

Efficient Matroid Bandit Linear Optimization Leveraging Unimodality

Aurélien Delage¹, Romaric Gaudel¹

¹ Univ. Rennes, Inria, CNRS IRISA - UMR 6074, F35000 Rennes, France
prenom.nom@irisa.fr

December 2, 2025

Abstract

We study the combinatorial semi-bandit problem under matroid constraints. The regret achieved by recent approaches is optimal, in the sense that it matches the lower bound. Yet, time complexity remains an issue for large matroids or for matroids with costly membership oracles (*e.g.* online recommendation that ensures diversity). This paper sheds a new light on the matroid semi-bandit problem by exploiting its underlying unimodal structure. We demonstrate that, with negligible loss in regret, the number of iterations involving the membership oracle can be limited to $\mathcal{O}(\log \log T)$. This results in an overall improved time complexity of the learning process. Experiments conducted on various matroid benchmarks show (i) no loss in regret compared to state-of-the-art approaches; and (ii) reduced time complexity and number of calls to the membership oracle.

1 Introduction

The semi-bandit framework models the interaction of an online learner that repeatedly selects one or multiple items to play from a known, and usually finite, groundset E . Stochastic feedback in terms of rewards is received for each selected item at each round. The goal is to maximize the sum of expected rewards, which is equivalent to minimizing the sum of regret for not selecting the optimal subset at each round. Matroid bandit problems constrain the subsets of E

that can be played. Applications range from learning routing networks [Kveton et al., 2014] and advertisement [Streeter et al., 2009], to task assignment [Chen et al., 2016], learning maximum spanning trees [Papadimitriou and Steiglitz, 1998], and leader-follower multi-agent planning [Clark et al., 2012, Lin et al., 2011].

The bandit algorithms designed for combinatorial problems with linear objective functions, such as CUCB [Chen et al., 2013] and ESCB [Combes et al., 2015] apply to matroid bandit problems. Their versions leveraging the strong matroid structure, respectively **OMM** [Kveton et al., 2014] and **KL-OSM** [Talebi and Proutiere, 2016], achieve better asymptotic regret, and **KL-OSM** is even optimal. However, time complexity still remains an issue, particularly for matroids with costly membership oracles [Tzeng et al., 2024]. The main bottleneck is the call to the greedy algorithm at each iteration. The latter suffers a $\mathcal{O}(|E|(\log |E| + \mathcal{T}_m))$ time complexity, where \mathcal{T}_m is the time complexity of the membership oracle determining whether a set $I \cup \{x\}$ belongs to the matroid, given that $I \subset E$ does. Recently, Tzeng et al. [2024] proposed **FasterCUCB**, a first attempt to reduce the overall time complexity. The authors approximate the greedy procedure using a dynamic maintenance of maximum-weight bases. Yet, this comes at the cost of a loss in regret; the introduction of domain-dependent update oracles; and an additional $\text{polylog}(T)$ term in the per-round time complexity.

Contributions We present MAUB, a unimodal bandit algorithm tailored for matroid optimization. Our main contributions are:

1. with negligible loss in regret, MAUB requires only $\mathcal{O}(\log \log(T))$ calls to the matroid structure, leading to a reduced overall time complexity compared to OMM, KL-OSM and FasterCUCB;
2. MAUB does not require instance-specific matroid oracles, which makes it more general; and
3. our alternative approach to matroid optimization highlights the interest of unimodal structures for both the analysis of combinatorial bandit problems, and practical implementation of simple, yet efficient, learning algorithms.

The article is organized as follows: Section 2 presents the related work and Section 3 gives the necessary background. Next, we introduce our proposed algorithm, namely MAUB, in Section 4. Its theoretical guarantees are studied in Section 5, and Section 6 illustrates its empirical behavior in numerical experiments.

2 Related Work

Unimodal Bandit Unimodality considers graph-structured bandit problems. It is assumed that unless the current arm is optimal, there always exists a strictly better one in its neighborhood. It defines an important subclass of the general bandit problem [Cope, 2009, Yu and Mannor, 2011, Combes and Proutiere, 2014, Trinh et al., 2020]. To handle this class of problems, Combes and Proutiere [2014]’s algorithm, namely OSUB, splits the optimization problem at each round into two parts: (i) determining the best arm so far; and (ii) deciding what to play to ensure optimism. The first problem is solved by examining *mean-statistics*, and the second one only looks at arms within the neighborhood of the current best arm. Trinh et al. [2020], Gauthier et al. [2021, 2022] extended OSUB to the semi-bandit setting to handle their target application, multiple-play online recommendation systems (ORS).

The restricted exploration set has been leveraged by Trinh et al. [2020], Gauthier et al. [2021, 2022] to design regret-optimal algorithms, by aligning neighborhood structures with the gap-dependent terms appearing in the logarithmic regret bound. While they did it to simplify algorithm design and regret analysis, in current paper we demonstrate that unimodality also path the way toward the reduction of computational costs. This reduction relies on two key factors: first, the neighborhood rarely changes, which minimizes the need for combinatorial optimization; and second, most of the time, decisions are restricted to a small neighborhood, keeping the per-step computation low.

Optimal Regret in Matroid Linear Optimization Table 1 summarizes regret upper-bounds and time complexities of state-of-the-art matroid bandit algorithms. Both optimistic matroid maximization (OMM) [Kveton et al., 2014] and KL-based efficient sampling for matroids (KL-OSM) [Talebi and Proutiere, 2016] rely on repeated calls to the greedy algorithm that computes the arm B^* with the highest expected value according to *optimistic statistics*. OMM corresponds to the direct application of CUCB [Chen et al., 2013] to matroids. It uses UCB-like bonuses, and achieves a $\mathcal{O}\left(\frac{|E|-D}{\Delta_{\min}} \log(T)\right)$ regret, where D is the rank of the matroid, and Δ_{\min} is the minimum gap between the mean value of an element in the optimal arm and any element outside it. KL-OSM addresses an extraneous multiplicative factor in this regret by replacing UCB bonuses with KL indices, yielding an optimal algorithm that matches the lower bound of Kveton et al. [2014]. However, reducing regret comes at a price: it limits applicability to reward distributions on $[0, 1]$ and increases computational cost.

Approximation to Reduce Time Complexity Having these regret-optimal algorithms, the remaining question is computation complexity. Tzeng et al. [2024] introduced FasterCUCB, an algorithm focusing on this aspect. The approach maintains a maximum-weight basis, which is updated across iterations. Calls to the membership oracle are thus replaced with calls to an update oracle, for which class-dependent

Table 1: Overall regret and time complexity achieved by OMM, KL-OSM, FasterCUCB and MAUB for the matroid bandit problem. \mathcal{T}_u is the time complexity of updating maximum-weight bases in FasterCUCB. Tzeng et al. [2024, Section 3] detail \mathcal{T}_u for uniform, graphic, partition, and transversal matroids. The mapping σ is defined in Section 4.1, and poly is a polynomial.

	Regret	Time Complexity (in $\mathcal{O}()$)
CUCB/OMM 1	$\sum_{e \notin B^*} \frac{16 \log T}{\min_{i \in B^*, \mu_i > \mu_e} \mu_i - \mu_e}$	$ E (\log E + \mathcal{T}_m)T$
KL-OSM 9	$\sum_{e \notin B^*} \frac{(\mu_e - \mu_{\sigma(e)}) \log T}{\text{KL}(\mu_e, \mu_{\sigma(e)})}$	$ E (\log E + \mathcal{T}_m)T$
FasterCUCB 10	$\sum_{e \notin B^*} \frac{12 \max_{i \in B^*, \mu_i > \mu_e} \text{supp}(\mu_i) \cdot \log T}{8 \cdot \log T}$	$D\mathcal{T}_u T \text{polylog } T$
MAUB (Th.5.5)	$\sum_{e \notin B^*} \frac{\mu_{\sigma(e)} - \mu_e}{\mu_{\sigma(e)} - \mu_e}$	$ E T + \binom{ E }{D} \text{poly}(E , D)\mathcal{T}_m \log \log T$

efficient implementations exist. The cost, however, is the introduction of a multiplicative constant in the overall regret, and the per-round time complexity suffers an additional multiplicative $\text{polylog}(T)$ factor. Furthermore, while the-round time complexity of FasterCUCB is sublinear in $|E|$ for various classes of matroids, it also contains a $\text{polylog}(T)$ factor, which makes it slower than MAUB on the long run (as soon as $D\mathcal{T}_u \text{polylog}(T) > |E|$).

For the sake of completeness, Perrault et al. [2019] study the matroid semi-bandit problem when there is a submodular function to optimize. The rationale is that such function is either the primary objective, or arises from the design of the optimistic term (*e.g.* ESCB [Degenne and Perchet, 2016]). Solving the corresponding combinatorial problem at each iteration would be too prohibitive, so they (approximately) solve it using either local search or a variant of the greedy algorithm. Yet, the time complexity is at least $\mathcal{O}(D|E|)$, and each submodular maximization does not decrease the number of calls to the membership oracle compared to OMM. In current paper we limit ourselves to linear objective for which such fancy optimism is not required to get regret-optimal algorithms.

3 Background

The following provides background and notations on general combinatorial bandit problems, the matroid structure, and the corresponding greedy algorithm.

We use bold symbols to denote collections (*e.g.* $\boldsymbol{\mu}$ stands for $(\mu_e)_{e \in E}$). Also, for any set I and element e , we write $I + e$ (resp. $I - e$) for $I \cup \{e\}$ (resp. $I \setminus \{e\}$).

Combinatorial Semi-Bandit Setting We consider a combinatorial semi-bandit problem in which arms are subsets of a groundset E of elementary actions. For any e in E , $X_e(t)$ is the random variable associated with the reward of e at time step t . All random variables $(X_e(t))_{t \geq 1}$ are assumed i.i.d., with unknown mean $\mu_e \in (0, \infty)$. We further assume that $\forall x, y \in E, \mu_x \neq \mu_y$. For any subset $I \subseteq E$, we define at time step t $X_I(t) \triangleq \sum_{e \in I} X_e(t)$, and the corresponding expectation $\mu_I \triangleq \sum_{e \in I} \mu_e$. Random variables $(X_I(t))_{I \subseteq E}$ are consequently correlated.

Semi-bandit interactions are done as follows: there is a set of allowed combinations $\mathcal{I} \subseteq 2^E$ such that at each time step t , the player plays a subset $P(t) \in \mathcal{I}$, observes $(X_e(t))_{e \in P(t)}$ and is rewarded $X_{P(t)}(t)$. In this scenario, learning algorithms aim at minimizing the (*pseudo-*)regret

$$R(t) \triangleq \sum_{s \leq t} \max_{I \in \mathcal{I}} \mu_I - \mu_{P(s)}. \quad (1)$$

Matroids Matroids are constraining the set \mathcal{I} .

Definition 3.1 (Matroid, groundset, independent subsets, rank, bases). A matroid \mathcal{M} is a pair (E, \mathcal{I}) , where E is a set of items, called the groundset, and $\mathcal{I} \subseteq 2^E$ is a subset of the powerset of E . Subsets I in \mathcal{I} are said independent.

The matroid structure requires that (i) $\emptyset \in \mathcal{I}$, (ii) hereditary property: every subset of an independent set is also independent, and (iii) augmentation property: for all I, J in $\mathcal{I}, |I| = |J| + 1$ implies that there exists $e \in I \setminus J$ such that $J + e \in \mathcal{I}$.

Axioms (ii) and (iii) imply that all maximal sets (for the inclusion order) have equal size, which defines the rank D of \mathcal{M} . These sets are the bases $B \in \mathcal{B}$ of \mathcal{M} .

From these axioms, it follows the *basis exchange property*, which underpins the unimodal graph explored by MAUB to identify an optimal independent set.

Lemma 3.2 (basis exchange property [Schrijver, 2003, Th. 10.2 p.177], [Talebi and Proutiere, 2016, Prop. 1]). *For any matroid $M = (E, \mathcal{I})$, it holds that:*

$$\forall X, Y \in \mathcal{B} \implies \forall x \in X \setminus Y, \exists y \in Y \setminus X, X - x + y \in \mathcal{B}.$$

Linear Optimization in Matroids All means μ_e for e in E being non-negative, and since \mathcal{B} is finite, an independent set B^* with highest value must be a base (namely $B^* \in \mathcal{B}$). In the following, we thus restrict the arms of the bandit problem to bases $B \in \mathcal{B}$. We further assume that B^* is uniquely defined.

If values $\boldsymbol{\mu}$ were known, one could compute the optimal base B^* through the *greedy* algorithm (Algorithm 1). In essence, greedy iteratively constructs an optimal basis B^* by considering elements $e_i \in E$ in decreasing order. At a step i , e_i is added to B^* if and only if it does not make B^* dependent (*i.e.* $B^* + e_i \in \mathcal{I}$).

Algorithm 1 Greedy Algorithm [Edmonds, 1971]

Require: element values $\boldsymbol{\mu}$

```

 $B^* \leftarrow \emptyset$ 
Find an ordering  $\mu_{e_1} \geq \dots \geq \mu_{e_{|E|}}$ 
for  $i = 1, \dots, K$  do
    if  $B^* + e_i \in \mathcal{I}$  then  $B^* \leftarrow B^* + e_i$  end if
end for
Return  $B^*$ 

```

Each independence property is tested by querying a membership oracle with the subset $B^* + e_i$. Hence the $\mathcal{O}(T|E|(\log |E| + \mathcal{T}_m))$ time complexity of OMM and KL-OSM, which are calling greedy at each time step (on optimistic estimates of $\boldsymbol{\mu}$).

4 Unimodal Approach to Learn in Matroids

Unimodal bandits enable decoupling the optimistic selection rule from the combinatorial optimization over the matroid. Concretely, optimization is performed on empirical means and only when the leader changes. Since leader changes are rare, this results in very few membership-oracle queries and thus low

computational cost. The arm to play is then chosen optimistically within a small candidate set of neighbors.

The following formally describes the underlying unimodal structure in matroid bandit problems as well as our unimodal bandit algorithm that leverages it.

4.1 Unimodality

Let us first define the graph associated to a matroid and then prove that it is unimodal. Let $M = (E, \mathcal{I})$ be a matroid of bases \mathcal{B} , whose elements are associated to expectations $\boldsymbol{\mu}$. Let $B \in \mathcal{B}$ and define the mapping $\sigma_{B, \boldsymbol{\mu}} : E \setminus B \rightarrow B$ such that $\forall e \in E \setminus B, \sigma_{B, \boldsymbol{\mu}}(e) = \arg \min_{\{x \in B : B - x + e \in \mathcal{B}\}} \mu_x$. $\sigma_{B, \boldsymbol{\mu}}$ maps the external elements e to the element x in B with lowest value among those that can be swapped with e . For the sake of conciseness, we denote σ the mapping $\sigma_{B^*, \boldsymbol{\mu}}$.

Using these mappings, we define the graph $G_{\boldsymbol{\mu}} = (S, A)$, where $S = \mathcal{B}$, and the neighbors of a basis B is $\mathcal{N}_{B, \boldsymbol{\mu}} \triangleq \cup_{e \in E \setminus B} B - \sigma_{B, \boldsymbol{\mu}}(e) + e$. We below show that unimodality holds in $G_{\boldsymbol{\mu}}$.

Theorem 4.1 (Unimodality¹). *Let $M = (E, \mathcal{I})$ be a matroid of bases \mathcal{B} , and let $\boldsymbol{\mu}$ denote the expectations on elements E . For any basis $B \in \mathcal{B}$ such that $\mu_B \neq \max_{I \in \mathcal{I}} \mu_I$, there exists $B^+ \in \mathcal{N}_{B, \boldsymbol{\mu}}$ with $\mu_{B^+} > \mu_B$.*

Proof. Let B be a basis of M and denote B^* the optimal basis. As $\mu_B \neq \max_{I \in \mathcal{I}} \mu_I$, $B \neq B^*$ and by rank-property of matroids, $B \setminus B^*$ is non-empty. let $x = \arg \min_{\tilde{x} \in B \setminus B^*} \mu_{\tilde{x}}$ be the element in $B \setminus B^*$ with lowest value. We shall prove that one can create a subset $B - x + \alpha$ (belonging to the neighborhood of B) with higher value than B .

Using Lemma 3.2, $\exists \alpha \in B^* \setminus B$ such that $B - x + \alpha \in \mathcal{B}$. By definition of x , $\sigma_{B, \boldsymbol{\mu}}(\alpha) = x$, and therefore $B - x + \alpha \in \mathcal{N}_{B, \boldsymbol{\mu}}$.

It holds that $\mu_{B - x + \alpha} - \mu_B = \mu_\alpha - \mu_x$, hence we wonder the sign of $\mu_\alpha - \mu_x$. Assume $\mu_\alpha < \mu_x$. The matroid axiom (iii) implies that $\exists \beta \in B \setminus (B^* -$

¹Talebi and Proutiere [2016] implicitly use this result, within the regret analysis of KL-OSM, when showing that sets $\mathcal{K}_i = \{l \in B^*, B^* - l + i \in \mathcal{B}\}$ are non-empty, allowing to consider their minima l_i , which satisfies $\mu_{l_i} > \mu_i$.

α , $B^* - \alpha + \beta \in \mathcal{B}$. But then, $\mu_\beta \geq \mu_x$ (by definition of x), so that $\mu_\beta \geq \mu_x > \mu_\alpha$. Consequently, $\mu_{B^* - \alpha + \beta} > \mu_{B^*}$, which is absurd. Hence $\mu_\alpha \geq \mu_x$, and even $\mu_\alpha > \mu_x$ as $\alpha \neq x$. Therefore $\mu_{B - x + \alpha} > \mu_B$, which concludes the proof.

4.2 MAUB

We now introduce MAUB, which follows the unimodal bandit framework of Combes and Proutiere [2014]: it maintains a *leader* (i.e., the best arm identified so far), and plays optimistically within its neighborhood.

In contrast to OSUB, the graph $G_{\hat{\mu}(t)}$ used by MAUB (defined below) is not static and may fail to be unimodal with respect to the true—only partially known—optimization problem. Nevertheless, we establish in the regret analysis (see Lemma 5.2) that $G_{\hat{\mu}(t)}$ most of the time contains the correct neighborhood. Before moving on to a detailed description of MAUB, we define several statistics of interest.

Definition 4.2 (Unimodal Bandit Statistics). *We respectively denote by $L(t)$ and $P(t)$ the leader and the arm played at time step t . Additionally, let:*

- $l_B(t) \triangleq \sum_{s=1}^t \mathbb{E}[\mathbf{1}_{L(s)=B}]$ be the number of times B was leader up to time t ;
- $\hat{\mu}_e(t)$ the average value of $(X_e(s))$ for all time steps $s \leq t$ at which e was played.

Let us then define $\bar{\mu}_B(t, L) \triangleq \sum_{e \in B} [\hat{\mu}_e(t-1) + \sqrt{2 \cdot \log(l_L(t))/N_e(t-1)}]$ be the optimistic estimator of B , where $N_e(t) = \sum_{s=0}^t \mathbf{1}_{e \in P(s)}$ is the number of times e was in a basis that was played at some time step $s \leq t$. $l_L(t)$ is a “local time”, in comparison with the classical $\log(t)$ term in UCB bonuses. It indicates how much time was spent on the current leader, and is tighter than using total time spent to decide whether external elements $e \in E \setminus L(t)$ should be tested in comparison with elements of the leader.

A pseudocode of MAUB is given in Algorithm 3. It repeatedly performs three steps: (i) if needed, compute the new current leader and its corresponding neighborhood (Algorithms 3 to 3); (ii) optimistically pick an arm to play within the current neighborhood

(Algorithms 3 to 3), and (iii) observe the results and update statistics (Algorithms 3 to 3).

Leader Computations Leader computations aim at identifying the best arm, according to current mean statistics. They boil down to a call to greedy. Whenever a new leader is computed, neighborhood shall also be updated, meaning that Algorithm 2 is called.

Algorithm 2 Neighborhood Computation

Require: base B , element values μ

- 1: Neighbors $\leftarrow \emptyset$
- 2: NotMapped $\leftarrow E \setminus B$
- 3: **for** $x \in B$ in order $\mu_{x_1} \leq \dots \leq \mu_{x_D}$ **do**
- 4: **for** $e \in \text{NotMapped}$ **do**
- 5: **if** $B - x + e \in \mathcal{B}$ **then**
- 6: Neighbors $\leftarrow \text{Neighbors} \cup \{B - x + e\}$
- 7: NotMapped $\leftarrow \text{NotMapped} - e$
- 8: **end if**
- 9: **end for**
- 10: **end for**
- 11: **Return** Neighbors

Picking the Arm to Play Identifying the arm to play within the current neighborhood of size at most γ (Line 17) can be implemented in linear time $\mathcal{O}(\gamma)$ by remarking that, $\forall t, \forall B = L(t) - x + y \in \mathcal{N}_{L(t), \hat{\mu}(t)}$,

$$\sum_{e \in L(t)} \bar{\mu}_e(t) - \sum_{e \in B} \bar{\mu}_e(t) = \bar{\mu}_x(t) - \bar{\mu}_y(t). \quad (2)$$

Leader Changes and Neighborhood Updates A significant difference with previous unimodal bandits ([Combes and Proutiere, 2014, Gauthier et al., 2021, 2022]) is that MAUB does not recompute a leader at each iteration. Instead, and based on the regret analysis, it is sufficient to stick with the same leader $L(t)$ as in previous iterations, as long as according to mean statistics $\hat{\mu}$ it has highest value within its neighborhood (Line 7). Additionally, as long as the leader remains the same, the neighborhood stays valid provided the order $\hat{\mu}_{e_1} \leq \dots \leq \hat{\mu}_{e_{|L|}}$ of the leader’s elements is unchanged (Line 10). Verifying the validity

of the leader and neighborhood is crucial, as it respectively saves $\mathcal{O}(|E|)$ and $\mathcal{O}(D(|E| - D))$ oracle calls whenever they remain unchanged.

5 Theoretical Analysis

Let now analyze both the regret and the time complexity of MAUB.

5.1 Regret Upper-bound

We start by restating the concentration inequality introduced by Combes and Proutiere [2014], that is at the core of the analysis of MAUB. It allows bounding the expected deviation of an arm at specific time steps.

Algorithm 3 MAUB

```

1: Play every element once
2: Denote  $\hat{\mu}(0)$  current estimates of  $\mu$ 
3:  $L \leftarrow \text{greedy}(\hat{\mu}(0))$ 
4:  $\mathcal{N} \leftarrow \text{ComputeNeighbors}(L, \hat{\mu}(0))$ 
5: for  $t = 1, 2, \dots$  do
6:   /* Update leader and neighborhood */
7:   if  $\hat{\mu}_L(t-1) < \max_{B \in \mathcal{N}} \hat{\mu}_B(t-1)$  then
8:      $L \leftarrow \text{greedy}(\hat{\mu}(t-1))$ 
9:      $\mathcal{N} \leftarrow \text{ComputeNeighbors}(L, \hat{\mu}(t-1))$ 
10:    else if for the elements of  $L$ , the order
11:       $\hat{\mu}_{e_1} \leq \dots \leq \hat{\mu}_{e_D}$  has changed then
12:         $\mathcal{N} \leftarrow \text{ComputeNeighbors}(L, \hat{\mu}(t-1))$ 
13:    end if
14:    /* Choose the arm to play */
15:    if  $l_L(t) - 1 \equiv 0 [|E| - D + 1]$  then
16:       $P \leftarrow L$ 
17:    else
18:       $P \leftarrow \arg \max_{B \in \{L\} \cup \mathcal{N}} \bar{\mu}_B(t, L)$ 
19:    end if
20:    /* Play and update statistics */
21:    Observe  $X_e$  for all  $e \in P$ 
22:     $\forall e \in P, N_e(t) \leftarrow N_e(t-1) + 1$ 
23:     $\forall e \in P, \hat{\mu}_e(t) \leftarrow \frac{N_e(t-1) \cdot \hat{\mu}_e(t-1) + X_e}{N_e(t)}$ 
24:     $L(t) \leftarrow L, \mathcal{N}(t) \leftarrow \mathcal{N}, P(t) \leftarrow P$ 
25: end for
```

Lemma 5.1. [Combes and Proutiere, 2014, Lemma B.1] Let $k \in E$ and $\epsilon > 0$. Define \mathcal{F}_n the σ -algebra generated by $(X_k(t))_{t \leq n, k \in E}$. Let $\Lambda \subseteq \mathbb{N}$ be a random set of instants. Assume that there exists a sequence of random sets $(\Lambda(s))_{s \geq 1}$ such that (i) $\Lambda \subseteq \cup_{s \geq 1} \Lambda(s)$, (ii) $\forall s \geq 1, \forall n \in \Lambda(s), t_k(n) \geq \epsilon s$, (iii) $\forall s, |\Lambda(s)| \leq 1$, and (iv) the event $\{n \in \Lambda(s)\}$ is \mathcal{F}_n -measurable. Then, $\forall \delta > 0$:

$$\mathbb{E} \left[\sum_{n \geq 1} \mathbb{1}\{n \in \Lambda, |\hat{\mu}_k(n) - \mu_k| > \delta\} \right] \leq \frac{1}{\epsilon \delta^2} \quad (3)$$

The following lemma builds upon this result to bound the number of time steps for which unimodality is not satisfied. In light of Theorem 4.1, this can only happen whenever the order according to mean statistics within the elements of current leader $L(t)$ is wrong.

Lemma 5.2. (Number of iterations with incorrect neighborhood, proof in App. B) Let $B \in \mathcal{B}$. Let $x^* = \arg \min_{\tilde{x} \in B \setminus B^*} [\mu_{\tilde{x}}]$. Let $y^* \in \{\tilde{y} \in B^* \setminus B, B - x^* + \tilde{y} \in \mathcal{B}\}$. It holds that

$$\mathbb{E} \left[\sum_{s=1}^T \mathbb{1}_{\{L(s)=B, B-x^*+y^* \notin \mathcal{N}_{B, \hat{\mu}(s)}\}} \right] = \mathcal{O}_{T \rightarrow \infty}(1).$$

Lemma 5.2 permits following a similar proof architecture as Combes and Proutiere [2014], since we may

now stick with iterations satisfying unimodality, and bound the average time spent in suboptimal arms.

Theorem 5.3. (*Time spent in suboptimal arms, Proof in App. C*) Any suboptimal arm is the leader $\mathcal{O}(\log \log(T))$ times in expectation.

The regret coming from suboptimal arms being leaders is thus negligible compared to the regret when B^* is the leader, which resembles a regular bandit problem on the neighborhood of B^* . The final regret analysis of MAUB follows from this observation and is formalized in the following theorem.

Theorem 5.4. (*Proof in App. D*) The regret of MAUB up to time step T is

$$\mathcal{O}\left(\sum_{e \notin B^*} \frac{8}{\mu_{\sigma(e)} - \mu_e} \log(T)\right) = \mathcal{O}\left(\frac{|E| - D}{\Delta_{\min}} \log T\right),$$

with $\Delta_{\min} \triangleq \min_{e \notin B^*} \mu_{\sigma(e)} - \mu_e$.

The constant $\sum_{e \notin B^*} \frac{8}{\mu_{\sigma(e)} - \mu_e}$ is slightly better than OMM's constant $\sum_{e \notin B^*} \frac{16}{\min_{i \in B^*, \mu_i > \mu_e} \mu_i - \mu_e}$ since for any element $e \notin B^*$, $\sigma(e) \in B^*$. Yet, we believe that a finer analysis of OMM's behavior would show that OMM actually achieves the same regret as MAUB.

5.2 Time Complexity

The time complexity analysis of MAUB mainly builds upon Theorem 5.3: suboptimal arms can only be leaders for a negligible amount of time steps. Hence B^* is the leader for most time steps, and there are few iterations involving a leader change or a neighborhood update, which drastically decreases the overall number of oracle calls compared to state of the art.

Corollary 5.5. The expected overall time complexity of MAUB is

$$\begin{aligned} & \mathcal{O}(|E| \cdot T + |\mathcal{B}|(|E| - D)^2 \log \log T \\ & \quad \cdot [|E|(\log |E| + \mathcal{T}_m) + D(|E| - D)\mathcal{T}_m]) \\ & = \mathcal{O}(|E| \cdot T + |\mathcal{B}| \text{poly}(|E|, D)\mathcal{T}_m \log \log(T)), \end{aligned}$$

where $|\mathcal{B}| \leqslant \binom{|E|}{D}$.

The first term comes from identifying the arm to play in the neighborhood and the second one corresponds to neighborhood computations. In fact, all iterations take at least $\mathcal{O}(|E| - D)$, and there are $\mathcal{O}(\log \log(T))$ iterations such that the per-round time complexity suffers an additional time, which is detailed in the proof.

Proof. First, the identification of the arm to play requires at each iteration: the computation of $|E|$ optimistic terms, and $|E| - D$ comparisons of two optimistic terms. This induces a total computation complexity of $\mathcal{O}(|E| \cdot T)$. Other tests and updates of statistics have the same computational complexity.

Second, let us account for the changes of leader. Let $T \in \mathbb{N}^* \setminus \{1\}$ and let $C^T = \{n \leq T \mid L(n-1) \neq L(n)\}$. Because of the assumption that all bases have distinct values, $C^T \subseteq D^T \cup E^T$, where:

- $D^T = \{n \leq T \mid \mu_{L(n-1)} > \mu_{L(n)}\}$ is the number of suboptimal leader changes;
- $E^T = \{n \leq T \mid \mu_{L(n-1)} < \mu_{L(n)}\}$ is the number of leader changes improving current value.

The proof relies on Theorem 5.3 and remarking that in both D^T and E^T , the leader changes either come from a suboptimal arm or lead to a suboptimal arm. It holds that $D^T \subseteq \{n \leq T \mid L(n) \neq B^*\}$, and $E^T \subseteq \{n \leq T \mid L(n-1) \neq B^*\}$. Hence, from Theorem 5.3, the expectation of the size of both D^T and E^T is $\mathcal{O}(\gamma^2 |\mathcal{B}| \log \log(T))$, and then, there are $\mathcal{O}(\gamma^2 |\mathcal{B}| \log \log T)$ leader changes, each requiring $\mathcal{O}(|E|(\log |E| + \mathcal{T}_m) + D(|E| - D)\mathcal{T}_m)$ time.

Finally, let us upper-bounds the expected computation cost of neighborhood updates without leader change (Line 11). The proof follows the same structure as the one on leader changes, except that we use Lemma 5.2. It leads to a $\mathcal{O}(D(|E| - D)\mathcal{T}_m)$ cost for T iterations, which is negligible with respect to other terms.

6 Experiments

In this section, MAUB is evaluated against OMM on four different matroid benchmarks, namely uniform, linear, graphic, and transversal matroids. Section 6.3

Table 2: Overall time complexity of oracle calls for different types of matroids.

Matroid	Time Comp.	Oracle
Uniform	$\mathcal{O}(1)$	cardinal test
Graphic	$\mathcal{O}(\log(E))$	cycle test
Transversal	$\mathcal{O}(D E)$	augmenting path on matching
Linear	$\mathcal{O}(D^2)$	Gaussian pivot step

discusses the experimental results, but we first detail the experimental protocol in Section 6.1, and give usefull background on benchmarked matroids in Section 6.2.

6.1 Experimental protocol

For each matroid, the learning algorithm observes, at time step t , realizations of Gaussian distributions $(X_e)_{e \in P(t)}$, where $\forall e, X_e \sim \mathcal{N}(\mu_e, \sigma^2)$. Mean values $\boldsymbol{\mu}$ are randomly set in $[0.5, 1]$ (except for linear matroid, see corresponding paragraph) at the beginning of the experiment and the standard deviation σ is set to 0.2 for all distributions.

Each experiment ran with 20 different seeds, and we capture (i) the regret (see Equation (1)), as a function of iterations, given in Figure 1; and (ii) overall time complexity, number of oracle and greedy calls, and MAUB’s number of order change within current leaders (Line 11) provided in Table 3.

Reproducibility The code to reproduce the experiments is available in supplementary material. All algorithms are implemented using Python3.11. We use the *Sagemath* library [The Sage Developers, 2025] to handle matroids. Graphic and uniform matroids are taken from Sagemath’s database, while the linear and transversal matroids are constructed as Sagemath matroids using external data. More detail is given in their corresponding paragraphs. The experiments were run using one core of a 2.20 GHz Intel® Xeon® Gold 5320 CPU, part of a cluster with 384 GiB available RAM.

6.2 Experimental Domains

We detail here the matroids considered in the experiments. The time complexity \mathcal{T}_m of the membership oracle for the four classes of matroid are given in Table 2.

Uniform Matroids In these experiments, all subsets of size at most D are independent, so the problem reduces to identifying the D elements with the highest values.

For such matroids, the independence oracle simply checks the size of a subset I , which is done in $\mathcal{O}(1)$ time.

Linear Matroid Learning in linear matroids is illustrated here by searching for set of independent movies with highest rating. Similarly to Kveton et al. (2014) 100 movies from the dataset *Movielens* are considered. The database attributes types, among 18 possible ones, to each movie. A set of movies I is independent if and only if the characteristic vectors $\{u_e \in \{0, 1\}^{18}, e \in I\}$ form a linearly independent family of \mathbb{R}^{18} . The latter constraint ensures diversity to some extent. Values of movies are given by users’ average ratings $\boldsymbol{\mu}$, with $\forall e, \mu_e \in [0, 5]$.

The independence oracle for linear matroids involves one step in Gaussian pivot, which takes $\mathcal{O}(D^2)$ time.

Graphic Matroids For any graph $G = (S, A)$, a graphic matroid is defined by the set of spanning trees of the graph. μ As Talebi and Proutiere (2016), we consider the graphic matroid K_N associated to the complete graph of size N .

Checking whether adding an element x to an independent set I creates a circuit can be done in $\mathcal{O}(\log(|E|))$ time, using union-find data structure.

Transversal Matroid Given a bipartite graph $G = (X \cup Y, A)$, one can define a matroid where the groundset is the vertices in X , and a subset $B \subseteq X$ is independent if and only if it admits a matching with Y . An arbitrary graph, available in supplementary material, with $|X| = 7$, $|Y| = 6$, and 17 edges is considered.

Table 3: Time statistics for MAUB and OMM on several matroids. Computation time is in seconds and other metrics correspond to numbers of calls (averaged over 20 runs). Neigh. Up. denotes the number of neighborhood updates performed by MAUB while the leader remains stable. MAUB significantly **reduces the number of oracle calls** compared to OMM, which translates into **lower overall computation time**.

(a) Uniform matroids					(b) Graphic matroids				
Algorithm	Time	Oracle Calls	Greedy Calls	Neigh. Up.	Algorithm	Time	Oracle Calls	Greedy Calls	Neigh. Up.
Uniform ($D = 7, E = 10, \mathcal{B} = 120$)									
OMM	3.45 s	$7 \cdot 10^5$	$1 \cdot 10^5$	-	OMM	11.28 s	$6.56 \cdot 10^5$	$1 \cdot 10^5$	-
MAUB	1.92 s	$4.97 \cdot 10^2$	$1.38 \cdot 10^1$	$1.22 \cdot 10^2$	MAUB	7.83 s	$7.97 \cdot 10^2$	$2.31 \cdot 10^1$	$6.04 \cdot 10^1$
Uniform ($D = 7, E = 15, \mathcal{B} = 6435$)									
OMM	4.44 s	$7 \cdot 10^5$	$1 \cdot 10^5$	-	OMM	22.24 s	$1.44 \cdot 10^6$	$1 \cdot 10^5$	-
MAUB	3.84 s	$2.68 \cdot 10^3$	$4.82 \cdot 10^1$	$2.47 \cdot 10^2$	MAUB	18.02 s	$7.03 \cdot 10^3$	$6.47 \cdot 10^1$	$1.66 \cdot 10^2$
Uniform ($D = 15, E = 20, \mathcal{B} = 15504$)									
OMM	6.87 s	$1.5 \cdot 10^6$	$1 \cdot 10^5$	-	OMM	131.89 s	$8.84 \cdot 10^6$	$1 \cdot 10^5$	-
MAUB	3.50 s	$3.33 \cdot 10^3$	$4.02 \cdot 10^1$	$5.07 \cdot 10^2$	MAUB	108.43 s	$5.82 \cdot 10^5$	$1.85 \cdot 10^2$	$1.58 \cdot 10^3$
Uniform ($D = 15, E = 30, \mathcal{B} = 1.55 \cdot 10^8$)									
OMM	8.84 s	$1.5 \cdot 10^6$	$1 \cdot 10^5$	-	OMM	262.55 s	$1.67 \cdot 10^7$	$1 \cdot 10^5$	-
MAUB	7.51 s	$1.28 \cdot 10^4$	$7.61 \cdot 10^1$	$7.05 \cdot 10^2$	MAUB	235.67 s	$3.5 \cdot 10^6$	$3.57 \cdot 10^2$	$3.77 \cdot 10^3$
(c) Linear matroid					(d) Transversal matroid				
Algorithm	Time	Oracle Calls	Greedy Calls	Neigh. Up.	Algorithm	Time	Oracle Calls	Greedy Calls	Neigh. Up.
Linear ($D = 16, E = 100, \mathcal{B} $ unknown)									
OMM	72.69 s	$9.7 \cdot 10^6$	$1 \cdot 10^5$	-	OMM	83.42 s	$6 \cdot 10^6$	$1 \cdot 10^6$	-
MAUB	60.49 s	$2.28 \cdot 10^5$	$4.96 \cdot 10^1$	$2.03 \cdot 10^2$	MAUB	17.44 s	$2.28 \cdot 10^2$	$2.56 \cdot 10^1$	$5.08 \cdot 10^1$

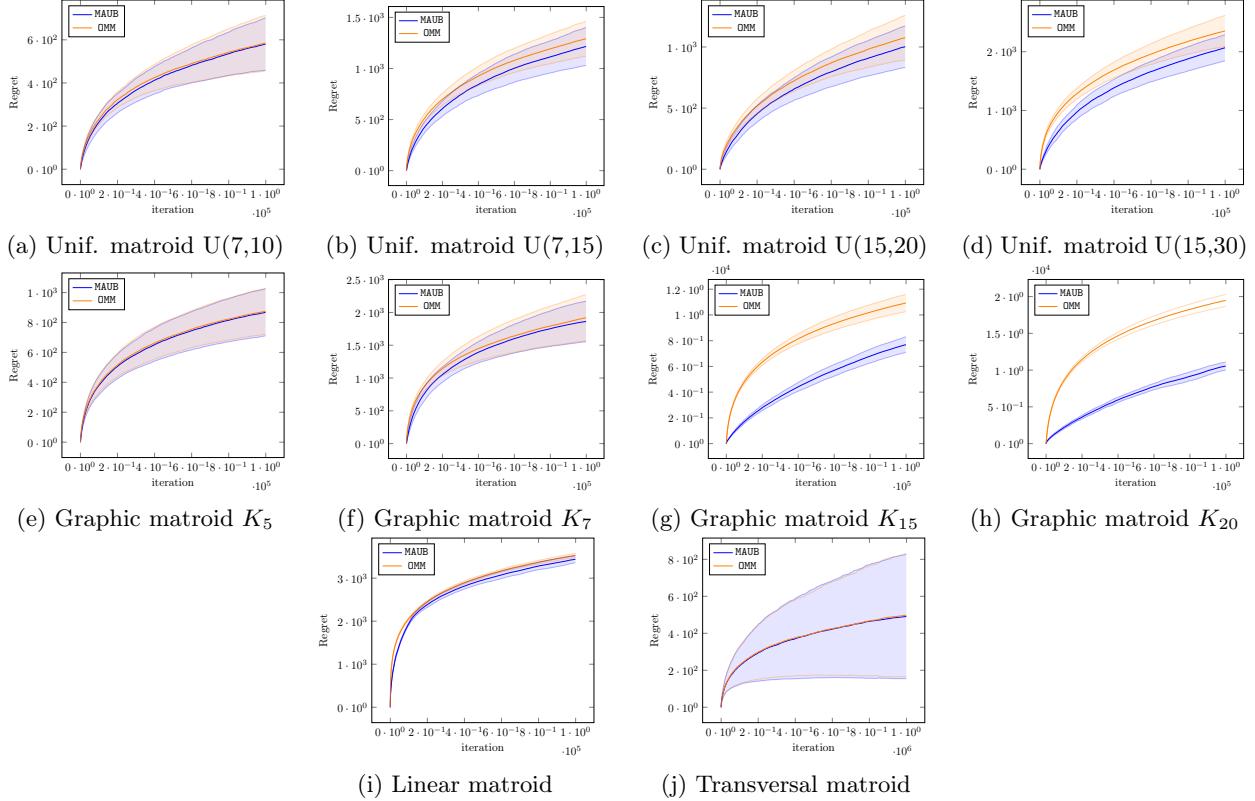


Figure 1: Regret vs iterations for uniform, linear, graphic and transversal matroids (the smaller, the better). $U(D, |E|)$ is the uniform matroid of rank D on $|E|$ elements. K_N it the graphic matroid associated to the complete graph of size N . Our algorithm MAUB consistently **matches or outperforms** OMM.

Assuming that I is independent, and that a matching M is stored, deciding whether $I + x$ is independent can be reduced to checking if an augmenting path in G with respect to M can be found. The independence oracle for transversal matroid thus takes $\mathcal{O}(D|E|)$ time.

6.3 Results

Figure 1 shows no loss in regret of MAUB compared to OMM in asymptotic regime, and an improved regret in early iterations for large matroids (*e.g.* $U(15; 30)$, linear matroid, K_{15} , K_{20}). We believe that this phenomenon is due MAUB’s restricted exploration, while OMM might test several suboptimal elements in the

same iteration.

The number of oracle calls given in Table 3 is drastically reduced compared to OMM (ranging from one order of magnitude (*e.g.* K_{20} , Linear) to three (*e.g.* $K_{3,3}$, transversal)). This is due to the number of iterations performed by MAUB requiring calls to the matroid structure (*i.e.* number of greedy calls plus number of order change requiring neighborhood computations) being significantly lower than the number of greedy calls performed by OMM. As expected, this smaller number of oracle calls translates into lower computation time.

A side observation is that constant C_M in Theorem 5.5 is largely overestimated, as pointed out by the relatively low number of greedy calls performed

by MAUB in experiments on matroids with intractable total number of bases (*e.g.* $U(15, 20)$, K_{20} , linear).

7 Discussion and Perspectives

We introduced MAUB, a unimodal bandit approach for learning in matroid combinatorial semi-bandit problems. While optimal regret is already achieved by the state of the art for this combinatorial structure, repeated queries to the underlying optimization problem can be a major computational bottleneck for certain classes of matroids. As supported by the theoretical analysis and the empirical experiments on various classes of matroids, MAUB leverages the unimodality of matroids to drastically reduce the number of oracle calls which results in an improved overall time complexity (for no loss in terms of regret). Beyond our study, we believe that the present contribution lays down a promising path to tackle time-complexity reduction in other semi-bandit problems.

Our work also underscores that the current unimodal bandit analysis lacks a fine-grained characterization of actual early trajectories taken by the learning algorithm. For instance, current theoretical analyses of combinatorial unimodal bandits ignore correlation between suboptimal arms, leading to a large overestimation of some constant terms in the overall regret and time complexity. Filling this theoretical gap would widen the relevance of the unimodal approach to the non-asymptotic regime, beyond the specific matroid structure considered in this paper.

References

- Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: fast combinatorial optimization with learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.
- Matthew Streeter, Daniel Golovin, and Andreas Krause. Online learning of assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009.
- Lijie Chen, Anupam Gupta, and Jian Li. Pure exploration of multi-armed bandit under matroid constraints. *CoRR*, abs/1605.07162, 2016.
- Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.
- Andrew Clark, Linda Bushnell, and Radha Poovendran. On leader selection for performance and controllability in multi-agent systems. In *IEEE conference on decision and control (CDC)*, 2012.
- Fu Lin, Makan Fardad, and Mihailo R Jovanović. Algorithms for leader selection in large dynamical networks: Noise-corrupted leaders. In *IEEE Conference on Decision and Control and European Control Conference (CDC)*, 2011.
- Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, 2013.
- Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, et al. Combinatorial bandits revisited. *Advances in neural information processing systems (NeurIPS)*, 2015.
- Mohammad Sadegh Talebi and Alexandre Proutiere. An optimal algorithm for stochastic matroid bandit optimization. In *International Conference on Autonomous Agents & Multiagent Systems (AA-MAS)*, 2016.
- Ruo-Chun Tzeng, Naoto Ohsaka, and Kaito Ariu. Matroid semi-bandits in sublinear time. In *International Conference on Machine Learning (ICML)*, 2024.
- Eric W Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control (TAC)*, 2009.
- Jia Yuan Yu and Shie Mannor. Unimodal bandits. In *International Conference on Machine Learning (ICML)*, 2011.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning (ICML)*, 2014.

Cindy Trinh, Emilie Kaufmann, Claire Vernade, and Richard Combes. Solving bernoulli rank-one bandits with unimodal thompson sampling. In *Algorithmic Learning Theory (ALT)*, 2020.

Camille-Sovanneary Gauthier, Romaric Gaudel, Elisa Fromont, and Boammani Aser Lompo. Parametric graph for unimodal ranking bandit. In *International Conference on Machine Learning (ICML)*. PMLR, 2021.

Camille-Sovanneary Gauthier, Romaric Gaudel, and Elisa Fromont. UniRank: Unimodal bandit algorithms for online ranking. In *International Conference on Machine Learning (ICML)*, 2022.

Pierre Perrault, Vianney Perchet, and Michal Valko. Exploiting structure of uncertainty for efficient matroid semi-bandits. In *International Conference on Machine Learning (ICML)*, 2019.

Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Alexander Schrijver. A course in combinatorial optimization. *CWI, Kruislaan*, 413, 2003.

Jack Edmonds. Matroids and the greedy algorithm. *Mathematical programming*, 1:127–136, 1971.

The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 8.1)*, 2025. <https://www.sagemath.org>.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 2002.

To help the reader, we provide below (Figure 2) a proof diagram summarizing the finite-time analysis of MAUB conducted in the appendix.

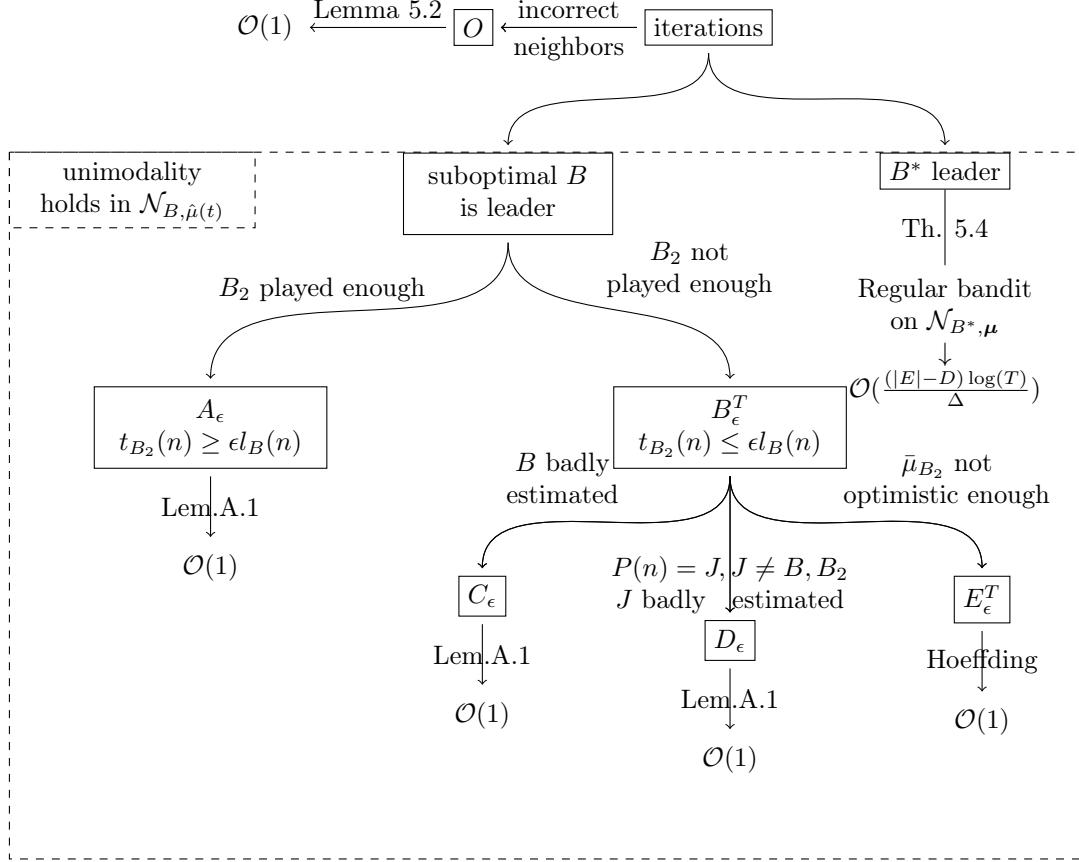


Figure 2: Proof diagram of finite-time analysis of MAUB. We remind the reader that whenever unimodality is satisfied, B_2 is an arm in $\mathcal{N}_{B,\mu}$ such that $\mu_{B_2} > \mu_B$.

A Concentration Inequality

We start by restating two core lemmas introduced by Combes and Proutiere [2014]. They allow bounding the expected number of high deviations of mean statistics at specific time steps.

Lemma A.1. [Combes and Proutiere, 2014, Lemma B.1] Let $k \in E$ and $\epsilon > 0$. Define \mathcal{F}_n the σ -algebra generated by $(X_k(t))_{t \leq n, k \in E}$. Let $\Lambda \subseteq \mathbb{N}$ be a random set of instants. Assume that there exists a sequence of random sets $(\Lambda(s))_{s \geq 1}$ such that (i) $\Lambda \subseteq \cup_{s \geq 1} \Lambda(s)$, (ii) $\forall s \geq 1, \forall n \in \Lambda(s), t_k(n) \geq \epsilon s$, (iii) $\forall s, |\Lambda(s)| \leq 1$,

and (iv) the event $\{n \in \Lambda(s)\}$ is \mathcal{F}_n -measurable. Then, $\forall \delta > 0$:

$$\mathbb{E} \left[\sum_{n \geq 1} \mathbb{1}\{n \in \Lambda, |\hat{\mu}_k(n) - \mu_k| > \delta\} \right] \leq \frac{1}{\epsilon \delta^2} \quad (4)$$

Lemma A.2. [Combes and Proutiere, 2014, Lemma B.2] Let $k, k' \in E$, $k \neq k'$ and $\epsilon > 0$. Define \mathcal{F}_n the σ -algebra generated by $(X_k(t))_{t \leq n, k \in \{1, \dots, |E|\}}$. Let $\Lambda \subset \mathbb{N}$ be a random set of instants. Assume that there exists a sequence of random sets $(\Lambda(s))_{s \geq 1}$ such that (i) $\Lambda \subset \cup_{s \geq 1} \Lambda(s)$, (ii) $\forall s \geq 1$, $\forall n \in \Lambda(s)$, $t_k(n) \geq \epsilon s$, and $t_{k'}(n) \geq \epsilon s$, (iii) $\forall s$, $|\Lambda(s)| \leq 1$ almost surely, (iv) $\forall n \in \Lambda$, $\mathbb{E}[\hat{\mu}_k(n)] \leq \mathbb{E}[\hat{\mu}_{k'}(n)] - \Delta_{k,k'}$, and (v) the event $\{n \in \Lambda(s)\}$ is \mathcal{F}_n -measurable. Then, $\forall \delta > 0$:

$$\mathbb{E} \left[\sum_{n \geq 1} \mathbb{1}\{n \in \Lambda, \hat{\mu}_k(n) > \hat{\mu}_{k'}(n)\} \right] \leq \frac{8}{\epsilon \Delta_{k,k'}^2} \quad (5)$$

B Incorrect Neighborhood

This section proves that the number of iterations for which unimodality is not ensured is finite, in expectation. The proof relies on remarking that Th.4.1 is not guaranteed to hold only for wrong ordering $\hat{\mu}_{e_{i_1}} < \dots < \hat{\mu}_{e_{i_{|L(t)|}}}$ of items in the current leader $L(t)$.

Lemma 5.2. (Number of iterations with incorrect neighborhood, originally stated on page 6) Let $B \in \mathcal{B}$. Let $x^* = \arg \min_{\tilde{x} \in B \setminus B^*} [\mu_{\tilde{x}}]$. Let $y^* \in \{\tilde{y} \in B^* \setminus B, B - x^* + \tilde{y} \in \mathcal{B}\}$. It holds that

$$\mathbb{E} \left[\sum_{s=1}^T \mathbb{1}\{L(s)=B, B-x^*+y^* \notin \mathcal{N}_{B,\hat{\mu}(s)}\} \right] = \mathcal{O}_{T \rightarrow \infty}(1).$$

Proof. Let us define

$$O \triangleq \{s \leq T \mid L(s) = B, B - x^* + y^* \notin \mathcal{N}_{B,\hat{\mu}(s)}\},$$

and denote $\delta = \min_{a,b \in E, a \neq b} |\mu_a - \mu_b|$ the minimal gap between two elements in E . By design of Algorithm 2,

$$\begin{aligned} O &= \{s \leq T \mid B = L(s), \exists x \in \{\tilde{x} \in B, B - \tilde{x} + y^* \in \mathcal{B}\}, \hat{\mu}_{x^*}(s) > \hat{\mu}_x(s)\} \\ &\subseteq \cup_{x \in B - x^*} \{s \leq T \mid L(s) = B, \hat{\mu}_{x^*}(s) > \hat{\mu}_x(s)\} \\ &\subseteq \cup_{x \in B - x^*} \{s \leq T \mid L(s) = B, |\hat{\mu}_{x^*}(n) - \mu_{x^*}| > \frac{\delta}{2} \vee |\hat{\mu}_x(n) - \mu_x| > \frac{\delta}{2}\} \\ &\subseteq \bigcup_{x \in B} \{s \leq T \mid L(s) = B, |\hat{\mu}_x(n) - \mu_x| > \frac{\delta}{2}\}. \end{aligned}$$

Let $x \in B$ and bound the expected size of $O_x \triangleq \{s \leq T \mid L(s) = B, |\hat{\mu}_x(n) - \mu_x| > \frac{\delta}{2}\}$. Let $\Lambda \triangleq \{n \leq T, L(n) = B\}$. It holds that $O_x = \{n \in \Lambda, |\hat{\mu}_x(n) - \mu_x| > \frac{\delta}{2}\}$. For all s , let us define $\Lambda(s) = \{n \leq T, L(n) = B, l_B(n) = s\}$ so that $\Lambda \subseteq \cup_s \Lambda(s)$. Since l_B is increased each time B is the leader, $\forall s$, $|\Lambda(s)| \leq 1$. x being in B , and by design of MAUB forcing the leader to be played enough, for any $s \geq 1$ and $n \in \Lambda(s)$, $t_x(n) \geq t_B(n) \geq l_B(n)/(\gamma + 1) = s/(\gamma + 1)$. Then, by Lemma A.1, for $\epsilon \triangleq 1/(\gamma + 1)$

$$\mathbb{E}[|O_x|] = \mathbb{E} \left[\sum_n \mathbb{1}\{n \in \Lambda, |\hat{\mu}_x(n) - \mu_x| > \frac{\delta}{2}\} \right] \leq \frac{4}{\epsilon \delta^2}.$$

Since this holds for any x ,

$$\mathbb{E}[|O|] \leq \sum_x \mathbb{E}[|O_x|] \leq |B| \frac{4}{\epsilon \delta^2} = \mathcal{O}(1).$$

C Suboptimal Arms

We are now ready to prove the key theorem for the regret analysis, stating that suboptimal arms can only be leaders for a $\mathcal{O}(\log \log T)$ amount of time. The rest of the time then must be spent with the optimal arm being the leader, which resembles a regular bandit.

Theorem 5.3. (*Time spent in suboptimal arms, originally stated on page 7*) Any suboptimal arm is the leader $\mathcal{O}(\log \log(T))$ times in expectation.

Proof. Let B be a suboptimal arm, and let $B^* = \arg \max_{B \in \mathcal{B}} \mu_B$ be the global optimum. Let $B_2 \triangleq B - x^* + y^* = \arg \max_{J \in \mathcal{N}_{B,\mu}} \mu_J$ be the best arm in the neighborhood of B .

It holds that

$$\{n \leq T \mid L(n) = B\} \subseteq V \cup A_\epsilon \cup B_\epsilon^T, \quad (6)$$

where:

- $V \triangleq \{n \mid L(n) = B, \exists x, y \in L(n), \hat{\mu}_x(n) > \hat{\mu}_y(n), \mu_x < \mu_y\}$; and
- for any $n \notin V$, it holds that $B_2 \in \mathcal{N}_{B,\hat{\mu}(n)}$, and A_ϵ and B_ϵ^T can be defined as follows:
 - $A_\epsilon \triangleq \{n \notin V \mid L(n) = B, t_{B_2}(n) \geq \epsilon l_B(n)\}$;
 - $B_\epsilon^T \triangleq \{n \leq T, n \notin V \mid L(n) = B, t_{B_2}(n) \leq \epsilon l_B(n)\}$.

Remark C.1. Note that V resembles O , the number of iterations which do not satisfy unimodality. We artificially also include in O the time steps with wrong ordering for at least two elements in B . The proof should hold using O , but V allows for a static neighborhood when bounding B_ϵ^T , which circumvents technical details, with no loss in asymptotic regret.

Similarly to bounding $|O|$, we show that $|V| = \mathcal{O}(1)$. We use the following decomposition:

$$V \subseteq \bigcup_{x,y \in B^2, \mu_x < \mu_y} \{n \in \Lambda_{x,y} \mid \hat{\mu}_x > \hat{\mu}_y\}; \quad (7)$$

where $\forall x, y, \forall s, \Lambda_{x,y}(s) \triangleq \{n \mid L(n) = B, l_B(n) = s\}$. For all x, y ,

- $\forall s, |\Lambda_{x,y}(s)| \leq 1$;
- $\forall s, \forall n \in \Lambda(s), t_x(n) \geq t_B(n) \geq l_B(n)/(\gamma + 1) = s/(\gamma + 1) \geq \epsilon s$, for some $\epsilon < 1/(\gamma + 1)$, and similarly for y .

Then, it holds that:

$$V \subseteq \bigcup_{x,y \in B^2, \mu_x < \mu_y} \{n \in \Lambda_{x,y} \mid \hat{\mu}_x > \mu_x + \frac{\delta}{2} \vee \hat{\mu}_y < \mu_y - \frac{\delta}{2}\} \quad (8)$$

$$\subseteq \bigcup_{x,y \in B^2, \mu_x < \mu_y} \{n \in \Lambda_{x,y} \mid \hat{\mu}_x > \mu_x + \frac{\delta}{2}\} \cup \{n \in \Lambda_{x,y} \mid \hat{\mu}_y < \mu_y - \frac{\delta}{2}\}, \quad (9)$$

where $\delta = \min_{a,b \in E, a \neq b} |\mu_a - \mu_b|$. It follows from Lemma A.1 that $|V| \leq D^2 \frac{8}{\epsilon \delta^2} = \mathcal{O}(1)$.

We below bound the number of iterations at which a suboptimal arm B is the leader, given that $B_2 \in \mathcal{N}_{B, \hat{\mu}(t)}$.

$\mathbb{E}[A_\epsilon] < \infty$ for any T Let $n \in A_\epsilon$, and let $s \triangleq l_B(n)$. By design of Alg. 3, $t_B(n) \geq s/(\gamma + 1)$, and by definition of A_ϵ , $t_{B_2}(n) \geq \epsilon l_B(n) = \epsilon s$. In light of those observations, the following applies Lemma A.2. For any s , let $\Lambda(s) = \{n \in A_\epsilon \mid l_B(n) = s\}$, and $\Lambda = \cup_{s \geq 1} \Lambda(s)$.

- As any n in A_ϵ is required to verify $L(n) = B$, and since $l_B(\cdot)$ increases each time B is leader, $|\Lambda(s)| \leq 1$ (the time step such that B is leader for the s -th time is unique).
- At any $p \in A_\epsilon$, either B was already leader before, either a leader change occurred and led to B . In case of a leader change, a greedy algorithm was performed (Line 8) and $\hat{\mu}_B$ was therefore better than $\hat{\mu}_{B_2}$. In case of B staying leader, $\hat{\mu}_B$ was better than $\hat{\mu}_{B_2}$ (Line 7). It therefore holds that $p \in A_\epsilon \implies p \in \{k \in \Lambda, \hat{\mu}_B(k) \geq \hat{\mu}_{B_2}(k)\}$, meaning that $A_\epsilon \subset \{k \in \Lambda, \hat{\mu}_B(k) \geq \hat{\mu}_{B_2}(k)\}$. Using the fact that $B_2 = B - x^* + y^*$, we have $A_\epsilon \subseteq \{p \in \Lambda, \hat{\mu}_{x^*}(p) \geq \hat{\mu}_{y^*}(p)\}$.
- Let $n \in \Lambda(s)$, for some s . By definition, $l_B(n) = s$. MAUB forces the leader to be played at least $\lfloor 1/(\gamma + 1) \rfloor \cdot s$ times, so that $t_B(n) \geq \lfloor 1/(\gamma + 1) \rfloor \cdot s$. Picking² epsilon sufficiently small with respect to the fixed γ constant, it holds that $t_B(n) \geq \epsilon \cdot s$. It follows that $t_{x^*}(n) \geq t_B(n) \geq \epsilon \cdot s$.
- Let $n \in \Lambda(s) \subset A_\epsilon$, for some s . By definition of A_ϵ , n is such that $t_{B_2}(n) \geq \epsilon l_B(n) = \epsilon \cdot s$. It follows that $t_{y^*}(n) \geq t_{B_2}(n) = \epsilon \cdot s$.
- Let $n \in \Lambda(s)$. By assumption on the bandit problem, $\mathbb{E}[\hat{\mu}_{B_2}] - \mathbb{E}[\hat{\mu}_B] = \mathbb{E}[\hat{\mu}_{y^*}] - \mathbb{E}[\hat{\mu}_{x^*}] = \mu_{y^*} - \mu_{x^*} \geq \delta$ (with $\delta = \min_{a,b \in E, a \neq b} |\mu_a - \mu_b|$).
- Finally, by design of MAUB, the event $\{n \in \Lambda(s)\}$ clearly is measurable w.r.t. $(X_e(t))_{e \in E, 1 \leq t \leq T}$.

We thus apply Lem. A.1, and get $\mathbb{E}[|A_\epsilon|] < \infty$, for any ϵ, T .

$\mathbb{E}[B_\epsilon^T] < \infty$ for any T Let:

- $C_\delta \triangleq \{n \notin V \mid L(n) = B, \exists e \in B, |\hat{\mu}_e(n) - \mu_e| > \delta\}$ be the set of instants at which B is badly estimated on at least one element e ;
- $D_\delta \triangleq \bigcup_{J \in \mathcal{N}_{B, \mu} \setminus \{B_2\}} D_{\delta, J}$, where $\forall J, D_{\delta, J} \triangleq \{n \notin V \mid L(n) = B, P(n) = J, \exists w \in J, |\hat{\mu}_w(n) - \mu_w| > \delta\}$ is the set of instants at which B is the leader, J is chosen to be played while being badly estimated on at least one element w ;

²Later, another constraint shall be put on the choice of ϵ , but both are coherent.

- $E^T \triangleq \{n \leq T, n \notin V \mid L(n) = B, \exists e \in B_2, \bar{\mu}_e(n) \leq \mu_e\}$ be the set of instants at which B is the leader and the optimistic statistic $\bar{\mu}$ underestimates at least one value μ_e for $e \in B_2$.

The first step to bound B_ϵ^T is to show that for a properly chosen ϵ , $|B_\epsilon^T| = 2\gamma(1+\gamma) \cdot [|C_\delta| + |D_\delta| + |E^T|] + \mathcal{O}_{T \rightarrow \infty}(1)$. Let $n \in B_\epsilon^T$ be a time step, let B be the leader at n , and let B_2 be the best arm in $\mathcal{N}_{B,\mu}$. It holds that

$$l_B(n) = t_{B,B_2} + \sum_{J \in \mathcal{N}_{B,\mu} \cup \{B\}} t_{B,J}(n). \quad (10)$$

In other words, the number of time that B was leader equals the sum, for all J in $\mathcal{N}_{B,\mu}$, of the number of times J was chosen while B was the leader. Some information are known for B_2 and B : $t_{B_2}(n) \leq \epsilon l_B(n)$ as $n \in B_\epsilon^T$, and $t_B(n) \geq l_B(n)/(\gamma+1)$. It follows that

$$(1 - \epsilon)l_B(n) \leq \sum_{J \in \mathcal{N}_{B,\mu} \cup \{B\} \setminus \{B_2\}} t_{B,J}(n). \quad (11)$$

Let $\epsilon < \frac{1}{2 \cdot (\gamma+1)}$ ³⁴. We show that it is impossible that both:

- $\forall J \in \mathcal{N}_{B,\mu} \setminus \{B_2\}, t_{B,J} < l_B(n)/(\gamma+1)$; and
- $t_{B,B} < (3/2)l_B(n)/(\gamma+1) + 1$

hold. Assume that both hold. Then, we show that the right-hand side of Eq.11 can be strictly upper-bounded by $(1 - \epsilon)l_B(n)$, which creates an absurdity.

$$\sum_{J \in \mathcal{N}_{B,\mu} \setminus \{B_2\}} t_{B,J}(n) < (3/2) \frac{l_B(n)}{(\gamma+1)} + 1 + \sum_{J \in \mathcal{N}_{B,\mu} \setminus \{B_2\}} t_{B,J}(n) \quad (12)$$

$$< (3/2) \frac{l_B(n)}{(\gamma+1)} + 1 + (\gamma-1) \cdot \frac{l_B(n)}{(\gamma+1)} \quad (\text{since } \gamma = |\mathcal{N}_{B,\mu}|) \quad (13)$$

$$\leq \frac{l_B(n)}{(\gamma+1)} \cdot [(3/2) + (\gamma-1)] \quad (14)$$

$$= \frac{l_B(n)}{2(\gamma+1)} \cdot [2 \cdot \gamma + 1]. \quad (15)$$

By assumption on ϵ , it also holds that:

$$(1 - \epsilon)l_B(n) > (1 - \frac{1}{2(\gamma+1)})l_B(n) \quad (16)$$

$$= (\frac{2\gamma+1}{2(\gamma+1)})l_B(n). \quad (17)$$

We conclude that $\sum_{J \in \mathcal{N}_{B,\mu} \setminus \{B_2\}} t_{B,J}(n) < (1 - \epsilon)l_B(n)$, which is absurd, given Equation (11).

³Note that this choice is coherent with the previous constraint on ϵ ($\epsilon < 1/(\gamma+1)$)

⁴The following aims at knowing either that B was played enough for it to be unlikely that it was chosen over B_2 (while B_2 is not overestimated, else, $n \in E^T$ holds!) or another J was played enough for it to be unlikely that J was picked over B_2 . B and J must be treated differently as $t_{B,B}(n) \geq s/(\gamma+1)$ by design of MAUB.

(a) Assume that $\exists J \in \mathcal{N}_{B,\mu} \setminus \{B_2\}$, $t_{B,J}(n) \geq l_B(n)/(\gamma + 1)$. Let J be such neighbor. Let $\phi(n)$ be the unique time step such that $L(\phi(n)) = B$, $P(\phi(n)) = J$, and $t_{B,J}(\phi(n)) = \lfloor l_B(n)/2 \cdot (\gamma + 1) \rfloor$. $\phi(n) < n$ exists as (i) $t_{B,J}$ is only incremented when B is the leader and J is selected; and (ii) $n \mapsto l_B(n)$ is increasing. The unicity of ϕ comes from $t_{B,J}(\cdot)$ being increased at $\phi(n)$. In the following, we assume $\phi(n) \notin C_\delta \cup D_\delta \cup E^T$. Let $(x_{B_2}^*, y_{B_2}^*, x_J, y_J) \in [B \times B_2] \times [B \times J]$ be uniquely defined by:

$$B_2 = B - x_{B_2}^* + y_{B_2}^*, \text{ and} \quad (18)$$

$$J = B - x_J + y_J. \quad (19)$$

The comparison between J and B_2 boils down to a study of those four elements at $\phi(n)$. Note that there is no reason for B_2 to be in $\mathcal{N}_{J,\mu}$, and conversely. Yet, $\phi(n)$ not belonging to $C_\delta \cup D_\delta \cup E^T$ implies that (i) x_J, y_J , and x_{B_2} are correctly estimated up to the constant δ , and (ii) $\bar{\mu}_e > \mu_e$ for all e in B_2 . By definition,

$$\bar{\mu}_J(\phi(n)) = \sum_{e \in J} \hat{\mu}_e(\phi(n)) + \sum_{e \in J} \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_e(\phi(n))}}, \quad (20)$$

so that

$$\bar{\mu}_J(\phi(n)) - \bar{\mu}_{B_2}(\phi(n)) \quad (21)$$

$$= -\hat{\mu}_{x_J}(\phi(n)) + \hat{\mu}_{y_J}(\phi(n)) + \hat{\mu}_{x_{B_2}^*}(\phi(n)) - \hat{\mu}_{y_{B_2}^*}(\phi(n)) \quad (22)$$

$$- \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_{B_2}^*}(\phi(n))}} - \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_{B_2}^*}(\phi(n))}} \quad (23)$$

$$< -(\mu_{x_J} - \delta) + \mu_{y_J} + \delta + \mu_{x_{B_2}^*} + \delta - \hat{\mu}_{y_{B_2}^*}(\phi(n)) \quad (24)$$

$$- \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_{B_2}^*}(\phi(n))}} - \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_{B_2}^*}(\phi(n))}} \quad (25)$$

$$= -\mu_{x_J} + \mu_{y_J} + 3\delta + \mu_{x_{B_2}^*} - \hat{\mu}_{y_{B_2}^*}(\phi(n)) \quad (26)$$

$$- \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_{B_2}^*}(\phi(n))}} - \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_{B_2}^*}(\phi(n))}} \quad (27)$$

(\downarrow since $\bar{\mu}_{y_{B_2}^*} > \mu_{y_{B_2}^*}$ by assumption on ϕ)

$$(28)$$

$$< \mu_{x_J} - \mu_{y_J} + \mu_{x_{B_2}^*} + 3\delta - \mu_{y_{B_2}^*} \quad (29)$$

$$- \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_{B_2}^*}(\phi(n))}} \quad (30)$$

$$< -\mu_{x_J} + \mu_{y_J} + \mu_{x_{B_2}^*} + 3\delta - \mu_{y_{B_2}^*} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_{B_2}^*}(\phi(n))}} \quad (31)$$

$$= \mu_{y_J} - \mu_{x_J} + \mu_{x_{B_2}^*} - \mu_{y_{B_2}^*} + 3\delta + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_J}(\phi(n))}} + \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_{B_2}^*}(\phi(n))}} \quad (32)$$

$$(33)$$

But now, since $\mu_{B_2} > \mu_J$, one can pick δ such that $\delta < ([\mu_{y_{B_2}^*} - \mu_{x_{B_2}^*}] - [\mu_{y_J} - \mu_{x_J}])/5$.

By design of MAUB, and by assumption (a), it respectively holds that $t_{x_{B_2}^*}(\phi(n)) \geq l_B(\phi(n))/(\gamma + 1) \geq l_B(\phi(n))/2(\gamma + 1)$, and $t_{y_J}(\phi(n)) \geq t_J(\phi(n)) \geq t_{B,J}(\phi(n)) \geq l_B(\phi(n))/2(\gamma + 1)$, we have:

$$\sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{x_{B_2}^*}}} < \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{l_B(n)/2(\gamma + 1)}} < \sqrt{\frac{2 \cdot \log(l_B(n))}{l_B(n)/2(\gamma + 1)}}, \text{ and} \quad (34)$$

$$\sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{t_{y_J}^*}} < \sqrt{\frac{2 \cdot \log(l_B(\phi(n)))}{l_B(n)/2(\gamma + 1)}} < \sqrt{\frac{2 \cdot \log(l_B(n))}{l_B(n)/2(\gamma + 1)}}. \quad (35)$$

But then, $n \mapsto \sqrt{\frac{2 \cdot \log(l_B(n))}{l_B(n)/2(\gamma + 1)}}$ converges to 0 as n goes to ∞ . Therefore, there must exist l_0 such that $l_B(n) \geq l_0$ implies that $\sqrt{\frac{2 \cdot \log(l_B(n))}{l_B(n)/2(\gamma + 1)}} < \delta$.

(b) Now assume that $t_{B,B} \geq (3/2)l_B(n)/(\gamma + 1) + 1$. B and B_2 can directly be compared through $x_{B_2}^*$ and $y_{B_2}^*$. There are at least $l_B(n)/2(\gamma + 1) + 1$ instants \tilde{n} such that B was selected normally (*i.e.* not forced). By the same reasoning as in (a), there exist a $\phi(n)$ such that $L(\phi(n)) = B$, $P(\phi(n)) = B$, $t_{B,B}(\phi(n)) = \lfloor l_B(n)/2(\gamma + 1) \rfloor$ and $(l_B(\phi(n)) - 1)$ is not a multiple of $1/(\gamma + 1)$. Thus, $\bar{\mu}_B(\phi(n)) \geq \bar{\mu}_{B_2}(\phi(n))$. Similar derivations as in (a), with $J = B$ produce an absurdity, which finally gives $\phi(n) \in C_\delta \cup D_\delta \cup E^T$.

Let us define $B_{\epsilon,l_0}^T = \{n \leq T \mid n \in B_\epsilon^T, l_B(n) \geq l_0\}$. We have $|B_{\epsilon,l_0}^T| \leq l_0 + |B_\epsilon^T|$. $\phi : n \mapsto \phi(n)$ is a mapping from B_{ϵ,l_0}^T to $C_\delta \cup D_\delta \cup E^T$. To bound the size of B_{ϵ,l_0}^T , we use the following decomposition:

$$\{n \mid n \in B_{\epsilon,l_0}^T, l_B(n) \geq l_0\} \subseteq \bigcup_{n' \in C_\delta \cup D_\delta \cup E^T} \{n \mid n \in B_{\epsilon,l_0}^T, \phi(n) = n'\}.$$

Let us fix n' . If $n \in B_{\epsilon,l_0}^T$ and $\phi(n) = n'$, then $\lfloor l_B(n)/2(\gamma + 1) \rfloor \in \bigcup_{J \in \mathcal{N}_{B,\mu} \setminus \{B_2\}} \{t_{B,J}(n')\}$ and $l_B(n)$ is incremented at time n because $L(n) = B$. Therefore:

$$|\{n \mid n \in B_{\epsilon,l_0}^T, \phi(n) = n'\}| \leq 2\gamma(\gamma + 1)$$

Using union bound, we obtain the desired result:

$$|B_\epsilon^T| \leq l_0 + |B_{\epsilon,l_0}^T| \leq \mathcal{O}(1) + 2\gamma(\gamma + 1)(|C_\delta| + |D_\delta| + |E^T|).$$

Bound on C_δ We apply Lem. A.1 with $\Lambda(s) \triangleq \{n \leq T \mid L(n) = B, l_B(n) = s\}$, and $\Lambda \triangleq \bigcup_{s \geq 1} \Lambda(s)$. It holds that:

$$C_\delta = \{n \in \bigcup_s \Lambda(s) \mid \exists x, |\hat{\mu}_x(n) - \mu_x| > \delta\} \quad (36)$$

$$\subseteq \bigcup_x \{n \in \bigcup_s \Lambda(s) \mid |\hat{\mu}_x(n) - \mu_x| > \delta\} \quad (37)$$

$$\triangleq \bigcup_x C_{\delta,x}. \quad (38)$$

Since each time B is leader, $l_B(n)$ increases, $\forall s$, $|\Lambda(s)| \leq 1$. By design of MAUB, it holds that $t_B(n) \geq 1/(\gamma + 1) \cdot l_B(n)$, so that $\forall s$, $t_B(s) > \epsilon \cdot s$ for some $\epsilon < 1/(\gamma + 1)$. Thus, for all $n, x \in C_\delta \cup B$, $t_x(n) \geq \epsilon s$. For all $x \in B$, Lem. A.1 gives $\mathbb{E}[|C_{\delta,x}|] = \mathbb{E}[\sum_n \mathbf{1}_{\{n \in \Lambda \mid |\hat{\mu}_x(n) - \mu_x| > \delta\}}] = \mathcal{O}(1)$. We thus get $\mathbb{E}[|C_\delta|] \leq \sum_{x \in B} \mathbb{E}[|C_{\delta,x}|] = \mathcal{O}(1)$.

Bound on D_δ For any $J \in \mathcal{N}_{B,\mu}$, for any $x \in J$ we apply Lem. A.1 with $\Lambda_{J,x}(s) \triangleq \{n \leq T | L(n) = B, P(n) = J, t_J(n) = s\}$. Then:

$$D_\delta \subseteq \cup_J \cup_x \{n \in \Lambda_{J,x}(s) \mid |\hat{\mu}_x - \mu_x| > \delta\} \quad (39)$$

Since each time B is leader and J is played, $t_J(n)$ increases, $\forall s, \forall x, |\Lambda_{J,x}(s)| \leq 1$. By definition of $\Lambda_{J,x}$, it holds that $\forall n \in \Lambda(s), t_x(n) = s > \epsilon \cdot s$, for some $\epsilon < 1$. For all J, x , Lem.A.1 gives $\mathbb{E}[\sum_n \{n \in \Lambda_{J,x}(s) \mid |\hat{\mu}_x - \mu_x| > \delta\}] = \mathcal{O}(1)$. Summing over the finite set J , and over the finite neighborhood $\mathcal{N}_{B,\mu}$, we finally get $\mathbb{E}[|D_\delta|] = \mathcal{O}(1)$.

Bound on E^T It holds that $E^T = \mathcal{O}(\log \log(T))$. For all x^* in B_2 , and for all t in \mathbb{N}^* , it holds that:

$$\mathbb{P}(\hat{\mu}_{x^*}(t) \leq \mu_{x^*} - \sqrt{\frac{2 \log(l_B(t))}{N_{x^*}(t)}}) \leq \exp(-2 \frac{N_{x^*}(t) 2 \log(l_B(t))}{N_{x^*}(t)}) \quad (\text{Hoeffding's inequality}) \quad (40)$$

$$= \exp(-4 \log(l_B(t))) \quad (41)$$

$$= l_B(t)^{-4}. \quad (42)$$

For all s , let $\Lambda(s) \triangleq \{n \leq T \mid l_B(n) = s, L(n) = B\}$ be the set of unique time step where B is the leader for the s -th time. We define $\Lambda \triangleq \cup_s \Lambda(s)$. For all s , let $\{\phi_s\} = \Lambda(s)$ if $\{1, \dots, T\} \cap \Lambda \neq \emptyset$ and $\phi_s = T + 1$ otherwise.

$$\mathbb{E}\left[\sum_{n=1}^T \mathbf{1}\{L(n) = B, \bar{\mu}_{x^*}(n) < \mu_{x^*}\}\right] \quad (43)$$

$$= \mathbb{E}\left[\sum_{n=1}^T \mathbf{1}\{n \in \Lambda, \bar{\mu}_{x^*}(n) < \mu_{x^*}\}\right] \quad (44)$$

$$\leq \mathbb{E}\left[\sum_{s \geq 1}^T \mathbf{1}\{\hat{\mu}_{x^*}(\phi_s) + \sqrt{\frac{2 \log(l_B(\phi_s))}{t_x(\phi_s)}} < \mu_{x^*}, \phi_s \leq T\}\right] \quad (45)$$

$$\leq \mathbb{E}\left[\sum_{s \geq 1} \mathbf{1}\{\hat{\mu}_{x^*}(\phi_s) + \sqrt{\frac{2 \log(s)}{t_x(\phi_s)}} < \mu_{x^*}\}\right] \quad (46)$$

$$(47)$$

All this put together gives $\mathbb{E}[B_\epsilon^T] = \mathcal{O}(\log \log(T))$, and adding the bound on $\mathbb{E}[A_\epsilon]$ yields $\mathbb{E}[l_B(T)] = \mathcal{O}(\log \log(T))$, which concludes the proof.

D Regret

Theorem 5.4. *Originally stated on page 7 The regret of MAUB up to time step T is*

$$\mathcal{O}\left(\sum_{e \notin B^*} \frac{8}{\mu_{\sigma(e)} - \mu_e} \log(T)\right) = \mathcal{O}\left(\frac{|E| - D}{\Delta_{\min}} \log T\right),$$

with $\Delta_{\min} \triangleq \min_{e \notin B^*} \mu_{\sigma(e)} - \mu_e$.

Proof. Given Th. 5.3, we restrict the study to the iterations for which B^* is the leader. MAUB behaves as a classical UCB on the finite neighborhood of B^* . We therefore lightly adapt [Auer et al., 2002]'s proof to our case. Let $B = B^* - \mu_{\sigma(e)} + \mu_e$.

Let us upper bound the number of times $t_B(n)$ that a suboptimal arm B was played up to time step n . Let $\ell \in \mathbb{N}^*$.

$$t_B(n) = \sum_{t=1}^n \mathbb{1}_{\{P(t)=B\}} \quad (48)$$

$$\leq \ell + \sum_{t=1}^n \mathbb{1}_{\{P(t)=B, t_B(t) \geq \ell\}}. \quad (49)$$

But now, for B to be played at time step t , it must hold that $\bar{\mu}_B(t) > \bar{\mu}_{B^*}(t)$, so that the event $\{P(t) = B\}$ is a subset of the event $\{\bar{\mu}_B(t) > \bar{\mu}_{B^*}(t)\}$. It follows that:

$$\mathbb{E}[t_B(n)] \leq \ell + \mathbb{E}\left[\sum_{t=1}^n \mathbb{1}_{\{\bar{\mu}_B(t) > \bar{\mu}_{B^*}(t), t_B(t) \geq \ell\}}\right] \quad (50)$$

$$\leq \ell + \sum_{t=1}^n \mathbb{P}\left(\hat{\mu}_e(t) + \sqrt{\frac{2 \log(l_{B^*}(t))}{N_e(t)}} > \hat{\mu}_{\sigma(e)}(t) + \sqrt{\frac{2 \log(l_{B^*}(t))}{N_{\sigma(e)}(t)}}, t_B(t) \geq \ell\right). \quad (51)$$

The event $\{\hat{\mu}_e(t) + \sqrt{\frac{2 \log(l_{B^*}(t))}{N_e(t)}} > \hat{\mu}_{\sigma(e)}(t) + \sqrt{\frac{2 \log(l_{B^*}(t))}{N_{\sigma(e)}(t)}}, t_B(t) \geq \ell\}$ is included in the event $\bigcup_{s=1}^n \bigcup_{s_i=\ell}^n \{\hat{\mu}_e + \sqrt{\frac{2 \log(l_{B^*}(t))}{s_i}} > \hat{\mu}_{\sigma(e)} + \sqrt{\frac{2 \log(l_{B^*}(t))}{s}}\}$. Now observe that $\{\hat{\mu}_e + \sqrt{\frac{2 \log(l_{B^*}(t))}{s_i}} > \hat{\mu}_{\sigma(e)} + \sqrt{\frac{2 \log(l_{B^*}(t))}{s}}\}$ implies that at least one of the following must hold:

$$\hat{\mu}_{\sigma(e)} \leq \mu_{\sigma(e)} - \sqrt{\frac{2 \log(l_{B^*}(t))}{s}} \quad (52)$$

$$\hat{\mu}_e \geq \mu_e + \sqrt{\frac{2 \log(l_{B^*}(t))}{s_i}} \quad (53)$$

$$\mu_{\sigma(e)} < \mu_e + 2\sqrt{\frac{2 \log(l_{B^*}(t))}{s_i}}. \quad (54)$$

For $\ell \geq \lfloor 8 \log(l_{B^*}(n)) / |\mu_{\sigma(e)} - \mu_e|^2 \rfloor + 1$, the last event can not happen.

Both other events are bounded as before (Equations 43 to 47) using Hoeffding's inequality. It follows that for all $B \in \mathcal{N}_{B^*, \mu}$, $t_B(n) = \mathcal{O}(8 \log(n) / |\mu_{\sigma(e)} - \mu_e|)$. Summing over $\mathcal{N}_{B^*, \mu}$ gives the overall regret $\mathcal{O}(\sum_{e \notin B^*} \frac{8 \cdot \log T}{\mu_{\sigma(e)} - \mu_e})$