Improved Learning Complexity in Combinatorial Pure Exploration Bandits

Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh, Ronald Ortner & Peter Bartlett

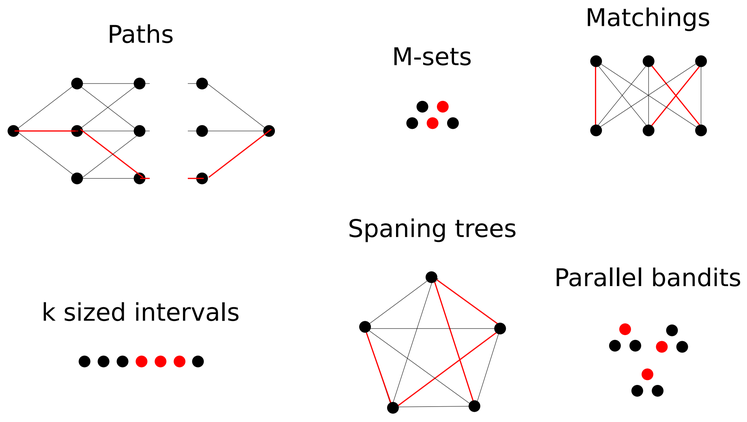
1) Abstract

- We construct a new measure of **complexity** that provably characterizes the learning performance of the algorithms we propose for the fixed confidence and fixed budget settings.
- We show that this complexity is never bigger than the one in the 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 computation, we provide a series of examples, including a planning problem, where the extra computation is not significant.

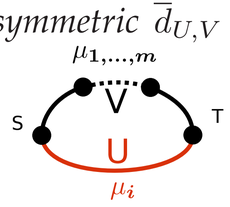
2) Problem Formulation

- $\mathcal{K}$ a set of K arms with initially unknown distributions $\nu_i \in [0,1]$ and an expected value μ_i that the learner aims to learn.
- The (combinatorial) decision space $\mathcal{C} \subseteq 2^K$ contains decision sets (sets of arms) $U \subseteq \mathcal{K}$

$$\mu_U = \sum_{i \in U} \mu_i .$$



- Gap: $\Delta_{U,V} = \mu_U - \mu_V$, and $U^* = \arg\max_{U \in \mathcal{C}} \mu_U$
- $U \oplus V = (U \backslash V) \cup (V \backslash U)$ the set of arms either in U or in V, but not in both.
- *(A)symmetric* $\overline{d}_{U,V} = |U \oplus V|$ $(d_{U,V} = |U \backslash V|)$



$\overline{d}_{U,V} = m + 1$ and $d_{V,U} = m$ and $d_{U,V} = 1$

The fixed budget setting: Given a budget n, the learner minimizes the probability $\widetilde{\delta}$ of not identifying the best decision set, i.e., $\widetilde{\delta} = \mathbb{P}\left[\widehat{U}^*(n) \neq U^*\right]$. See paper for the fixed confidence setting.

3) Sample Complexities

• Discriminate two decision sets U and V

Simply uniformly allocate pulls over the arms in $U \oplus V$ and stop at the first step t when the lower-bound on the gap is positive, i.e.,

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

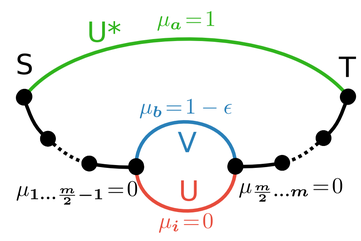
This stopping time suggests the idea to define the **sample complexity** for U, V as

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

- inverse dependency on $\Delta_{U,V}$
- $\overline{d}_{U,V}$ is like a variance term

• C_U: the complement of U

When trying to discard a suboptimal set U from $\mathcal{C}$ it may be easier to compare it to a set $V \neq U^*, \mu_V > \mu_U$ with a smaller complexity $H_{U,V}$.



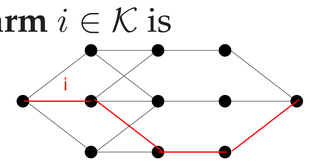
For any decision $U \neq U^*$, its complement is

$$\mathsf{C}_U = \arg\min_{V \in \mathcal{C}: \mu_V > \mu_U} H_{U,V}.$$

For any arm $i \in \mathcal{K}$ different sets U may force different requirements on the number of times i should be pulled.

Thus the **complexity of an arm** $i \in \mathcal{K}$ is

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



The **global complexity** is $H = \sum_{i \in \mathcal{K}} H_i$.

The *simplicity* is $G_{U,V} = \frac{\Delta_{U,V}}{\overline{d}_{U,V}}$.

4) Fixed Budget Algorithm

The algorithm is an extension of Audibert et. al. [1]. It is composed of K phases. At the end of each phase, we compute the empirical $\widehat{\mu}_i(k)$, $\widehat{\Delta}_{U,V}(k)$, $\widehat{U}^*(k)$, $\widehat{G}_{U,V}(k) = \widehat{\Delta}_{U,V}(k)/\overline{d}_{U,V}$, $\widehat{\mathsf{C}}_U(k) = \arg\max_{V \in \mathcal{C}: \widehat{\mu}_V(k) > \widehat{\mu}_U(k)} \widehat{G}_{U,V}(k)$, $\widehat{G}_i(k) = \min_{U \in \mathcal{C}: i \in U \oplus \widehat{\mathsf{C}}_U(k)} \widehat{G}_{\widehat{\mathsf{C}}_U(k),U}(k)$. We set $n_k = \left\lceil \frac{n-K}{\overline{\log}(K)(K+1-k)} \right\rceil$, $k \in \mathcal{K}$.

Parameters: number of rounds n, set of arms $\mathcal{K}$, decision set $\mathcal{C}$ and cumulative pulls schedule $n_0, n_1, \dots, n_K$.
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 arms j_k, i.e., set $\mathcal{K}_{k+1} = \mathcal{K}_k \backslash j_k$.
  $k \longleftarrow k + 1$
end while
Return $J_n = \arg\max_{U \in \mathcal{C}} \widehat{\mu}_U(n)$

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

with $\overline{H} = \max_{i \in \mathcal{K}} i H_{(i)}$, $H_{(1)} \geq H_{(2)} \geq \dots \geq H_{(K)}$. As noted in [1], it holds that $\overline{H} \leq H \leq \overline{H}\overline{\log}(K)$.

Sketch of the proof. By **double** induction through the phases of the algorithm:

(i) The arms that are rejected are well classified as belonging or not to U^*

(ii) The deactivated arm has been sufficiently sampled so that if will be used later as a "good" reference arm (as part of some C_U).

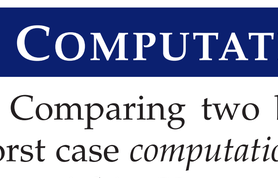
5) Comparison with Chen et al.

Chen et al. (2014) [2] proposed a different measure of sample complexity of arm $i \in \mathcal{K}$:

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

The complexity is $H_i^{\odot} = \frac{width(\mathcal{C})^2}{(\Delta_i^{\odot})^2}$.

- Our complexity uses the **complement** C_U instead of U^*.
- $width(\mathcal{C})$ defines a **distance** $\overline{d}$ directly for $\mathcal{C}$ while we define it for every pair U, V:



$width(\mathcal{C}) \approx m$

$H_i^{\odot} = \frac{\text{width}(\mathcal{C})^2}{(\Delta^{\odot}_{U^*,V})^2} = \frac{m^2}{1/m^2} = m^4$

$H_i = \frac{\overline{d}^2_{U^*,V}}{\Delta^2 U^*,V} = \frac{2^2}{(1/m)^2} = 4m^2 \leq H_i^{\odot}$

For all $i \in \mathcal{K}$, $H_i^{\odot} \geq H_i$

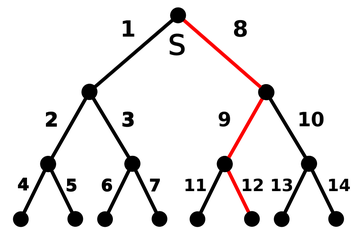
Sketch of Proof. In the first part we show that the definition H_i is indeed better when comparing sets U to their complements C_U instead of U^*. Consider the drawing on the left. There are three decision sets U, V and U^* containing the three edges/arms i, b and a. $V = \mathsf{C}_U$, $U^* = \mathsf{C}_V$ and $b \in V$. As $b \in V \oplus \mathsf{C}_V$, by definition the complexity of b should be at least H_{V,C_V}. As $b \in U \oplus \mathsf{C}_U$, by definition the complexity of b should be at least H_{U,C_U}. This last requirement is an extra requirement compared to the complexity defined by Chen et al. Fortunately, it is shown that $H_{U,\mathsf{C}_U} \leq H_{V,\mathsf{C}_V}$.

6) Computational Complexity

Comparing two by two every solution has a worst case *computational complexity* that can be exponential in K.

Examples of tractability:

Using *independence* in the multi-bandit.



Tree-planning:

Tree-planning case: $K \approx |\mathcal{C}|$

- as maximizing the expected sum of rewards over m consecutive actions from a starting state when the state dynamics is deterministic and the reward distributions are unknown.
- the set of decisions (paths) $V \neq U$ can be clustered into m groups of sets depending on which is 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.
- We can also show a planning tree example where the learning complexity is significantly improved over [2].

7) Discussion

- Is our sample complexity optimal? Solution of an optimization problem?
- Asymmetric distance $d_{U,V}$ instead of $\overline{d}_{U,V}$? σ_i^2/Δ_i^2 instead of the standard $(\sigma_i^2 + \sigma_{i^*}^2)/\Delta_i^2$ in the single bandit fixed budget setting?

[1] J.-Y. Audibert, S. Bubeck, and R. Munos. Best Arm Identification in Multi-Armed Bandits. In *COLT*, 2010.

[2] S. Chen, T. Lin, I. King, M. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. In *NIPS*, 2014.