}$, $X_{j+1} = \mathsf{C}_{X_j}$, and $X_p = U^*$. This sequence is well-defined and has at least two elements, since $U \neq U^*$ and U^* is unique. If we prove that for any $j \in \{2, \ldots, p\}$, we have that for all $i \in (X_1 \oplus X_j)$, $\mathcal{H}_i^1 \geq H_{X_1,X_2}$, then $j = p$ will gives us the proof of the proposition. Now let us prove this statement. The proof is by induction on j.

Base Step: $j = 2$. In this case, the claim follows directly from Proposition 5.

Inductive Step: Here we assume that for $j = j'$, we have that $\mathcal{H}_i^1 \geq H_{X_1,X_2}, \forall i \in (X_1 \oplus X_{j'})$, and we want to show that $\mathcal{H}_i^1 \geq H_{X_1,X_2}, \forall i \in (X_1 \oplus X_{j'+1})$. From Proposition 5 and the construction of the sequence, we have $\mathcal{H}_i^1 \geq H_{X_{j'},X_{j'+1}}, \forall i \in (X_{j'}, X_{j'+1})$. By repeated application of Proposition 4, we can show that $\forall i \in (X_1 \oplus X_{j'}) \cup (X_{j'} \oplus X_{j'+1})$, we have $\mathcal{H}_i^1 \geq \min(H_{X_1,X_{j'}}, H_{X_{j'},X_{j'+1}}) \geq H_{X_1,X_2}$. Moreover, from Proposition 2, we know that $X_1 \oplus X_{j'+1} \subseteq (X_1 \oplus X_{j'}) \cup (X_{j'} \oplus X_{j'+1})$, and thus, we obtain $\forall i \in (X_1 \oplus X_{j'+1})$ that $\mathcal{H}_i^1 \geq H_{X_1,X_2}$, which proves the inductive step. □

We are now ready to prove the main result of this section, the equivalence of the different notions of arm complexity.

Lemma 3. *For any arm $i \in \mathcal{K}$, we have $H_i = \mathcal{H}_i^1 = \mathcal{H}_i^2$.*

[11] Note that such a sequence is finite, because by the definition of complement of a decision set, we have $\mu_{X_{j+1}} > \mu_{X_j}$, the number of decision sets is finite, and $i \in U^*$.

Proof. **Step 1:** We first prove that $H_i = \mathcal{H}_i^1, \forall i \in \mathcal{K}$.

From the definition of $\mathcal{H}_i^1$, it is immediate to see that $\mathcal{H}_i^1 \leq \max_{U \in \mathcal{C}: i \in U \oplus \mathsf{C}_U} H_{U,\mathsf{C}_U} = H_i$, and from Proposition 5, we may write $H_i = \max_{U \in \mathcal{C}: i \in U \oplus \mathsf{C}_U} H_{U,\mathsf{C}_U} \leq \mathcal{H}_i^1$. These together prove Step 1.

Step 2: We now want to prove $\mathcal{H}_i^1 = \mathcal{H}_i^2, \forall i \in \mathcal{K}$.

From the definitions of $\mathcal{H}_i^1$ and $\mathcal{H}_i^2$, it is immediate to write $\mathcal{H}_i^1 \leq \mathcal{H}_i^2$. To prove the reverse, we consider the following two cases:

Case 1) $i \notin U^*$

In this case, we may write

$$\mathcal{H}_i^2 = \max_{U \in \mathcal{C}: i \in U} H_{U,\mathsf{C}_U} = \max_{U \in \mathcal{C}: i \in (U \oplus U^*)} H_{U,\mathsf{C}_U} \overset{\text{(a)}}{\leq} \mathcal{H}_i^1,$$

where **(a)** is from Proposition 6.

Case 2) $i \in U^*$

In this case, we may write

$$\mathcal{H}_i^2 = \max_{U \in \mathcal{C}: i \notin U} H_{U,\mathsf{C}_U} = \max_{U \in \mathcal{C}: i \in (U \oplus U^*)} H_{U,\mathsf{C}_U} \overset{\text{(a)}}{\leq} \mathcal{H}_i^1,$$

where **(a)** is from Proposition 6.

The two cases together prove Step 2. □

E Proof of Lemma 2

Let $U \in \mathcal{C}$ be a decision set with complement C_U. Let b be an exchange set that satisfies constraints **(b)–(e)** of Definition 2 for the decision set pair (U, C_U). Let $V = U \pm b$ be the decision set resulted from applying the transformation b to U. We now define the exchange set $d = (C_U \backslash V, V \backslash C_U)$ as the exchange set that completes the transformation of U to C_U after applying b to U. It is easy to show that $d = \big((C_U \backslash U) \backslash b_+, (U \backslash C_U) \backslash b_-\big)$. We now prove the following two propositions that are used in the proof of Lemma 2.

Proposition 7. *For any decision set $U \in \mathcal{C}$, any exchange set b that satisfies constraints **(b)–(e)** of Definition 2 for the decision set pair (U, C_U), and any exchange set d that completes the transformation of U to C_U after applying b to U, i.e., $d = \big((C_U \backslash U) \backslash b_+, (U \backslash C_U) \backslash b_-\big)$, we have*

$$\Delta_{C_U,U} = \Delta_{b_+,b_-} + \Delta_{d_+,d_-} > 0, \tag{24}$$

$$\overline{d}_{U,C_U} = \overline{d}_{b_+,b_-} + \overline{d}_{d_+,d_-}, \tag{25}$$

so that $H_{U,C_U} = \frac{(\overline{d}_{b_+,b_-} + \overline{d}_{d_+,d_-})^2}{(\Delta_{b_+,b_-} + \Delta_{d_+,d_-})^2}$.

Proof. We begin with the proof of (24). By definition of C_U, $\mu_{C_U} > \mu_U$, so that Δ_{b_+,b_-} and Δ_{d_+,d_-} cannot be both negative. Now to prove the equality, first note that $\mu_{d_+} = \mu_{C_U \backslash U} - \mu_{b_+}$ and $\mu_{d_-} = \mu_{U \backslash C_U} - \mu_{b_-}$ from the definition of d and the fact that $b_+ \subseteq C_U \backslash U$. Further we have $b_- \subseteq U \backslash C_U$ from constraints (b) and (c) of Definition 2. Therefore,

$$\Delta_{C_U,U} = \mu_{C_U} - \mu_U = \mu_{C_U \backslash U} - \mu_{U \backslash C_U} = \underbrace{(\mu_{C_U \backslash U} - \mu_{b_+})}_{\mu_{d_+}} + \mu_{b_+} - \underbrace{(\mu_{U \backslash C_U} - \mu_{b_-})}_{\mu_{d_-}} - \mu_{b_-} = \Delta_{b_+,b_-} + \Delta_{d_+,d_-},$$

which proves (24).

Now let us turn to showing (25):

$$\begin{aligned}
\overline{d}_{b_+,b_-} + \overline{d}_{d_+,d_-} = |b_+ \oplus b_-| + |d_+ \oplus d_-| &\overset{\text{(a)}}{=} |b_+| + |b_-| + |d_+| + |d_-| \\
&\overset{\text{(b)}}{=} |b_+| + |b_-| + |C_U \backslash U \backslash b_+| + |U \backslash C_U \backslash b_-| \\
&\overset{\text{(c)}}{=} |b_+| + |b_-| + |C_U \backslash U| - |b_+| + |U \backslash C_U| - |b_-| = |U \oplus C_U| = \overline{d}_{U,C_U},
\end{aligned}$$

where **(a)** comes from the fact that $b_+ \cap b_- = d_+ \cap d_- = \varnothing$, **(b)** is from the definition of d, and **(c)** follows from constraints (b) and (c) of Definition 2. □

We are now ready to prove Lemma 2.

Proof of Lemma 2. The proof is by contradiction. Suppose $U \not\perp \mathsf{C}_U$. Since independence is symmetric, we also have $\mathsf{C}_U \not\perp U$. This means that there exists a non-empty exchange set $b = (b_+, b_-)$, different than the independent exchange set $(C_U \backslash U, U \backslash C_U)$ of (U, C_U), that satisfies constraints **(b)–(e)** of Definition 2. From the exchange set b, we define the exchange set d that completes the transformation of U to C_U after applying b to U as $d = \big((C_U \backslash U) \backslash b_+, (U \backslash C_U) \backslash b_-\big)$.

Since b satisfies constraints (b) and (c), we may write

$$\mu_{C_U \mp b} = \mu_{C_U} - \mu_{b_+} + \mu_{b_-} = \mu_{C_U} - \Delta_{b_+,b_-},$$

which gives

$$\Delta_{C_U \mp b, U} = \mu_{C_U \mp b} - \mu_U = \Delta_{C_U,U} - \Delta_{b_+,b_-}. \tag{26}$$

Since b is not empty, $C_U \mp b$ is closer to U than C_U, and hence, $\overline{d}_{U,C_U \mp b} < \overline{d}_{U,C_U}$. Now consider the following three cases (note that as shown in Proposition 7, Δ_{b_+,b_-} and Δ_{d_+,d_-} cannot be both negative):

Case 1) $\Delta_{b_+,b_-} \leq 0$

In this case, by (26) we may write

$$H_{U,C_U \mp b} = \frac{\overline{d}^2_{U,C_U \mp b}}{\Delta^2_{U,C_U \mp b}} = \frac{\overline{d}^2_{U,C_U \mp b}}{(\Delta_{U,C_U} - \Delta_{b_+,b_-})^2} \leq \frac{\overline{d}^2_{U,C_U \mp b}}{\Delta^2_{U,C_U}} \overset{\text{(a)}}{<} \frac{\overline{d}^2_{U,C_U}}{\Delta^2_{U,C_U}} \overset{\text{(b)}}{=} \min_{V \in \mathcal{C}: \mu_V > \mu_U} H_{U,V}, \tag{27}$$

where **(a)** comes from the fact that $\overline{d}_{U,C_U \mp b} < \overline{d}_{U,C_U}$ and **(b)** is from the definition of the complement C_U. Moreover, in this case, from (26), we have $\Delta_{C_U,U} \leq \Delta_{C_U \mp b,U}$, which gives us $\mu_{C_U} \leq \mu_{C_U \mp b}$. Since $\mu_U < \mu_{C_U}$ by definition, we have that $(C_U \mp b) \in \{V \in \mathcal{C} : \mu_V > \mu_U\}$ and hence $H_{U,C_U \mp b} \geq H_{U,C_U}$ by definition of C_U, which contradicts equation (27).

Case 2) $\Delta_{b_+,b_-} > 0$ and $\Delta_{d_+,d_-} \leq 0$

Here we first show that

$$\mu_{U \pm b} = \mu_{U \backslash b_- \cup b_+} \overset{\text{(a)}}{=} \mu_U - \mu_{b_-} + \mu_{b_+} = \mu_U + \Delta_{b_+,b_-},$$

where **(a)** comes from constraints (b) and (c) of Definition 2, which gives us

$$\Delta_{U \pm b, U} = \mu_{U \pm b} - \mu_U = \Delta_{b_+,b_-}. \tag{28}$$

It is also straightforward to see that

$$\overline{d}_{U,U \pm b} = |U \oplus U \pm b| = \overline{d}_{b_+,b_-}. \tag{29}$$

Now similar to (27), we may write

$$H_{U,U \pm b} = \frac{\overline{d}^2_{U,U \pm b}}{\Delta^2_{U,U \pm b}} \overset{\text{(a)}}{=} \frac{\overline{d}^2_{b_+,b_-}}{\Delta^2_{b_+,b_-}} \overset{\text{(b)}}{<} \frac{(\overline{d}_{b_+,b_-} + \overline{d}_{d_+,d_-})^2}{(\Delta_{b_+,b_-} + \Delta_{d_+,d_-})^2} \overset{\text{(c)}}{=} H_{U,C_U} = \min_{V \in \mathcal{C}: \mu_V > \mu_U} H_{U,V}, \tag{30}$$

where **(a)** comes from (28) and (29), **(b)** is from the fact that $\overline{d}_{d_+,d_-} > 0$ and $\Delta_{d_+,d_-} \leq 0$, and finally **(c)** is from Proposition 7. Moreover, since $\Delta_{b_+,b_-} > 0$, from (28) we have $\mu_U < \mu_{U \pm b}$, which means that $(U \pm b) \in \{V \in \mathcal{C} : \mu_V > \mu_U\}$, and thus, $H_{U,U \pm b}$ should be bigger than or equal to H_{U,C_U}, which contradicts equation (30).

Case 3) $\Delta_{b_+,b_-} > 0$ and $\Delta_{d_+,d_-} > 0$

From Proposition 7, we have

$$H_{U,C_U} = \frac{(\overline{d}_{b_+,b_-} + \overline{d}_{d_+,d_-})^2}{(\Delta_{b_+,b_-} + \Delta_{d_+,d_-})^2} \overset{\text{(a)}}{\geq} \min\left(\frac{\overline{d}^2_{b_+,b_-}}{\Delta^2_{b_+,b_-}}, \frac{\overline{d}^2_{d_+,d_-}}{\Delta^2_{d_+,d_-}}\right) = \min(H_{b_+,b_-}, H_{d_+,d_-}), \tag{31}$$

where **(a)** comes from footnote 10 Appendix C. The inequality in (31) is strict whenever $H_{b_+,b_-} \neq H_{d_+,d_-}$. We now consider the following three cases that all end up contradicting that $\mathsf{C}_U = \underset{V \in \mathcal{C}: \mu_V > \mu_u}{\arg\min} H_{U,V}$.

Case 3.1) $H_{b_+,b_-} < H_{d_+,d_-}$

From equation 31, we have $H_{U,C_U} > H_{b_+,b_-} \overset{\text{(a)}}{=} H_{U\pm b,U}$, where **(a)** is from (28) and (29). At the same time we have $H_{U,C_U} = \underset{V:\mu_V>\mu_U}{\arg\min}\, H_{U,V} \overset{\text{(b)}}{\leq} H_{U\pm b,U}$, which leads to a contradiction. Note that **(b)** holds because $\Delta_{b_+,b_-} > 0$, and thus, $\mu_U < \mu_{U\pm b}$ from (28).

Case 3.2) $H_{b_+,b_-} > H_{d_+,d_-}$

From equation 31, we have $H_{U,C_U} > H_{d_+,d_-}$. Since $d_- \subseteq U$ and $d_+ \cap U = \varnothing$ from the definition of d, we may write

$$\Delta_{U\pm d,U} = \mu_{U\pm d} - \mu_U = \mu_U - \mu_{d_-} + \mu_{d_+} - \mu_U = \Delta_{d_+,d_-},$$

and

$$\overline{d}_{U\pm d,U} = \overline{d}_{d_+,d_-},$$

which gives us $H_{U,C_U} > H_{d_+,d_-} = H_{U\pm d,U}$. At the same time $H_{U,C_U} = \underset{V:\mu_V>\mu_U}{\arg\min}\, H_{U,V} \leq H_{U\pm d,U}$, which leads to a contradiction.

Case 3.3) $H_{b_+,b_-} = H_{d_+,d_-}$

From equation 31, we have $H_{U,C_U} \geq H_{b_+,b_-}$. If the inequality is strict, then we obtain the contradiction as in Case 3.1. Thus let us assume that $H_{U,C_U} = H_{b_+,b_-}$. From (28) and (29), we have $H_{U,C_U} = H_{b_+,b_-} = H_{U,U\pm b}$, and from (28) and the fact that $\Delta_{b_+,b_-} > 0$, we have $\mu_{U\pm b} > \mu_U$, and from (25) and (29), we have $\overline{d}_{U\pm b,U} < \overline{d}_{U,C_U}$. This creates a contradiction because in case $H_{U,C_U} = H_{U,U\pm b}$, according to Definition 4, the tie should be broken in favor of the set with the smaller symmetric distance, and thus, C_U should be $U \pm b$. □

F Proof of Theorem 3

We begin this section with the definition of the $*$-complement of a decision set. For this, let $Q(U)$ be the set of decision sets V such that $U \perp V$ and the exchange set $b = (V\backslash U, U\backslash V)$ satisfies constraints **(b)–(e)** of Definition 2 for the pair of decision sets (U, U^*).[12]

Definition 8. *The $*$-complement of a decision set $U \in \mathcal{C}$ with $U \neq U^*$, denoted by C^*_U, is defined as*

$$\mathsf{C}^*_U = \underset{V\in Q(U)}{\arg\min}\, H_{U,V}.$$

We have to show that the argument of the *argmin* in Definition 8 is not empty, i.e., $Q(U) \neq \varnothing$. For this purpose, we build a sequence of decision sets $\{V_1, \ldots, V_p\}$ such that $V_1, \ldots, V_{p-1}$ are all not independent of U and $U \perp V_p$, that is, we stop the sequence as soon as we reach a decision set V_p independent from U. To construct such a sequence, we start with $V_1 = U^*$ and for $k \in \{1, \ldots, p-1\}$, we generate $V_{k+1} = U \pm b_{k+1}$, where $b_1 = (b_{1,+}, b_{1,-}) = (U^*\backslash U, U\backslash U^*)$ and $b_{k+1} = (b_{k+1,+}, b_{k+1,-}) \subset b_k = (b_{k,+}, b_{k,-})$ is an exchange set that satisfies constraints (b)-(e) of Definition 2 for the pair of decision sets (U, V_k). Note that b_{k+1} exists by definition as V_k is not independent of U and this is why we can build iteratively the sequence $\{V_1, \ldots, V_p\}$ until we find V_p with $U \perp V_p$. Since $V_{k+1} = U \pm b_{k+1}$, we have $|V_{k+1} \oplus U| = |(b_{k+} \oplus b_{k-})| < |V_k \oplus U|$, which means that the size of the exchange sets b_k is decreasing, and thus, the sequence eventually has to end. From the construction of the b_k's, it is clear that they are all subsets of $b_1 = (U^*\backslash U, U\backslash U^*)$, and thus, $(V_p\backslash U, U\backslash V_p)$ satisfies constraints **(b)–(e)** of Definition 2 for the pair of decision sets (U, U^*). This proves that $Q \neq \varnothing$ and the argument of the *argmin* in Definition 8 is not empty. Also, note that $\mu_{\mathsf{C}^*_U} > \mu_U$ as intuitively C^*_U is a decision set made by replacing parts of U by parts of U^*.

We are now ready to give the proof of Theorem 3.

Proof of Theorem 3. We only consider the case where $i \notin U^*$ in detail, the case $i \in U^*$ is symmetric. Let $H*_i$ and $\mathcal{H}*^2_i$ be defined as H_i and $\mathcal{H}^2_i$, respectively, but using the $*$-complement C^*_U of Definition 8 instead of C_U. Then similar to Lemma 3 one can show the equivalence of the $*$-complement complexities, i.e., $\mathcal{H}*^2_i = H*_i$.

[12]Note that since $U \perp V$, the exchange set $b = (V\backslash U, U\backslash V)$ is the only non-empty exchange set that satisfies constraints **(b)–(e)** of Definition 2 for the pair of decision sets (U, C_U).

Therefore, we have the following series of inequalities

$$H_i \overset{(a)}{=} \mathcal{H}_i^2 \overset{(b)}{=} \max_{U \in \mathcal{C}: i \in U} H_{U,\mathsf{C}_U} \overset{(c)}{\leq} \max_{U \in \mathcal{C}: i \in U} H_{U,\mathsf{C}_U^*} \overset{(d)}{=} \max_{U \in \mathcal{C}: i \in U \setminus \mathsf{C}_U^*} H_{U,\mathsf{C}_U^*} \overset{(e)}{=} H_{U_i,V_i},$$

where **(a)** holds by Lemma 3, **(b)** uses the definition of $\mathcal{H}_i^2$, **(c)** uses $H_{U,C_U} = \operatorname{arg\,min}_{V:\mu_V > \mu_U} H_{U,V} \leq H_{U,\mathsf{C}_U^*}$ as $\mu_{\mathsf{C}_U^*} > \mu_U$, **(d)** uses the equivalence of complexity based on the $*$-complement, **(e)** introduces U_i to denote the decision set attaining the maximum in the above equation and $V_i = \mathsf{C}_{U_i}^*$. By the definition of the $*$-complement, $b' = (V_i \backslash U_i, U_i \backslash V_i)$ satisfies constraints **(b)**-**(e)** of Definition 2 for the pair of decision sets (U, U^*). As a result, $U^* \mp b' \in \mathcal{C}$ and $i \in U^* \mp b'$ (see constraint (e) in Definition 2 and $i \in b'\_$). By the definition of b', we have $\mu_{U_i} = \mu_{V_i} + \mu_{b'_+} - \mu_{b'_-}$, and thus, we may write

$$\Delta_{V_i,U_i} = \mu_{b'_+} - \mu_{b'_-} = \mu^* - (\mu^* - \mu_{b'_+} + \mu_{b'_-}) = \Delta_{U^*,U^* \mp b'} \geq \min_{U: i \in U} \Delta_{U^*,U},$$

where the last inequality follows from the fact that $i \in U^* \mp b'$. We note that for any independent pair of sets such as V_i, U_i, any well defined exchange class $\mathcal{B}$ should include the (unique) exchange set $b' = (V_i \backslash U_i, U_i \backslash V_i)$ that allows to move from one set to another. As a result for any exchange class $\mathcal{B}$, $\text{width}(\mathcal{B}) = \max_{(b_+,b_-) \in \mathcal{B}} |b_+| + |b_-| \geq |b'_+| + |b'_-| = |U_i \oplus V_i| = \overline{d}_{U_i,V_i}$. Therefore, $\text{width}(\mathcal{C}) = \min_{\mathcal{B} \in \text{Exchange}(\mathcal{C})} \text{width}(\mathcal{B}) \geq \overline{d}_{U_i,V_i}$, which together with $\Delta_{V_i,U_i} \geq \min_{U: i \in U} \Delta_{U^*,U}$ leads to the desired outcome

$$H_i \leq \frac{\overline{d}_{U_i,V_i}^2}{\Delta_{V_i,U_i}^2} \leq \frac{\text{width}(\mathcal{C})^2}{\min_{U: i \in U} \Delta_{U^*,U}^2} = H_i^{\odot}.$$

□

G Fixed Budget Results: Proof of Theorem 1

In this section, we provide a proof of Theorem 1. In the following, we will mainly work with the complexity $\mathcal{H}_i^2$ (and the corresponding simplicity G) as defined in equation (7). Recall that this formulation is equivalent to H_i. In the following, we use $[N]$ to denote the set $\{1, 2, \ldots, N\}$. We also introduce two numerical constants $0 < c_1 < 1$ and $0 < c_2 < 1/2$ such that $c_2 \geq \frac{c_1}{1-2c_2} \geq c_2$, whose exact values will be chosen later. Finally, we consider a permutation π of the arms that orders the arms with respect to the values G_i, that is, $G_{\pi(1)} \geq G_{\pi(2)} \geq \ldots G_{\pi(K)}$. To simplify notation, in the following, we will simply write $G_{(i)}$ instead of $G_{\pi(i)}$.

We now introduce a high-probability event which serves as a basis for the proof of the correctness of the algorithm. This event states that at the end of each phase k the estimated values of the arms will differ from their real values by at most $G(k)$.

Lemma 4. *Let $G_{(1)} \geq G_{(2)} \geq \ldots \geq G_{(K)}$ be an ordering of arms by decreasing complexity [13]. The event ξ defined as*

$$\xi = \left\{ \forall i \in \mathcal{K}, k \in [K], \ \ |\widehat{\mu}_i(k) - \mu_i| \leq c_1 G_{(k)} \right\} \tag{32}$$

holds with probability

$$\mathbb{P}(\xi) \geq 1 - 2K^2 \exp\left(-\frac{2c_1^2(n-K)}{\overline{\log}(K)\overline{H}} \right).$$

Proof. By Hoeffding's inequality and a union bound, the probability of the complementary event $\overline{\xi}$ of ξ can be bounded as

[13] Notice that the i-th simplest arm is the $(K+1-i)$-th most complex arm.

Parameters: number of rounds n, set of arms $\mathcal{K}$, decision set $\mathcal{C}$, and cumulative pulls scheme $n_0, n_1, \ldots, n_K$.
Let $\mathcal{K}_1 = \mathcal{K}$, $k = 1$, and $J_n = \varnothing$
while $|\mathcal{K}_k| \geq 1$ **do**
  Pull each arm $i \in \mathcal{K}_k$ for $n_k - n_{k-1}$ rounds.
  Compute $\widehat{U}^*(k) = \arg\max_{U \in \mathcal{C}} \widehat{\mu}_U(k)$.
  Find $j_k = \max_{i \in \mathcal{K}_k} \widehat{G}_i(k)$.
  if $j_k \in \widehat{U}^*(k)$ **then**
    The arm j_k is accepted and $J_n = J_n \cup \{j_k\}$.
  end if
  Deactivate arms j_k, i.e., set $\mathcal{K}_{k+1} = \mathcal{K}_k \backslash j_k$.
  $k \longleftarrow k + 1$
end while
Return J_n

Figure 5: The modified fixed budget algorithm.

follows, provided we use the proposed pulls scheme $n_k = \left\lceil \frac{n-K}{\overline{\log}(K)(K+1-k)} \right\rceil$, $k \in \mathcal{K}$:

$$\begin{aligned}
\mathbb{P}(\bar{\xi}) &= \sum_{i=1}^{K} \sum_{k=1}^{K} \mathbb{P}\left(|\widehat{\mu}_i(k) - \mu_i| > c_1 G_{(k)}\right) \\
&\leq \sum_{i=1}^{K} \sum_{k=1}^{K} 2\exp\left(-2n_k c_1^2 G_{(k)}^2\right) \\
&\leq \sum_{i=1}^{K} \sum_{k=1}^{K} 2\exp\left(-\frac{2c_1^2(n-K)}{\overline{\log}(K)(K+1-k)H_{(K+1-k)})}\right) \\
&\leq 2K^2 \exp\left(-\frac{2c_1^2(n-K)}{\overline{\log}(K)\overline{H}}\right).
\end{aligned}$$

□

For the proof of Theorem 1 we analyze a slightly modified algorithm described in Figure 5, where for each arm that is deactivated it is immediately (and not only after all arms have been deactivated) decided, whether it shall be contained in the set returned by the algorithm at the end. On the event ξ, the correctness of both algorithms is the same, which can be deduced from statement (ii) of the induction hypothesis in Definition 9 below.

We will now prove Theorem 1 by showing that on event ξ, the optimal set U^* is identified at the end of the phases. That is, the algorithm neither accepts an arm not in U^*, nor is any arm in U^* rejected.

G.1 The Induction Hypothesis

The proof proceeds by induction over the phases of the algorithm. We first introduce the induction hypothesis.

Definition 9. *The induction hypothesis is defined by the two following properties. At the beginning of phase k we have:*

(i) *All accepted arms belong to the optimal set, i.e., $J_n(k-1) \subseteq U^*$, and all rejected arms (i.e., arms which have been deactivated but never accepted) do not belong to the optimal set, i.e., $(\mathcal{K}_k \backslash J_n(k-1)) \cap U^* = \varnothing$.*

(ii) *If arm $i \notin \mathcal{K}_k$ and it has been deactivated during phase $l \in [k-1]$, then $G_i \geq (1 - 2c_2)G_{(l)}$.*

Statement (i) is the classical desired property, while statement (ii) is specific to our approach and implies that by having been pulled n_l times the arm i has been sampled sufficiently often w.r.t. its complexity. Indeed, recall that a set is not necessarily compared to the optimal set U^* whose arms most probably belongs to $\mathcal{K}_k$. Therefore we need to show that an arm i contained in a set V that is likely to be used as a complement set by some "active" set U has been sampled often enough (i.e., proportionally to its complexity H_i), especially if it has been removed in a previous phase $l < k$.

We continue with a few properties implied by the induction hypothesis together with the high-probability event ξ of Lemma 4. We start with concentration inequalities on event ξ for $\widehat{\mu}$, $\widehat{\Delta}$, and $\widehat{G}$.

Proposition 8. *Assume that the induction hypothesis (Definition 9) at the beginning of phase k as well as event ξ hold. Then for any arm $i \in \mathcal{K}$,*

$$|\widehat{\mu}_i(k) - \mu_i| \leq c_1 \max\left\{\frac{G_i}{1-2c_2}, G_{(k)}\right\}. \tag{33}$$

Furthermore for any pair $(U,V) \in \mathcal{C}^2$ such that $V = \mathsf{C}_U$ we have

$$\left|\widehat{\Delta}_{V,U}(k) - \Delta_{V,U}\right| \leq c_2 \overline{d}_{U,V} \max\{G_{V,U}, G_{(k)}\}$$
$$\text{and} \quad \left|\widehat{G}_{V,U}(k) - G_{V,U}\right| \leq c_2 \max\left\{G_{V,U}, G_{(k)}\right\}.$$

where $\widehat{\Delta}_{V,U}(k)$ and $\widehat{G}_{V,U}(k)$ are the gaps and the simplicity computed at the end of the phase k. Finally, for the special case of pairs (U^*, U),

$$\left|\widehat{\Delta}_{U^*,U}(k) - \Delta_{U^*,U}\right| \leq c_2 \overline{d}_{U,U^*} \max\{G_{V,U}, G_{(k)}\}$$
$$\text{and} \quad \left|\widehat{G}_{U^*,U}(k) - G_{U^*,U}\right| \leq c_2 \max\{G_{V,U}, G_{(k)}\},$$

with $V = \mathsf{C}_U$.

Proof. First note that if $i \in \mathcal{K}_k$, then on event ξ we have $|\widehat{\mu}_i(k) - \mu_i| \leq c_1 G_{(k)}$. Thus, let us assume that $i \notin \mathcal{K}_k$. Let l be the phase at which arm i has been deactivated, with $l \in \{1, \dots, k-1\}$. We have

$$|\widehat{\mu}_i(k) - \mu_i| \overset{(a)}{\leq} c_1 G_{(l)} \overset{(b)}{\leq} \frac{c_1}{1-2c_2} G_i,$$

where **(a)** is implied by the event ξ and the fact that $\widehat{\mu}_i(k) = \widehat{\mu}_i(l)$, **(b)** uses property (ii) of the induction hypothesis. Summarizing, independent of whether $i \in \mathcal{K}_k$ or not, we have for any i

$$|\widehat{\mu}_i(k) - \mu_i| \leq c_1 \max\left\{\frac{G_i}{1-2c_2}, G_{(k)}\right\},$$

which shows the first claim. Let us now focus on a pair of sets U, V such that $i \in U \oplus V$ and $V = \mathsf{C}_U$ or $V = U^*$. Then by the previous inequality,

$$\begin{aligned}
|\widehat{\mu}_i(k) - \mu_i| &\leq c_1 \max\left\{\frac{G_i}{1-2c_2}, G_{(k)}\right\} \\
&\overset{(a)}{\leq} \max\left\{c_2 G_i, c_1 G_{(k)}\right\} \\
&\overset{(b)}{\leq} c_2 \max\left\{G_{V,U}, G_{(k)}\right\},
\end{aligned}$$

where **(a)** follows from the choice of the constants such that $c_2 \geq \frac{c_1}{1-2c_2}$. For **(b)** we use Proposition 5 for $V = \mathsf{C}_U$, the definition of $G_{V,U}$, the fact that $\Delta_{V,U} > 0$ as well as $c_1 \leq c_2$. As a result we obtain

$$\begin{aligned}
\left|\widehat{\Delta}_{V,U}(k) - \Delta_{V,U}\right| &\leq \sum_{i \in U \oplus V} |\widehat{\mu}_i(k) - \mu_i| \\
&\leq c_2 \overline{d}_{V,U} \max\{G_{V,U}, G_{(k)}\},
\end{aligned}$$

which proves the first part of the second statement. The second part then simply follows from

$$\begin{aligned}
\widehat{G}_{V,U}(k) &= \frac{\widehat{\Delta}_{V,U}(k)}{\overline{d}_{U,V}} \\
&\overset{(c)}{\geq} \frac{\Delta_{V,U} - c_2 \overline{d}_{V,U} \max\{G_{V,U}, G_{(k)}\}}{\overline{d}_{V,U}} \\
&= G_{V,U} - c_2 \max\{G_{V,U}, G_{(k)}\},
\end{aligned}$$

where **(c)** follows from the first statement. The missing inequality to conclude the second part of the second statement can be obtained analogously. Finally, the last statement follows along the same lines, only replacing Proposition 5 in step **(b)** above by Proposition 6. □

The following Proposition shows that at the beginning of each phase k there is an arm a_k which has a larger simplicity than the k-th largest simplicity. In the remaining proof this arm will serve as a reference arm, as this is the arm that should be deactivated at the end of phase k.

Proposition 9. *Let*

$$a_k = \arg\min_{i \in \mathcal{K}_k} H_i = \arg\max_{i \in \mathcal{K}_k} G_i$$

be the simplest arm among those left at the beginning of phase k. Then $G_{a_k} \geq G_{(k)}$.

Proof. At the beginning of phase k, only $K - |\mathcal{K}_k| = k - 1$ arms have been deactivated, i.e., $|\mathcal{K}_k| = K + 1 - k$. Hence the simplest arm left in $\mathcal{K}_k$ (i.e., a_k) cannot be more difficult than the arm that would be left if all the $k-1$ simpler arms were deactivated (i.e., $G_{(k)}$), which gives the claimed $G_{(k)} \leq G_{a_k}$. □

The following Proposition shows that at the end of each phase k, the reference arm a_k belongs to $\widehat{U}^*$ if and only if it actually belongs to U^*. This allows us to show that the simplicity of a_k can be well estimated at the end of phase k.

Proposition 10. *Assume that the induction hypothesis (Definition 9) at the beginning of phase k as well as event ξ hold. Then $a_k \in U^*$ if and only if $a_k \in \widehat{U}^*(k)$, where $\widehat{U}^*(k)$ is the estimated optimal set at the end of phase k.*

Proof. We prove in detail that $a_k \notin U^*$ implies $a_k \notin \widehat{U}^*(k)$. The reverse can be shown along a similar line of arguments. The proof is by contradiction. Thus, let us assume that $a_k \notin U^*$ and $a_k \in \widehat{U}^*(k)$. Note that this implies that $\widehat{U}^*(k) \neq U^*$. Let $W = \mathsf{C}_{\widehat{U}^*(k)}$ be the complement of the estimated optimal set (note that W exists, since $\widehat{U}^*(k)$ is not optimal). We have

$$\begin{aligned}
\widehat{\Delta}_{W,\widehat{U}^*(k)}(k) &\overset{(a)}{\geq} \Delta_{W,\widehat{U}^*(k)} - c_2 \overline{d}_{W,\widehat{U}^*(k)} \max\left\{G_{W,\widehat{U}^*(k)}, G_{(k)}\right\} \\
&= \overline{d}_{W,\widehat{U}^*(k)} \left(G_{W,\widehat{U}^*(k)} - c_2 \max\left\{G_{W,\widehat{U}^*(k)}, G_{(k)}\right\}\right),
\end{aligned}$$

where **(a)** uses Proposition 8. We consider two cases.

Case 1) $G_{W,\widehat{U}^*(k)} \geq G_{(k)}$

In this case,

$$\begin{aligned}
\widehat{\Delta}_{W,\widehat{U}^*(k)}(k) &\geq \overline{d}_{W,\widehat{U}^*(k)} \left(G_{W,\widehat{U}^*(k)} - c_2 G_{W,\widehat{U}^*(k)}\right) \\
&= \overline{d}_{W,\widehat{U}^*(k)} G_{W,\widehat{U}^*(k)} (1 - c_2) > 0,
\end{aligned}$$

where the last inequality follows from the fact that $0 < c_2 < 1$ and $G_{W,\widehat{U}^*(k)} > 0$ by definition of W. It follows that $\widehat{\mu}_{\widehat{U}^*(k)}(k) < \widehat{\mu}_W(k)$, which contradicts that $\widehat{U}^*(k)$ is the empirical best set.

Case 2) $G_{W,\widehat{U}^*(k)} < G_{(k)}$)

We write

$$\begin{aligned}
\widehat{\Delta}_{W,\widehat{U}^*(k)}(k) &\geq \overline{d}_{W,\widehat{U}^*(k)} \left(G_{W,\widehat{U}^*(k)} - c_2 G_{(k)}\right) \\
&\overset{(b)}{\geq} \overline{d}_{W,\widehat{U}^*(k)} \left(G_{a_k} - c_2 G_{(k)}\right) \\
&\overset{(c)}{\geq} \overline{d}_{W,\widehat{U}^*(k)} G_{(k)} (1 - c_2) > 0,
\end{aligned}$$

where **(b)** holds since $a_k \in \widehat{U}^*(k)$, so that $G_{a_k} = \min_{U: a_k \in U} \max_{V: \mu_V > \mu_U} G_{V,U} \leq \max_{V: \mu_V > \mu_{\widehat{U}^*(k)}} G_{V,\widehat{U}^*(k)} = G_{W,\widehat{U}^*(k)}$. Concerning **(c)**, this holds by Proposition 9, as $G_{a_k} \geq G_{(k)}$. Similar as before we obtain the contradiction $\widehat{\mu}_{\widehat{U}^*(k)}(k) < \widehat{\mu}_W(k)$. □

The following Proposition gives a lower bound on the estimated simplicity of the reference arm a_k at the end of phase k depending on $G(k)$. This will be used later to show that the algorithm does not remove other arms than a_k in each phase k.

Proposition 11. *Assume that the induction hypothesis (Definition 9) at the beginning of phase k as well as event ξ hold. Then $\widehat{G}_{a_k}(k) \geq (1-c_2)G_{(k)}$.*

Proof. We give a detailed proof for the case where $a_k \notin \widehat{U}^*(k)$. The other case follows from symmetric arguments. Let U_k be a set that defines the estimated simplicity of a_k, that is

$$U_k \in \arg\min_{U: a_k \in U} \widehat{G}_{\widehat{\mathsf{C}}_U(k),U}(k),$$

and let $W_k = \mathsf{C}_{U_k}$. Note that W_k is well-defined because $U_k \neq U^*$: Indeed, from Proposition 10, since $a_k \notin \widehat{U}^*(k)$, also $a_k \notin U^*$, so that because $a_k \in U_k$, we have $U_k \neq U^*$.

Then we have

$$\begin{aligned}
\widehat{G}_{a_k}(k) &= \widehat{G}_{\widehat{\mathsf{C}}_U(k),U_k}(k) \\
&= \max_{V: \widehat{\mu}_V(k) > \widehat{\mu}_{U_k}(k)} \widehat{G}_{V,U_k}(k) \\
&\geq \widehat{G}_{W_k,U_k}(k) \\
&\overset{(a)}{\geq} G_{W_k,U_k} - c_2 \max\left\{G_{W_k,U_k}, G_{(k)}\right\},
\end{aligned}$$

where **(a)** follows from Proposition 8. We have, as $a_k \notin U^*$, that

$$G_{a_k} = \min_{U: a_k \in U} G_{\mathsf{C}_U,U} \leq G_{W_k,U_k}.$$

Furthermore, from Proposition 9 we have that $G_{(k)} \leq G_{a_k}$ and thus $G_{(k)} \leq G_{W_k,U_k}$. Thus the previous expression simplifies to

$$\widehat{G}_{a_k}(k) \geq G_{W_k,U_k} - c_2 G_{W_k,U_k} \geq (1-c_2)G_{(k)}.$$

□

The following lemma is rather technical. It shows that if the estimated simplicity of a sub-optimal decision U_k is defined with respect to V_k, then this estimated simplicity will be larger than the true simplicity of the decision set V_k. The proof shows that if the estimated simplicity of U_k were smaller than the true simplicity of V_k, then it would surely be also smaller than the estimated simplicity of V_k defined with respect to W_k, leading to a contradiction. For notational convenience, in the following we will drop the dependency of the estimated quantities on the phase k (e.g., write $\widehat{G}_{V_k,U_k}$ instead of $\widehat{G}_{V_k,U_k}(k)$).

Proposition 12. *Assume that the induction hypothesis (Definition 9) at the beginning of phase k as well as event ξ hold. Further assume that $U_k, V_k, W_k \in \mathcal{C}$ such that $U_k \neq \widehat{U}^*(k)$, $V_k = \widehat{\mathsf{C}}_{U_k}(k) \neq U^*$, $W_k = \mathsf{C}_{V_k}$, and $G_{W_k,V_k} \geq G_{(k)}$. Then $\widehat{G}_{V_k,U_k}(k) \geq (1-c_2)G_{W_k,V_k}$.*

Proof. We start by showing that $\widehat{\mu}_{U_k} < \widehat{\mu}_{V_k} < \widehat{\mu}_{W_k}$. First, $\widehat{\mu}_{U_k} < \widehat{\mu}_{V_k}$ comes from the definition of V_k as the (estimated) complement of U_k. Furthermore,

$$\begin{aligned}
\widehat{\Delta}_{W_k,V_k} &\overset{(a)}{\geq} \Delta_{W_k,V_k} - c_2 \overline{d}_{W_k,V_k} \max\{G_{W_k,V_k}, G_{(k)}\} \\
&\overset{(b)}{\geq} \Delta_{W_k,V_k} - c_2 \overline{d}_{W_k,V_k} G_{W_k,V_k} \\
&\overset{(c)}{=} \Delta_{W_k,V_k} - c_2 \Delta_{W_k,V_k} > 0,
\end{aligned}$$

where **(a)** follows from Proposition 8 and the fact that $\Delta_{W_k,V_k} > 0$ (since W_k is the (exact) complement of V_k), **(b)** follows from the assumption that $G_{W_k,V_k} \geq G_{(k)}$, and **(c)** is obtained from the definition of simplicity G_{W_k,V_k} and the fact that $0 < c_2 < 1$. This completes the proof of the claim that $\widehat{\mu}_{U_k} < \widehat{\mu}_{V_k} < \widehat{\mu}_{W_k}$.

Next, we show that $\widehat{G}_{W_k,V_k} \leq \widehat{G}_{V_k,U_k}$. First note that by Proposition 3[14] we obtain from $\widehat{\mu}_{U_k} < \widehat{\mu}_{V_k} < \widehat{\mu}_{W_k}$ that $\widehat{G}_{W_k,U_k} \geq \min\left\{\widehat{G}_{W_k,V_k}, \widehat{G}_{V_k,U_k}\right\}$, where the inequality is strict whenever $\widehat{G}_{W_k,V_k} \neq \widehat{G}_{V_k,U_k}$. Now if we assume that $\widehat{G}_{W_k,V_k} > \widehat{G}_{V_k,U_k}$, the previous inequality becomes strict as $\widehat{G}_{W_k,U_k} > \widehat{G}_{V_k,U_k}$. Then we would obtain the contradiction

$$\widehat{G}_{V_k,U_k} = \max_{V:\widehat{\mu}_V > \widehat{\mu}_U} \widehat{G}_{V,U_k} \geq \widehat{G}_{W_k,U_k} > \widehat{G}_{V_k,U_k},$$

thus implying the claimed $\widehat{G}_{W_k,V_k} \leq \widehat{G}_{V_k,U_k}$.

We now can conclude with

$$\begin{aligned}\widehat{G}_{V_k,U_k} &\geq \widehat{G}_{W_k,V_k} \\ &\overset{(a)}{\geq} G_{W_k,V_k} - c_2 \max\{G_{W_k,V_k}, G_{(k)}\} \\ &\overset{(b)}{\geq} G_{W_k,V_k} - c_2 G_{W_k,V_k},\end{aligned}$$

where **(a)** follows from Proposition 8 and the fact that $\Delta_{W_k,V_k} > 0$ and **(b)** holds due to the assumption that $G_{W_k,V_k} \geq G_{(k)}$. □

G.2 The Induction Step

We now move on to prove the induction step. We do this in two separate lemmas, one for each property in the induction hypothesis.

Lemma 5. *Assume that the induction hypothesis at the beginning of phase k as well as event ξ hold. Then property (i) of Definition 9 holds at phase $k+1$ as well.*

Proof. Property (i) states that no error is made until the beginning of phase k. To prove that this is still true at the beginning of phase $k+1$ we need to prove that during phase k no error is made. An error occurs when either the algorithm deactivates and rejects an arm $j_k \in U^*$, or the algorithm deactivates and accepts an arm $j_k \notin U^*$. We show by contradiction that both cases cannot happen.

We give a detailed proof for the case where $j_k \in U^*$ is rejected. The argument for the second source of error, when $j \notin U^*$ is accepted, is similar.

The strategy of the proof will be to compare the estimated simplicity of the rejected arm to the simplicity of the arm a_k, that is, the arm with the highest simplicity at the end of phase k, and which should be the targeted arm to be deactivated. Using Proposition 11, we know that the simplicity of a_k is of order of $G_{(k)}$. Therefore it remains to prove that G_{j_k} is smaller than $G_{(k)}$.

As j_k is rejected, $j_k \notin \widehat{U}^*(k)$ and hence, $U^* \neq \widehat{U}^*$. Furthermore, $a_k \neq j_k$, as otherwise Proposition 10 would imply that $j_k \in \widehat{U}^*$, a contradiction to the assumption that arm j_k is rejected.
As j_k has been deactivated during phase k, we have $\widehat{G}_{j_k} \geq \widehat{G}_{a_k}$, since the algorithm deactivates the arm with the largest simplicity, that is, $j_k = \arg\max_{i \in \mathcal{K}_k} \widehat{G}_i$. Let $V_k = \widehat{\mathsf{C}}_{U^*}$ be the estimated complement of the optimal set (which exists, as $U^* \neq \widehat{U}^*$) and $W_k = \mathsf{C}_{V_k}$ be the (exact) complement of V_k. Then

$$\begin{aligned}\widehat{G}_{j_k} &= \min_{U: j \in U} \max_{V:\widehat{\mu}_V > \widehat{\mu}_U} \widehat{G}_{V,U} \\ &\overset{(a)}{\leq} \widehat{G}_{V_k,U^*} \\ &\overset{(b)}{\leq} G_{V_k,U^*} + c_2 \max\left\{G_{W_k,V_k}, G_{(k)}\right\} \\ &\overset{(c)}{\leq} c_2 \max\left\{G_{W_k,V_k}, G_{(k)}\right\},\end{aligned}$$

where **(a)** is because $j_k \in U^*$ and V_k is the (estimated) complement of U^*, **(b)** holds by the last statement of Proposition 8, and **(c)** follows from $\mu_{V_k} < \mu^*$, whence $\Delta_{V_k,U^*} < 0$ and $G_{V_k,U^*} < 0$.

[14] More precisely, we rely on an equivalent version based on estimated values and estimated simplicity.

We now show by contradiction that $G_{W_k,V_k} \leq G_{(k)}$. Thus, assume that $G_{W_k,V_k} > G_{(k)}$. Then by Proposition 12, $\widehat{G}_{V_k,U^*} \geq (1-c_2)G_{W_k,V_k}$. Together with the previous inequality this gives

$$(1-c_2)G_{W_k,V_k} \leq \widehat{G}_{V_k,U^*} \leq c_2 \max\{G_{W_k,V_k}, G_{(k)}\} \leq c_2 G_{W_k,V_k},$$

which is a contradiction, since $c_2 < 1/2$, and we conclude that $G_{W_k,V_k} \leq G_{(k)}$.

Now using inequality (c) from before, we get $\widehat{G}_{j_k} \leq c_2 G_{(k)}$. Using Proposition 11, for a_k we have that $\widehat{G}_{a_k} \geq (1-c_2)G_{(k)}$. Since $c_2 < 1/2$, it holds that $\widehat{G}_{j_k} \leq c_2 G_{(k)} < (1-c_2)G_{(k)} \leq \widehat{G}_{a_k}$. As a result $\widehat{G}_{j_k} < \widehat{G}_{a_k}$, which contradicts the fact that j_k would be the deactivated arm during phase k, as by definition of $j_k = \arg\max_{i \in \mathcal{K}_k} \widehat{G}_i$. Thus we can conclude that the algorithm does not reject any arm from U^*.

□

Lemma 6. *Assume that the induction hypothesis at the beginning of phase k as well as event ξ hold. Then property (ii) of Definition 9 holds at phase $k+1$ as well.*

Proof. Let j_k be the arm which is deactivated at the end of phase k, i.e., $j_k = \arg\max_{i \in \mathcal{K}_k} \widehat{G}_i(k)$. Again, we only consider the case where $j_k \notin U^*$, as the proof for the case $j_k \in U^*$ is symmetrical. We have to show that j_k satisfies $G_{j_k} \geq (1-2c_2)G_{(k)}$, which will be done in five steps.

Step 1. We first notice that Lemma 5 implies that no error is made during phase k, since property (i) still holds at the beginning of phase $k+1$. As a result, we have that $j_k \in U^*$ if and only if $j_k \in \widehat{U}^*$ which means in our current case that $j_k \notin \widehat{U}^*$. Let U_k and V_k be the sets which define the (exact) simplicity of j_k, that is,

$$U_k = \arg \min_{U : j_k \in U} G_{\mathsf{C}_U, U}$$

and $V_k = \mathsf{C}_{U_k}$, so that $G_{j_k} = G_{V_k,U_k}$. Further, let $W_k = \widehat{\mathsf{C}}_{U_k}$, noting that W_k is well defined since $U_k \neq \widehat{U}^*$. Indeed, $j_k \notin U^*$ and $j_k \in U_k$, whence $U_k \neq U^*$.

We claim that $G_{W_k,U_k} \leq G_{j_k}$. Indeed, if $\mu_{W_k} \leq \mu_{U_k}$ then we trivially have $G_{W_k,U_k} \leq 0 \leq G_{j_k}$. Furthermore, if $\mu_{W_k} > \mu_{U_k}$ we have by definition of U_k and V_k,

$$G_{W_k,U_k} \leq \max_{W : \mu_W > \mu_{U_k}} G_{W,U_k} = G_{V_k,U_k} = G_{j_k},$$

which proves our claim $G_{W_k,U_k} \leq G_{j_k}$.

Step 2. Next, we note that

$$\widehat{G}_{j_k} = \min_{U : j_k \in U} \widehat{G}_{\widehat{\mathsf{C}}_U, U} \leq \max_{\widehat{\mu}_V > \widehat{\mu}_{U_k}} \widehat{G}_{V,U_k} = \widehat{G}_{W_k,U_k}. \tag{34}$$

Step 3. Here we show that $G_i \leq \max\{G_{j_k}, G_{Z_k,W_k}\}$ for $i \in W_k \oplus U_k$. We distinguish the following cases:

Case 1) $i \in W_k \backslash U_k$
Case 1.1) $i \in U^*$: In this case $i \in U_k \oplus U^*$ and thus we can apply Proposition 6 to U_k and $\mathsf{C}_{U_k} = V_k$ and obtain $G_i \leq G_{V_k,U_k} = G_{j_k}$.
Case 1.2) $i \notin U^*$: Let $Z_k = \mathsf{C}_{W_k}$, noting that since i is in W_k but not in U^*, we have $W_k \neq U^*$. Then

$$G_i = \min_{U : i \in U} G_{\mathsf{C}_U, U} \leq G_{\mathsf{C}_{W_k}, W_k} = G_{Z_k,W_k},$$

where the first equality follows from the fact that $i \notin U^*$, while the inequality is due to the fact that $i \in W_k$.

Case 2) $i \in U_k \backslash W_k$
Case 2.1) $i \in U^*$: Let $Z_k = \mathsf{C}_{W_k}$, noting that since i is in U^* but not in W_k, it holds that $W_k \neq U^*$. Then

$$G_i = \min_{U : i \notin U} G_{\mathsf{C}_U, U} \leq G_{\mathsf{C}_{W_k}, W_k} = G_{Z_k,W_k},$$

where the first equality follows from the fact that $i \in U^*$, while the inequality is due to the fact that $i \notin W_k$.
Case 2.2) $i \notin U^*$: In this case we have by definition that

$$G_i = \min_{U: i \in U} G_{\mathsf{C}_U, U} \leq G_{\mathsf{C}_U, U_k} = G_{V_k, U_k} = G_{j_k},$$

which completes the proof that $G_i \leq \max\{G_{j_k}, G_{Z_k, W_k}\}$ for $i \in W_k \oplus U_k$.

Step 4. Next, we are going to show that $\widehat{G}_{j_k} \leq G_{j_k} + c_2 \max\{G_{j_k}, G_{(k)}\}$, a version of Proposition 8 for arm j_k. We distinguish two cases:

Case 1) $W_k = U^*$

In this case, we have

$$\widehat{G}_{j_k} \overset{(a)}{\leq} \widehat{G}_{W_k, U_k} \overset{(b)}{\leq} G_{W_k, U_k} + c_2 \max\{G_{V_k, U_k}, G_{(k)}\} \overset{(c)}{\leq} G_{j_k} + c_2 \max\left\{G_{j_k}, G_{(k)}\right\},$$

where **(a)** is obtained from equation (34), **(b)** is a result of Proposition 8, and **(c)** follows from $G_{W_k, U_k} \leq G_{j_k}$ of Step 1.

Case 2) $W_k \neq U^*$

Let $Z_k = \mathsf{C}_{W_k}$. We prove by contradiction that $G_{Z_k, W_k} \leq \max\left\{G_{V_k, U_k}/(1-2c_2), G_{(k)}\right\}$. Thus, assume that $G_{Z_k, W_k} > \max\left\{G_{V_k, U_k}/(1-2c_2), G_{(k)}\right\}$. Then

$$\begin{aligned}
\widehat{G}_{W_k, U_k} &\overset{(a)}{\leq} G_{W_k, U_k} + c_1 \frac{1}{\overline{d}_{U_k \oplus W_k}} \sum_{i \in U_k \oplus W_k} \max\left\{\frac{G_i}{1-2c_2}, G_{(k)}\right\} \\
&\overset{(b)}{\leq} G_{W_k, U_k} + c_1 \max\left\{\frac{G_{j_k}}{1-2c_2}, \frac{G_{Z_k, W_k}}{1-2c_2}, G_{(k)}\right\} \\
&\overset{(c)}{\leq} G_{V_k, U_k} + c_1 \frac{G_{Z_k, W_k}}{1-2c_2} \\
&\overset{(d)}{\leq} G_{V_k, U_k} + c_2 G_{Z_k, W_k}, \qquad (35)
\end{aligned}$$

where **(a)** is using for all $i \in U_k \oplus W_k$ the Equation 33 in Proposition 8, **(b)** follows from Step 3, and **(c)** is obtained by $G_{W_k, U_k} \leq G_{j_k} = G_{V_k, U_k}$ (Step 1) and the assumption on G_{Z_k, W_k}, finally **(d)** is obtained by $\frac{c_1}{1-2c_2} \leq c_2$. By Proposition 12, we have $\widehat{G}_{W_k, U_k} \geq (1-c_2) G_{Z_k, W_k}$, which together with (35) gives

$$(1-c_2) G_{Z_k, W_k} \leq \widehat{G}_{W_k, U_k} \leq G_{V_k, U_k} + c_2 G_{Z_k, W_k} \leq c_2 G_{Z_k, W_k},$$

which is a contradiction due to $0 < c_2 < 1/2$. This finishes the proof of $G_{Z_k, W_k} \leq \max\{G_{V_k, U_k}/(1-2c_2), G_{(k)}\}$.

By (35) and the results of Steps 2 and 1, we finally get

$$\widehat{G}_{j_k} \leq \widehat{G}_{W_k, U_k} \leq G_{W_k, U_k} + c_1 \max\left\{\frac{G_{j_k}}{1-2c_2}, \frac{G_{Z_k, W_k}}{1-2c_2}, G_{(k)}\right\} \leq G_{j_k} + c_2 \max\left\{G_{j_k}, G_{(k)}\right\}.$$

Step 5. From Proposition 11 and the fact that j_k is the deactivated arm (i.e., $j_k = \arg\max_{i \in \mathcal{K}_k} \widehat{G}_i$) we have

$$(1-c_2) G_{(k)} \leq \widehat{G}_{a_k} \leq \widehat{G}_{j_k} \leq G_{j_k} + c_2 \max\left\{G_{j_k}, G_{(k)}\right\}.$$

We conclude by considering the two possible cases for the $\max$ term.
Case 1) $G_{j_k} > G_{(k)}$
We have $(1-c_2) G_{(k)} \leq G_{j_k} + c_2 G_{j_k}$. Since $\frac{1-c_2}{1+c_2} \geq 1 - 2c_2$, we get

$$G_{j_k} \geq \frac{1-c_2}{1+c_2} G_{(k)} \geq (1-2c_2) G_{(k)}.$$

Case 2) $G_{j_k} \leq G_{(k)}$
Here we have $(1-c_2) G_{(k)} \leq G_{j_k} + c_2 G_{(k)}$, whence

$$(1-2c_2) G_{(k)} \leq G_{j_k},$$

which concludes the proof. □

G.3 Proof of Theorem 1

With the results of the previous sections, the proof of Theorem 1 is immediate. First, assume that event ξ holds. Note that properties (i) and (ii) of the induction assumption hold for phase $k = 1$. Lemmas 5 and 6 prove the induction step, showing that properties (i) and (ii) hold for all phases k. It remains to consider the error probability for event ξ. This is handled by Lemma 4, where we finally choose $c_1 = 1/8$, $c_2 = 1/4$ so that $0 < c_1 < 1$, $0 < c_2 < 1/2$ and $c_2 \geq \frac{c_1}{1-2c_2}$. □

H Fixed Confidence Results: Proof of Theorem 2

We first introduce a high-probability event corresponding to the confidence bounds used by our algorithm.

Lemma 7. *The event ξ defined as*

$$\xi = \{\forall i \in \mathcal{K}, \forall t > 0, |\widehat{\mu}_i(t) - \mu_i| \leq \beta_i(t-1)\} \tag{36}$$

holds with probability $1 - \delta$, where

$$\widehat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{t=1}^{T_i(t)} X_{i,t} \quad \textit{and} \quad \beta_i(t-1) = \sqrt{\frac{\log \frac{4Kt^2}{\delta}}{2T_i(t)}}.$$

Proof. By Chernoff-Hoeffding's inequality, the definition of the confidence intervals $\beta_i(t-1)$, and a union bound over all $T_i(t) \in \{0, \dots, t\},\ t = 1, \dots, \infty$. □

Recall that

$$\mathcal{U}'_t = \left\{U : \forall V \in \mathcal{C}, \widehat{\Delta}^+_{U,V}(t) > -\overline{d}_{U,V} \max_{W \in \mathcal{C}} \widehat{G}^+_{W,U}(t)/2\right\}$$

and

$$\widehat{G}^+(t) = \max_{U \in \mathcal{U}'_t, V \in \mathcal{C}} \widehat{G}^+_{V,U}(t) \text{ and } (U_t, V_t) = \operatorname*{arg\,max}_{U \in \mathcal{U}'_t, V \in \mathcal{C}, U \neq V} \widehat{G}^+_{V,U}(t).$$

The following lemma gives a lower bound on $\widehat{G}^+(t)$.

Lemma 8. *On the event ξ, for all time steps t, $\widehat{G}^+(t) \geq \frac{1}{2} G_{I(t)}$.*

Proof. First note that on event ξ, for any t and any pair of decisions $U, V \in \mathcal{C}$, we have $\widehat{\Delta}^+_{U,V}(t) \geq \Delta_{U,V}$ and consequently $\widehat{G}^+_{U,V} \geq G_{U,V}$. The proof proceeds by distinguishing two main cases. We show the details for the case when $I(t) \notin U^*$. The case $I(t) \in U^*$ can be dealt with using similar arguments.

Case 1) $I(t) \in U_t$
We introduce $W_t = \mathsf{C}_{U_t}$, which exists since $U_t \neq U^*$, as $I(t) \notin U^*$ and $I(t) \in U_t$. We have

$$\begin{aligned}\widehat{G}^+(t) = \max_{U \in \mathcal{U}'_t, V \in \mathcal{C}} \widehat{G}^+_{V,U}(t) \geq \widehat{G}^+_{W_t,U_t}(t) &= \frac{\widehat{\Delta}^+_{W_t,U_t}(t)}{\overline{d}_{U_t,W_t}} \\ \geq \frac{\Delta_{W_t,U_t}}{\overline{d}_{U_t,W_t}} &\overset{(a)}{\geq} \min_{U : I(t) \in U} \frac{\Delta_{\mathsf{C}_U,U}}{\overline{d}_{U,\mathsf{C}_U}} \overset{(b)}{=} G_{I(t)},\end{aligned}$$

where **(a)** follows from the fact that $W_t = \mathsf{C}_{U_t}$ and $I(t) \in U_t$, and **(b)** is due to $I(t) \notin U^*$ so that its complexity is defined as the minimum over decisions U to which it belongs.

Case 2) $I(t) \in V_t$
Let $W_t = \mathsf{C}_{V_t}$, noting that W_t is well-defined since $V_t \neq U^*$, as $I(t) \notin U^*$ and $I(t) \in V_t$.
Case 2.1) $V_t \in \mathcal{U}'_t$: Similar to Case 1, we have

$$\begin{aligned}\widehat{G}^+(t) = \max_{U \in \mathcal{U}'_t, V \in \mathcal{C}} \widehat{G}^+_{V,U}(t) &\overset{(a)}{\geq} \widehat{G}^+_{W_t,V_t}(t) = \frac{\widehat{\Delta}^+_{W_t,V_t}(t)}{\overline{d}_{V_t,W_t}} \\ \geq \frac{\Delta_{W_t,V_t}}{\overline{d}_{V_t,W_t}} &\overset{(b)}{\geq} \min_{U : I(t) \in U} \frac{\Delta_{\mathsf{C}_U,U}}{\overline{d}_{U,\mathsf{C}_U}} \overset{(c)}{=} G_{I(t)},\end{aligned}$$

where **(a)** holds by $V_t \in \mathcal{U}'_t$ and the definition of $\widehat{G}^+(t)$, **(b)** follows from $W_t = \mathsf{C}_{V_t}$ and our assumption $I(t) \in V_t$, and **(c)** is due to $I(t) \notin U^*$ so that its complexity is defined as the minimum over decisions U to which it belongs.

Case 2.2) $V_t \notin \mathcal{U}'_t$: In this case, by definition of $\mathcal{U}'(t)$ there exists a decision set Z_t such that $\widehat{G}^+_{V_t,Z_t}(t) \leq -\frac{1}{2}\max_{W\in\mathcal{C}} \widehat{G}^+_{W,V_t}(t)$. Therefore, we have

$$\widehat{G}^+_{V_t,Z_t}(t) \leq -\frac{1}{2}\max_{W\in\mathcal{C}} \widehat{G}^+_{W,V_t}(t) \leq -\frac{1}{2}\widehat{G}^+_{W_t,V_t}(t) \tag{37}$$

$$\leq -\frac{1}{2}G_{W_t,V_t} \leq -\frac{1}{2}G_{I(t)}, \tag{38}$$

where the last equality follows from the fact that $I(t) \notin U^*$ and the definition of $G_{I(t)}$ in that case. We now focus on the three decision sets V_t, U_t, Z_t and define the value $\dot{\mu}_i$ associated to arms $i \in V_t \cup U_t \cup Z_t$ as

$$\dot{\mu}_i = \begin{cases} \widehat{\mu}^-_i(t) & \text{if } i \in U_t, \\ \widehat{\mu}^+_i(t) & \text{if } i \in (V_t \cup Z_t)\backslash U_t. \end{cases}$$

We also define $\dot{G}_{Z_t,U_t}$, $\dot{G}_{V_t,U_t}$, $\dot{G}_{Z_t,V_t}$ as well as $\dot{\Delta}_{Z_t,U_t}$, $\dot{\Delta}_{V_t,U_t}$, $\dot{\Delta}_{Z_t,V_t}$ obtained by using $\dot{\mu}_i$ instead of μ_i in their computation. Then we get

$$\begin{aligned}
\dot{\Delta}_{Z_t,V_t}(t) &= \sum_{i\in Z_t\backslash V_t} \dot{\mu}_i - \sum_{i\in V_t\backslash Z_t} \dot{\mu}_i = \sum_{i\in (Z_t\backslash V_t)\backslash U_t} \dot{\mu}_i + \sum_{i\in (Z_t\backslash V_t)\cap U_t} \dot{\mu}_i - \sum_{i\in (V_t\backslash Z_t)\backslash U_t} \dot{\mu}_i - \sum_{i\in (V_t\backslash Z_t)\cap U_t} \dot{\mu}_i \\
&= \sum_{i\in (Z_t\backslash V_t)\backslash U_t} \widehat{\mu}^+_i(t) + \sum_{i\in (Z_t\backslash V_t)\cap U_t} \widehat{\mu}^-_i(t) - \sum_{i\in (V_t\backslash Z_t)\backslash U_t} \widehat{\mu}^+_i(t) - \sum_{i\in (V_t\backslash Z_t)\cap U_t} \widehat{\mu}^-_i(t) \\
&\geq \sum_{i\in (Z_t\backslash V_t)\backslash U_t} \widehat{\mu}^-_i(t) + \sum_{i\in (Z_t\backslash V_t)\cap U_t} \widehat{\mu}^-_i(t) - \sum_{i\in (V_t\backslash Z_t)\backslash U_t} \widehat{\mu}^+_i(t) - \sum_{i\in (V_t\backslash Z_t)\cap U_t} \widehat{\mu}^+_i(t) \\
&= -\left(\sum_{i\in (V_t\backslash Z_t)\backslash U_t} \widehat{\mu}^+_i(t) + \sum_{i\in (V_t\backslash Z_t)\cap U_t} \widehat{\mu}^+_i(t) - \sum_{i\in (Z_t\backslash V_t)\backslash U_t} \widehat{\mu}^-_i(t) - \sum_{i\in (Z_t\backslash V_t)\cap U_t} \widehat{\mu}^-_i(t) \right) \\
&= -\Delta^+_{V_t,Z_t}(t).
\end{aligned} \tag{39}$$

Furthermore,

$$\widehat{G}^+_{Z_t,U_t}(t) \overset{(a)}{=} \dot{G}_{Z_t,U_t} \overset{(b)}{\geq} \min\left(\dot{G}_{Z_t,V_t}, \dot{G}_{V_t,U_t}\right) \tag{40}$$

$$\overset{(c)}{\geq} \min\left(\dot{G}_{Z_t,V_t}, \widehat{G}^+_{V_t,U_t}(t)\right) \tag{41}$$

$$\overset{(d)}{\geq} \min\left(-\widehat{G}^+_{V_t,Z_t}(t), \widehat{G}^+_{V_t,U_t}(t)\right) \tag{42}$$

$$\overset{(e)}{\geq} \min\left(\frac{1}{2}G_{I(t)}, \widehat{G}^+_{V_t,U_t}(t)\right),$$

where **(a)** and **(c)** are obtained by the definition $\dot{G}$, **(b)** follows from $\dot{\mu}_{U_t} < \dot{\mu}_{V_t} < \dot{\mu}_{Z_t}$ and an analogue of Proposition 3 with strict inequality in case of $\dot{G}_{V_t,U_t} \neq \dot{G}_{Z_t,V_t}$, **(d)** uses equation (39), and **(e)** is by equation (37).

Now let us assume $\widehat{G}^+(t) = \widehat{G}^+_{V_t,U_t}(t) < \frac{1}{2}G_{I(t)}$, from which we will derive a contradiction. From equation (37) and by definition of $\dot{G}$, we have $\widehat{G}^+(t) = \widehat{G}^+_{V_t,U_t}(t) < \frac{1}{2}G_{I(t)} \leq \dot{G}_{Z_t,V_t}$, so we have strict inequality in **(b)** of equation (40). Consequently, we can derive from (40) the contradiction

$$\widehat{G}^+_{Z_t,U_t}(t) > \min\left(\widehat{G}^+_{V_t,U_t}(t), \frac{1}{2}G_{I(t)}\right) \geq \widehat{G}^+_{V_t,U_t}(t) = \max_{V\in\mathcal{C}} \widehat{G}^+_{V,U_t}(t),$$

which completes the proof of $\widehat{G}^+(t) \geq \frac{1}{2}G_{I(t)}$. □

The following Lemma shows that any set U in $\mathcal{U}_t$ also is in $\mathcal{U}'_t$.

Lemma 9. *On the event ξ, for all time steps t, $\mathcal{U}_t \subset \mathcal{U}'_t$.*

Proof. It is obviously sufficient to show that the threshold $-\mathcal{T}_{U,V}(t) \leq 0$. On the event ξ, we have $\max_{W \in \mathcal{C}} \widehat{G}^+_{W,U}(t) \geq \widehat{G}^+_{U^*,U}(t) \geq G_{U^*,U} \geq 0$. By definition of $\mathcal{T}_{U,V}(t)$, this implies that $-\mathcal{T}_{U,V}(t) \leq 0$. □

The following Lemma gives an upper bound on $\widehat{G}^+(t)$.

Lemma 10. *Assume that event ξ holds. Then $8\beta_{I(t)}(t-1) \geq \widehat{G}^+(t)$ for all t.*

Proof. First note that

$$\widehat{G}^+_{U_t,V_t}(t) \overset{(a)}{>} -\frac{1}{2}\max_{W \in \mathcal{C}} \widehat{G}^+_{W,U_t}(t) \overset{(b)}{=} -\frac{1}{2}\widehat{G}^+_{V_t,U_t}(t), \tag{43}$$

where **(a)** follows from $U_t \in \mathcal{U}'_t$ and the definition of $\mathcal{U}'_t$ and **(b)** from the definition of U_t and V_t. Moreover, we have

$$\begin{aligned}\widehat{G}^+_{V_t,U_t}(t) &\overset{(a)}{=} \frac{2}{\overline{d}_{U_t,V_t}} \sum_{i \in U_t \oplus V_t} \beta_i(t-1) - \widehat{G}^+_{U_t,V_t}(t) \\ &\overset{(b)}{\leq} \frac{2}{\overline{d}_{U_t,V_t}} \sum_{i \in U_t \oplus V_t} \beta_i(t-1) + \frac{1}{2}\widehat{G}^+_{V_t,U_t}(t),\end{aligned}$$

where **(a)** is because $\widehat{\Delta}^+_{V_t,U_t}(t) + \Delta^+_{U_t,V_t}(t) = 2\sum_{i \in U_t \oplus V_t} \beta_i(t-1)$, and **(b)** is because of equation (43). Hence, we obtain

$$\widehat{G}^+(t) = \widehat{G}^+_{V_t,U_t}(t) \leq \frac{4}{\overline{d}_{U_t,V_t}} \sum_{i \in U_t \oplus V_t} \beta_i(t-1). \tag{44}$$

Moreover, as we will demonstrate in the following, we have

$$\frac{1}{\overline{d}_{U_t,V_t}} \sum_{i \in U_t \oplus V_t} \beta_i(t-1) \leq 2\beta_{I(t)}(t-1). \tag{45}$$

We show this in detail for the case when $I(t) \in U_t$, the case $I(t) \in V_t$ is similar. For $I(t) \in U_t$, we have $\sum_{i \in U_t \setminus V_t} \beta_i(t-1) \geq \sum_{i \in V_t \setminus U_t} \beta_i(t-1)$ and consequently,

$$\sum_{i \in U_t \oplus V_t} \beta_i(t-1) = \sum_{i \in U_t \setminus V_t} \beta_i(t-1) + \sum_{i \in V_t \setminus U_t} \beta_i(t-1) \leq 2\sum_{i \in U_t \setminus V_t} \beta_i(t-1).$$

Since for $I(t) \in U_t$, we have for all $i \in U_t \setminus V_t$ that $\beta_i(t-1) \leq \beta_{I(t)}(t-1)$, we therefore obtain

$$\sum_{i \in U_t \oplus V_t} \beta_i(t-1) \leq 2\sum_{i \in U_t \setminus V_t} \beta_i(t-1) \leq 2d_{U_t,V_t}\beta_{I(t)}(t-1)$$

and consequently,

$$\frac{1}{\overline{d}_{U_t,V_t}} \sum_{i \in U_t \oplus V_t} \beta_i(t-1) \leq 2\,\frac{d_{U_t,V_t}}{\overline{d}_{U_t,V_t}}\beta_{I(t)}(t-1) \leq 2\beta_{I(t)}(t-1),$$

which proves equation (45). Finally, combining (44) and (45) gives the claim of the lemma. □

Finally, we are ready to give the proof of Theorem 2.

Proof of Theorem 2. First note that by Lemma 9, on event ξ we have $\mathcal{U}_t \subseteq \mathcal{U}'_t$, so the algorithm is well-defined, since as long as $|\mathcal{U}_t| > 1$ also $|\mathcal{U}'_t| > 1$ and there is always an arm to be pulled.

Next, we show that with probability of at least $1 - \delta$, the algorithm returns the optimal set U^*. Indeed, assume that ξ holds and that U^* is rejected from $\mathcal{U}_t$ at some step t. Then there exists a set V such that $0 \geq \widehat{\Delta}^+_{U^*,V}(t) = \mu^+_{U^*}(t) - \mu^-_V(t) \geq \mu_{U^*} - \mu_V > 0$, which is a contradiction. Hence, the claim follows from Lemma 7.

Finally, let us consider the sample complexity of our algorithm. By Lemma 10 and Lemma 8 we have for any pulled arm i that $8\beta_i(t) \geq \widehat{G}^+(t) \geq \frac{1}{2}G_i$. Summing over all arms $i \in \mathcal{K}$, this gives for each t that $T_i(t) \leq 128H_i \log(4Kt^2/\delta)$. Therefore,

$$\sum_{i \in \mathcal{K}} T_i(t) = t \leq \sum_{i \in \mathcal{K}} 128H_i \log(4Kt^2/\delta) \leq 128H \log(4Kt^2/\delta).$$

Thus, as soon as t reaches $t \geq 128H \log(4Kt^2/\delta)$, the algorithm stops. Denoting this step by $\tilde{n}$ and using Lemma 8 of Antos et al. [2010] in order to solve this equation gives $\tilde{n} \leq O\left(H \log(HK/\delta)\right)$. □