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

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

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.,

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

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*

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

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]

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

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*

$$H = \sum_{i\in\mathcal{K}} H_i. \tag{5}$$

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

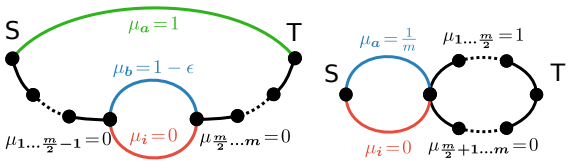
$$\Delta_i^{\odot} = \begin{cases} \mu^* - \max_{U \in \mathcal{C}: i \in U} \mu_U & \text{if } i \notin U^*, \\ \mu^* - \max_{U \in \mathcal{C}: i \notin U} \mu_U & \text{if } i \in U^*. \end{cases}$$

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.

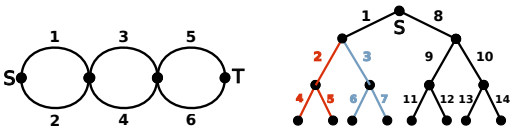
m, since the largest exchange set needs to remove all the (black) arms in the top (bottom) path and add all the (black) arms in the bottom (top) path (e.g., consider the exchange sets needed to move from U^* to $\{\boldsymbol{a}, m/2+1, \dots, m\}$). Hence, $H_i^{\odot} = \frac{\text{width}(\mathcal{C})^2}{(\Delta^{\odot}_{U^*,V})^2} = \frac{m^2}{(1/m)^2} = m^4$, 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

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

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 $\min_{\mathbb{E}[T_i(t)],\mathcal{B}} \quad \sum_{i\in\mathcal{K}} \mathbb{E}\big[T_i(t)\big]\log(1/4\delta)$, 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

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

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.