Solving Bernoulli Rank-One Bandits with Unimodal Thompson Sampling

Cindy Trinh, Emilie Kaufmann, Claire Vernade, Richard Combes

Solving Bernoulli Rank-One Bandits with Unimodal Thompson Sampling

Cindy Trinh CINDY.TRINH.SRIDYKHAN@GMAIL.COM
ENS Paris-Saclay

Emilie Kaufmann EMILIE.KAUFMANN@UNIV-LILLE.FR
CNRS, Université de Lille, Inria SequeL

Claire Vernade VERNADE@GOOGLE.COM
DeepMind, London

Richard Combes RICHARD.COMBES@SUPELEC.FR
CentraleSupélec, Gif-sur-Yvette

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

Stochastic Rank-One Bandits Katariya et al. (2017a,b) are a simple framework for regret minimization problems over rank-one matrices of arms. The initially proposed algorithms are proved to have logarithmic regret, but do not match the existing lower bound for this problem. We close this gap by first proving that rank-one bandits are a particular instance of unimodal bandits, and then providing a new analysis of Unimodal Thompson Sampling (UTS), initially proposed by Paladino et al. (2017). We prove an asymptotically optimal regret bound on the frequentist regret of UTS and we support our claims with simulations showing the significant improvement of our method compared to the state-of-the-art.

Keywords: Multi-armed bandits, unimodal bandits, rank-one bandits.

1. Introduction

We consider *Stochastic Rank-One Bandits*, a class of bandit problems introduced by Katariya et al. (2017b). These models provide a clear framework for the exploration-exploitation problem of adaptively sampling the entries of a rank-one matrix in order to find the largest one. Consider for instance the problem of finding the best design of a display, say for example a colored shape to be used as a button on a website. One may have at hand a set of different shapes, and a set of different colors to be tested. A display is a combination of those two attributes, and a priori the tester has as many options as there are different pairs of shapes and colors. Now let us assume the effect of each factor is independent of the other factor. Then, the value of a combination, say for instance the click rate on the constructed button, is the product of the values of each of its attributes. The better the shape, the higher the rate, and similarly for the color. This type of independence assumptions is ubiquitous in click models such as the *position-based model* Chuklin et al. (2015); Richardson et al. (2007). It is also closely related to online learning to rank Zoghi et al. (2017) where sequential duels allow to find the optimal ordering of a list of options. We review the related literature in Section 4.

We formalize our example above into a Bernoulli rank-one bandit model (Katariya et al., 2017a): this model is parameterized by two nonzero vectors $\boldsymbol{u} = (u_1, u_2, ..u_K) \in [0,1]^K$ and $\boldsymbol{v} =$

$(v_1, v_2, ..v_L) \in [0,1]^L$. There are $K \times L$ arms, indexed by $(i,j) \in [K] \times [L]$, where we use the notation $[p] := \{1, \dots, p\}$ for any positive integer p. Each arm (i,j) is associated with a Bernoulli distribution with mean $\mu_{(i,j)} := u_i v_j$. Observe that the matrix of means $\boldsymbol{\mu} = \boldsymbol{u}\boldsymbol{v}^T$ has rank one, hence the name. We denote Θ the class of all such instances $(\boldsymbol{u} \times \boldsymbol{v})$. At each time step t an agent selects an arm $K(t) = (I(t), J(t)) \in [K] \times [L]$ and receives a reward $r(t) \sim \mathcal{B}(\mu_{(I(t),J(t))})$, independently from previous rewards. To select $K(t)$, the agent may exploit the knowledge of previous observations and possibly some external randomness $U(t)$. Denoting by $\mathcal{F}_t$ the σ-field generated by $K(1), r(1), K(2), r(2), \dots, K(t), r(t)$, $K(t)$ is measurable with respect to $\sigma(\mathcal{F}_{t-1}, U(t))$.

The objective of the learner is to adjust their selection strategy to maximize the expected total reward accumulated. The oracle or optimal strategy here is to always play the arm with largest mean. Thus, maximizing rewards is equivalent to designing a strategy $\mathcal{A}$ with small *regret*, where the T-step regret $R_{\boldsymbol{\mu}}(T, \mathcal{A})$ is defined as the difference between the expected cumulative rewards of the oracle and the cumulative rewards of the strategy $\mathcal{A}$:

$$R_{\boldsymbol{\mu}}(T, \mathcal{A}) = \sum_{t=1}^{T} \left[\max_{(i,j)\in[K]\times[L]} \mu_{(i,j)} - \mathbb{E}_{\boldsymbol{\mu}}[\mu_{(I(t),J(t))}] \right] . \tag{1}$$

Letting $i_\star = \text{argmax}_i u_i$ and $j_\star = \text{argmax}_j v_j$, we assume that $u_{i_\star} > u_i$ for all $i \neq i_\star$ and $v_{j_\star} > v_j$ for all $j \neq j_\star$. This assumption is equivalent to assuming that the rank-one bandit instance has a unique optimal action, which is $(i_\star, j_\star) = \text{argmax}_{(i,j)\in[K]\times[L]} \mu_{(i,j)}$. We let $\Theta_\star$ denote this class of rank-one instance with a unique optimal arm. In this paper, we will furthermore restrict our attention to rank-one models for which either $\boldsymbol{u} \succ 0$ or $\boldsymbol{v} \succ 0$. This assumption is not very restrictive, but it rules out the possibility that $u_i = 0$ and $v_j = 0$ for a certain arm (i,j) (i.e. neither shape i nor color j attract any user). We found this assumption to be necessary to exhibit a unimodal structure in rank-one bandits.

An algorithm is called *uniformly efficient* if its regret is sub-polynomial in any instance $(\boldsymbol{u}\times\boldsymbol{v}) \in \Theta$. That is, for all $\alpha > 0$, for all $(\boldsymbol{u} \times \boldsymbol{v}) \in \Theta$, $\mathcal{R}(T) = o(T^\alpha)$. In their paper, Katariya et al. (2017b) provide the first uniformly efficient algorithm, `Rank1Elim`, for stochastic rank-one bandits, and Katariya et al. (2017a) propose an adaptation of this algorithm tailored for Bernoulli rewards, `Rank1ElimKL`. They also provide a problem-dependent asymptotic lower bound on the regret in the line of Lai and Robbins (1985). This type of result gives a precise characterization of the regret for a specific instance of the problem that one should expect for any uniformly efficient algorithm. We report their result below.

Proposition 1 *For any algorithm $\mathcal{A}$ which is uniformly efficient and for any Bernoulli rank-one bandit problem, $(\boldsymbol{u} \times \boldsymbol{v}) \in \Theta_\star$,*

$$\liminf_{T\to\infty} \frac{R_{\boldsymbol{\mu}}(\mathcal{A}, T)}{\log(T)} \geq \sum_{i\in[K]\setminus i_\star} \frac{\mu_{i_\star,j_\star} - \mu_{i,j_\star}}{\text{kl}(\mu_{i,j_\star}, \mu_{i_\star,j_\star})} + \sum_{j\in[L]/j_\star} \frac{\mu_{i_\star,j_\star} - \mu_{i_\star,j}}{\text{kl}(\mu_{i_\star,j}, \mu_{i_\star,j_\star})}.$$

where $\text{kl}(x,y) = x\ln(x/y) + (1-x)\ln((1-x)/(1-y))$ is the binary relative entropy.

In contrast to this result, the Lai and Robbins (1985) lower bound, which applies to algorithms that are uniformly efficient for *any* reward matrix $\boldsymbol{\mu}$, involves a sum over all matrix entries $(i,j) \in [K] \times [L]$ instead of restricting to arms in the best row and in the best column of the matrix. Thus a good algorithm for the rank-one problem should manage to select all entries (i,j) that are not

in this best row and column only $o(\ln(T))$ times. However, neither `Rank1Elim` nor `Rank1ElimKL` achieve the asymptotic performance of Proposition 1: the regret upper bounds provided by Katariya et al. (2017b,a) show much larger constants, and the empirical performance is not much tighter. A natural question one might ask then is: Is that lower bound achievable ?

Contributions The main contribution of this paper is to close this existing gap. To do so, we notice and prove that a stochastic rank-one bandit satisfying $\boldsymbol{u} \succ 0$ or $\boldsymbol{v} \succ 0$ is a particular instance of *Unimodal Bandits* (Combes and Proutière, 2014). Interestingly, when derived in the specific rank-one bandits setting, the `OSUB` algorithm proposed in the latter reference achieves the optimal asymptotic regret of Proposition 1. Unifying those two apparently independent lines of work sheds a new light on stochastic rank-one bandits.

Indeed, follow-up works on unimodal bandits sought ways to construct more efficient algorithms than `OSUB`. In particular Paladino et al. (2017) propose `UTS`, a Bayesian strategy based on Thompson Sampling (Thompson, 1933). Unfortunately, the theoretical analysis they provide does not allow to conclude an upper bound on the performance of their algorithm. We shall comment on that in Section 2.3. Thus, a second major contribution of the present work is a new finite-time analysis of the frequentist regret of `UTS` for Bernoulli stochastic rank-one bandits. Doing so, we provide an optimal regret bound for an efficient and easy-to-implement rank-one bandit algorithm.

Finally, our analysis provides new insights on the calibration of the leader exploration parameter which is present in other algorithms.

Outline The paper is organised as follows. Section 2 proves that rank-one bandits are an instance of unimodal bandits, and describes the `UTS` algorithm. The regret upper bound is proved in Section 3. In order to perform a fair empirical comparison with existing rank-one bandit algorithms, we give more background on this literature in Section 4. Finally, experiments in Section 5 provide empirical evidence of the optimality of `UTS` and show an improvement of an order of magnitude compared to the state-of-the-art `Rank1ElimKL`.

2. Rank-One Bandits, a particular case of Unimodal Bandits

In this section, we explain why the rank-one bandit model can be seen as a graphical unimodal bandit model as introduced by Yu and Mannor (2011); Combes and Proutière (2014). For completeness, we recall the relevant definition.

Definition 2 *Given a undirected graph $G = (V, E)$, a vector $\boldsymbol{\mu} = (\mu_k)_{k\in V}$ is* unimodal with respect to G *if (i) there exists a unique $k_\star \in V$ such that $\mu_{k_\star} = \max_i \mu_i$ and (ii) from any $k \neq k_\star$, we can find an increasing path to the optimal arm: formally, $\forall k \neq k_\star$, there exists a path $p = (k_1 = k, k_2, ..., k_{m_k} = k_\star)$ of length m_k, such that for all $i = 1, ..., m_k - 1$, $(k_i, k_{i+1}) \in E$, and $\mu_{k_i} < \mu_{k_{i+1}}$. We denote by $\mathcal{U}(G)$ the set of vectors $\boldsymbol{\mu}$ that are unimodal with respect to G.*

A bandit instance is unimodal with respect to an undirected graph $G = (V, E)$ if its vector of means $\boldsymbol{\mu} = (\mu_k)_{k\in V}$ is unimodal with respect to G: $\boldsymbol{\mu} \in \mathcal{U}(G)$. For a unimodal instance, we define the set of neighbors of an arm $k \in V$ as $\mathcal{N}(k) = \{\ell : (k, \ell) \in E\}$. Without loss of generality, we can assume that E does not contain self-edges (k, k) (which do not contribute to increasing paths), therefore $k \notin \mathcal{N}(k)$. The extended neighborhood of k is defined as $\mathcal{N}^+(k) = \mathcal{N}(k) \cup \{k\}$. In a unimodal bandit problem, the learner knows the graph G (hence the neighborhoods $\mathcal{N}(k), \mathcal{N}^+(k)$ for all $k \in V$), but not its parameters $\boldsymbol{\mu}$.

2.1. Rank-One Bandits are Unimodal

We define the undirected graph $G_1 = (V, E_1)$ as the graph with vertices $V = \{1, \dots, K\} \times \{1, \dots, L\}$ and such that $((i,j),(k,\ell)) \in E_1$ if and only if $(i,j) \neq (k,\ell)$ and ($i = k$ or $j = \ell$). In words, viewing the vertices as a $K \times L$ matrix, two distinct entries are neighbors if they belong to the same line or to the same column. In particular it can be observed that the graph G_1 has diameter two, and we shall exhibit below increasing paths of length at most two between any sub-optimal arm (i,j) and the best arm $(i_\star, j_\star)$. The main result of this section is Proposition 3. It allows us to build on the existing results for unimodal bandits in order to close the remaining theoretical gap in the understanding of rank-one bandits.

Proposition 3 *Let $\boldsymbol{u} = (u_1, u_2, ..u_K)$ and $\boldsymbol{v} = (v_1, v_2, ..v_L)$ be two nonzero vectors such that $\boldsymbol{u} \succ 0$ or $\boldsymbol{v} \succ 0$. A rank-one bandit instance parameterized by $\boldsymbol{u}, \boldsymbol{v}$ satisfies $\boldsymbol{\mu} \in \mathcal{U}(G_1)$.*

Proof Let us denote the best arm by $k_\star = (i_\star, j_\star)$. Then for any $(i,j) \in V$ with $(i,j) \neq k_\star$, one can find several increasing paths in G_1 from (i,j) to $(i_\star, j_\star)$. If $i = i_\star$ or $i = j_\star$, then $(i,j) \to (i_\star, j_\star)$ is valid as $((i,j), k_\star) \in E_1$ and $\mu_{(i,j)} < \mu_{k_\star}$. Otherwise, either $v_j \neq 0$ or $u_i \neq 0$. In the first case $(i,j) \to (i_\star, j) \to (i_\star, j_\star)$ is a valid increasing path. Indeed, $u_i < u_{i_\star}$ and $0 < v_j < v_{j_\star}$ implies that $\mu_{(i,j)} = u_i v_j < u_{i_\star} v_j = \mu_{(i_\star, j)} < u_{i_\star} v_{j_\star} = \mu_{(i_\star, j_\star)}$. In the second case, one can similarly show that $(i,j) \to (i, j_\star) \to (i_\star, j_\star)$ is a valid increasing path.

Figure 1 below illustrates a possible optimal path in a rank-one bandit with $K = L = 4$ and also shows the neighbors of a particular arm in the graph G_1.

$$\begin{bmatrix} (u_1v_1) & (u_1v_2) & \boldsymbol{(u_1v_3)} & (u_1v_4) \\ (u_2v_1) & (u_2v_2) & \boldsymbol{(u_2v_3)} & (u_2v_4) \\ \boldsymbol{(u_3v_1)} & \boldsymbol{(u_3v_2)} & \boxed{(u_3v_3)} & \boldsymbol{(u_3v_4)} \\ (u_4v_1) & (u_4v_2) & \boldsymbol{(u_4v_3)} & (u_4v_4) \end{bmatrix} \qquad \begin{bmatrix} \boldsymbol{(u_1v_1)} & (u_1v_2) & \boldsymbol{(u_1v_3)} & (u_1v_4) \\ (u_2v_1) & (u_2v_2) & (u_2v_3) & (u_2v_4) \\ (u_3v_1) & (u_3v_2) & \boldsymbol{(u_3v_3)} & (u_3v_4) \\ (u_4v_1) & (u_4v_2) & (u_4v_3) & (u_4v_4) \end{bmatrix}$$

Figure 1: $\mathcal{N}((3,3))$ in bold (left). Increasing path from $(3,3)$ to $(i_\star = 1, j_\star = 1)$ (right).

2.2. Solving Unimodal Bandits

In their initial paper, Yu and Mannor (2011) propose an algorithm based on sequential elimination that does not efficiently exploit the graph structure. Combes and Proutière (2014) tackle the unimodal bandit problem and provide an analysis of the achievable regret in that setting. Their Theorem 4.1 states an asymptotic regret lower bound that we recall below for Bernoulli rewards.

Proposition 4 *Let $G = (V, E)$ define a Bernoulli unimodal bandit problem, with $\mathcal{N}_G(k) = \{\ell : (k,\ell) \in E\}$ denoting the set of neighbors of arm $k \in V$. Let $\mathcal{A}$ be a uniformly efficient algorithm for every Bernoulli bandit instance with means in $\mathcal{U}(G)$. Then*

$$\forall \boldsymbol{\mu} \in \mathcal{U}(G), \quad \liminf_{T \to \infty} \frac{R_{\boldsymbol{\mu}}(\mathcal{A}, T)}{\ln(T)} \geq \sum_{k \in \mathcal{N}_G(k_\star)} \frac{\mu_{k_\star} - \mu_k}{\mathrm{kl}\left(\mu_k, \mu_{k_\star}\right)}.$$

In the particular case $G = G_1$, $\mathcal{N}_{G_1}((i_\star, j_\star)) = \{(i,j) : i = i_\star \text{ or } j = j_\star\} \backslash \{(i_\star, j_\star)\}$ we recover Proposition 1. An asymptotically optimal algorithm for unimodal bandits therefore particularizes into an asymptotically optimal algorithm for rank-one bandits.

2.3. Candidate algorithms and their analysis

There exists only a few optimal algorithms for unimodal bandits. Combes and Proutière (2014) propose OSUB, a computationally efficient algorithm that is proved to have the best achievable regret. Paladino et al. (2017) propose a Bayesian alternative, however for reasons detailed below we believe their regret analysis does not hold as is. Another valid algorithm would be OSSB (Combes et al., 2017), a generic method for structured bandits, however its implementation for rank-one bandits is not obvious (the matrix of empirical mean $\hat{\boldsymbol{\mu}}(t)$ would need to have rank one), and its generality often makes it less empirically efficient when compared to algorithms exploiting a particular structure, like here the rank-one structure.

Notation We now present the existing algorithms for unimodal bandits with respect to some undirected graph $G = (V, E)$. For $k \in V$, we let $N_k(t) = \sum_{s=1}^{t} \mathbb{1}_{(K(s)=k)}$ be the number of selections of arm k up to round t and $\hat{\mu}_k(t) = \frac{1}{N_k(t)} \sum_{s=1}^{t} X_s \mathbb{1}_{(K(s)=k)}$ be its empirical mean of rewards. We define the (empirical) leader $L(t) = \operatorname{argmax}_{k \in V} \hat{\mu}_k(t)$ and denote by $\ell_k(t) = \sum_{s=1}^{t} \mathbb{1}_{(L(s)=k)}$ the number of times arm k has been leader up to round t.

Optimal Sampling for Unimodal Bandits (OSUB) OSUB (Combes and Proutière, 2014) is the adaptation of the kl-UCB algorithm of Cappé et al. (2013), an asymptotically optimal algorithm for (unstructured) Bernoulli bandits. The vanilla kl-UCB algorithm uses as upper confidence bounds the indices $\mathrm{u}_k(t) = \max\{q : N_k(t)\mathrm{kl}(\hat{\mu}_k(t), q) \leq f(t)\}$, and selects at each round the arm with largest index. The idea of OSUB is to restrict kl-UCB to the neighborhood of the leader while adding a *leader exploration mechanism* to ensure that the leader gets "checked" enough and can eventually be trusted. Letting $\widetilde{\mathrm{u}}_k(t) = \max\left\{q : N_k(t)\mathrm{kl}\left(\hat{\mu}_k(t), q\right) \leq f(\ell_{L(t)}(t))\right\}$, OSUB selects at time $t+1$

$$A_{t+1} = \begin{cases} L(t) & \text{if } \ell_{L(t)}(t) \equiv 1[\gamma], \\ \operatorname*{argmax}_k \widetilde{\mathrm{u}}_k(t) & \text{else.} \end{cases} \tag{2}$$

The parameter γ quantifies how often the leader should be checked. OSUB is proved to be asymptotically optimal when γ is equal to the maximal degree in $G + 1$, which yields $\gamma = K + L - 1$ for rank-one bandits. Compared to kl-UCB, the alternative exploration rate $f(\ell_{L(t)}(t))$ that appears in the index $\widetilde{\mathrm{u}}_k(t)$ makes the analysis of OSUB quite intricate.

Unimodal Thompson Sampling (UTS) For classical bandits, Thompson Sampling (TS) is known to be a good alternative to kl-UCB as it shares its optimality property for Bernoulli distributions (Kaufmann et al., 2012; Agrawal and Goyal, 2013) without the need to tune any confidence interval and often with better performance. Paladino et al. (2017) therefore naturally proposed Unimodal Thompson Sampling (UTS). The algorithm, described in detail in Section 3, consists in running Thompson Sampling instead of kl-UCB in the neighborhood of the leader, while keeping a leader exploration mechanism similar to the one in (2). The exploration parameter γ should also be set to $K + L - 1$ in the rank-one case in order to prove the asymptotic optimality of UTS.

The analysis proposed by Paladino et al. (2017) (detailed in Appendix A of the extended version Paladino et al. (2016)) hinges on adapting some elements of the Thompson Sampling proof of Kaufmann et al. (2012) and is not completely satisfying. Our main objection is the upper bound that is proposed on the number of times a sub-optimal arm k is the leader (term $\mathcal{R}_2$ of the second equation on page 8). To deal with this term, a quite imprecise reduction argument is given (definition of $\hat{L}_{k,t}$) showing that one essentially needs to control the quantity $\sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_k(t) \geq \hat{\mu}_{k_2}(t)\right)$ for Thompson

Sampling playing in $\mathcal{N}(k)$ and k_2 being the element with largest mean in this neighborhood. However, we do not believe this quantity can be easily controlled for Thompson Sampling, as we have to handle a random number of observations (that may be small) from both k and $k_\star$. Besides, the upper bound on $\mathcal{R}_2$ proposed by Paladino et al. (2017) holds for the choice $\gamma = K + L - 1$ in the rank-one case, which we show is unnecessary.

Due to the lack of accuracy of the existing proof, we believe that a new, precise analysis of Unimodal Thompson Sampling is needed to corroborate its good empirical performance for rank-one bandits. Our analysis borrows elements from both the TS analysis of Agrawal and Goyal (2013) and that of Kaufmann et al. (2012). It also reveals that unlike what was previously believed, the leader exploration parameter can be set to an arbitrary value $\gamma \geq 2$.

3. Analysis of Unimodal Thompson Sampling

In this section, we present the *Unimodal Thompson Sampling* algorithm (UTS) for Bernoulli rank-one bandits, and we state our main theorem proving a problem-dependent regret upper bound for this algorithm, which extends to the graphical unimodal case.

3.1. UTS for Rank-One Bandits

UTS is a very simple computationally efficient, anytime algorithm. Its pseudo-code for Bernoulli rank-one bandits is given in Algorithm 1. It relies on one integer parameter $\gamma \geq 2$ controlling the fraction of rounds spent exploring the leader. After an initialization phase where each entry is pulled once, at each round $t > K \times L$, the algorithm computes the leader $L(t)$, that is the empirical best entry in the matrix. If the number of times $L(t)$ has been leader is multiple of γ, UTS selects the empirical leader. The rest of the time, it draws a posterior sample for every entry in the same row and column as the leader, and selects the entry associated to the largest posterior sample. Any tie is broken at random. This can be viewed as performing Thompson Sampling in $\mathcal{N}^+_{G_1}(L(t))$, the augmented neighborhood of the leader in the graph G_1 defined in Section 2.

3.2. Regret upper bound and asymptotic optimality

UTS can be easily extended to any graphical unimodal bandit problem with respect to a graph G, by performing Thompson Sampling on $\mathcal{N}^+_G$ instead of $\mathcal{N}^+_{G_1}$. For this more general algorithm, we state the following theorem, which is our main technical contribution.

Theorem 5 *Let $\boldsymbol{\mu}$ be a unimodal bandit instance with respect to a graph G. For all $\gamma \geq 2$, UTS with parameter γ satisfies, for every $\varepsilon > 0$,*

$$\mathcal{R}_{\boldsymbol{\mu}}(T, \texttt{UTS}(\gamma)) \leq (1+\varepsilon) \sum_{k \in \mathcal{N}(k_\star)} \frac{(\mu_\star - \mu_k)}{\mathrm{kl}(\mu_k, \mu_\star)} \ln(T) + C(\boldsymbol{\mu}, \gamma, \varepsilon),$$

where $C(\boldsymbol{\mu}, \gamma, \varepsilon)$ is some constant depending on the environment $\boldsymbol{\mu}$, on ε and on γ.

As a consequence $\limsup_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(T,\texttt{UTS}(\gamma))}{\ln(T)} \leq \sum_{k\in\mathcal{N}(k_\star)} \frac{(\mu_\star-\mu_k)}{\mathrm{kl}(\mu_k,\mu_\star)}$ for every parameter $\gamma \geq 2$. Therefore, $\texttt{UTS}(\gamma)$ is asymptotically optimal for any unimodal bandit problem. Particularizing this result to rank-one bandits, one obtains that Algorithm 1 has a regret which is asymptotically matching the lower bound in Proposition 1.

Algorithm 1 UTS for Bernoulli rank-one bandits

Input: $\gamma \in \mathbb{N}, \gamma \geq 2$.
  for $(i,j) \in [K] \times [L]$ **do**
    $N_{(i,j)} = 1$. $L_{(i,j)} = 0$.
    Draw arm (i,j), receive reward R and let $S_{(i,j)} = R$.
  end
  for $t = KL+1, \dots, T$ **do**
    Compute the entry-wise empirical leader $L(t) = (I_L(t), J_L(t)) = \underset{(i,j)\in[K]\times[L]}{\operatorname{argmax}} \hat{\mu}_{i,j}(t)$;
    Update the leader count $L_{L(t)} \leftarrow L_{L(t)} + 1$
    if $L_{L(t)} \equiv 0[\gamma]$ **then**
      $(I(t), J(t)) = L(t)$
    else
      for $k \in \{(I_L(t), j) : j \in [L]\} \bigcup \{(i, J_L(t)) : i \in [K]\}$ **do**
        $\theta_k \sim \text{Beta}\left(S_k + 1, N_k - S_k + 1\right)$
      end
      $(I(t), J(t)) = \underset{k}{\operatorname{argmax}}\, \theta_k$.
    end
    Receive reward $R_t \sim \mathcal{B}(\mu_{(I_t,J_t)})$
    $N_{(I(t),J(t))} \leftarrow N_{(I(t),J(t))} + 1$. $S_{(I(t),J(t))} \leftarrow S_{(I(t),J(t))} + R_t$
  end

Unlike previous work, in which logarithmic regret is proved only for the choice $\gamma = K + L - 1$ in the rank-one case[1], we emphasize that this result holds for any choice of the leader exploration parameter. We conjecture that UTS without any leader exploration scheme is also asymptotically optimal. However, our experiments of Section 5 reveal that this particular kind of "forced exploration" is not hurting for rank-one bandits, and that the choice $\gamma = 2$ actually leads to the best empirical performance.

3.3. Proof of Theorem 5

We consider a general K-armed graphical unimodal bandit problem with respect to some graph G and let $K(t)$ denote the arm selected at round t. We recall some important notations defined in Section 2.3: the number of arms selections $N_k(t)$, the empirical means $\hat{\mu}_k(t)$, the leader as $L(t) = \operatorname{argmax}_k \hat{\mu}_k(t)$, and the number of times arm k has been the leader up to time t: $\ell_k(t) = \sum_{s=1}^{t} \mathbb{1}\left(L(s) = k\right)$. Observe that the leader exploration scheme ensures that $\forall k \in \{1, \dots, K\}, \forall t \in \mathbb{N}, N_k(t) \geq \lfloor \ell_k(t)/\gamma \rfloor$.

Introducing the gap $\Delta_k = \mu_\star - \mu_k$, recall that the regret rewrites $\sum_{k \neq k_\star} \Delta_k \mathbb{E}_{\boldsymbol{\mu}}[N_k(T)]$. Just like in the analysis of Combes and Proutière (2014); Paladino et al. (2017), we start by distinguishing

1. For general unimodal bandits, OSUB sets γ to be the maximal degree of an arm, whereas UTS adaptively sets γ to be the degree of the current leader. Both parameterization coincide for rank-one bandits.

the times when the leader is the optimal arm and when it is not:

$$\mathcal{R}_{\boldsymbol{\mu}}(T, \text{UTS}(\gamma)) = \sum_{k \neq k_\star} \Delta_k \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k)\right]$$
$$= \underbrace{\sum_{k \neq k_\star} \Delta_k \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k, L(t) = k_\star)\right]}_{\mathcal{R}_1(T)} + \underbrace{\sum_{k \neq k_\star} \Delta_k \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k, L(t) \neq k_\star)\right]}_{\mathcal{R}_2(T)} .$$

To upper bound $\mathcal{R}_1(T)$, it can be noted that when $k_\star$ is the leader, the selected arm k is necessarily in the neighborhood of $k_\star$, hence the sum can be restricted to the neighborhood of $k_\star$. Therefore, we expect to upper bound $\mathcal{R}_1(T)$ by the same quantity which upper bounds the regret of Thompson Sampling restricted to $\mathcal{N}^+(k_\star)$. Such an argument is used for KL-UCB and Thompson Sampling by Combes and Proutière (2014) and Paladino et al. (2017) respectively, without much justification. However, a proper justification does need some care, as between two times the leader is $k_\star$, UTS may update the posterior of some arms in $\mathcal{N}^+(k_\star)$ for they belong to the neighborhoods of other potential leaders.

In this work, we carefully adapt the analysis Agrawal and Goyal (2013) to get the following upper bound. The proof can be found in Appendix B.

Lemma 6 *For all $\varepsilon > 0$ and all $T \geq 1$,*

$$\mathcal{R}_1(T) \leq (1+\varepsilon) \sum_{k \in N(k_*)} \frac{\Delta_k}{\text{kl}(\mu_k, \mu_\star)} \ln(T) + \tilde{C}(\boldsymbol{\mu}, \varepsilon),$$

for some quantity $\tilde{C}(\boldsymbol{\mu}, \varepsilon)$ which depends on the means $\boldsymbol{\mu}$ and on ε but not on T.

We now upper bound $\mathcal{R}_2(T)$, which can be related to the probability of choosing any given suboptimal arm k as the leader:

$$\begin{aligned}
\mathcal{R}_2(T) &\leq \sum_{\ell \neq k_\star} \sum_{k \neq k_\star} \Delta_k \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k, L(t) = \ell)\right] \\
&\leq \sum_{\ell \neq k_\star} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(L(t) = \ell) \sum_{k \neq k_\star} \mathbb{1}(K(t) = k)\right] = \sum_{k \neq k_\star} \sum_{t=1}^{T} \mathbb{P}\left(L(t) = k\right).
\end{aligned}$$

For each $k \neq k_\star$, we define the set of best neighbors of k, $\mathcal{B}_{\mathcal{N}(k)} = \text{argmax}_{\ell \in \mathcal{N}(k)} \mu_\ell$. Due to the unimodal structure, we know this set is nonempty because there exists at least one arm $\ell \in \mathcal{N}(k)$ such that $\mu_\ell > \mu_k$ (such an arm belongs to the path from k to $k_\star$). All arms belonging to $\mathcal{B}_{\mathcal{N}(k)}$ have same mean, that we note $\mu_{k_2} = \max_{\ell \in \mathcal{N}(k)} \mu_\ell$. We also introduce $\tilde{B} = \max_{k \in [K] \setminus \{k_\star\}} |\mathcal{B}_{\mathcal{N}(k)}|$, the maximal number of best arms in the neighborhood of all sub-optimal arms, which is bounded by

the maximum degree of the graph. With these notations, one can write, for any $b \in (0,1)$,

$$\sum_{t=1}^{T} \mathbb{P}\left(L(t)=k\right) = \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(L(t)=k, \exists k_2 \in \mathcal{B}_{\mathcal{N}(k)}, N_{k_2}(t) > (\ell_k(t))^b\right)}_{\mathcal{T}_1^k(T)} + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(L(t)=k, \forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, N_{k_2}(t) \leq (\ell_k(t))^b\right)}_{\mathcal{T}_2^k(T)}$$

The first term can be easily upper bounded by using the fact that if both arm k and one of its best neighbors $k_2 \in \mathcal{B}_{\mathcal{N}(k)}$ are selected enough, it is unlikely that $\hat{\mu}_k(t) \geq \hat{\mu}_{k_2}(t)$.

On the event $\{L(t)=k\}$, the empirical mean of the k-th arm is necessarily greater than that of the other arms (especially those in $\mathcal{B}_{\mathcal{N}(k)}$) . Therefore, letting $\delta_k = \frac{\mu_{k_2}-\mu_k}{2}$,

$$\begin{aligned}
\mathcal{T}_1^k(T) &= \sum_{t=1}^{T} \mathbb{P}\left(L(t)=k, \exists k_2 \in \mathcal{B}_{\mathcal{N}(k)}, \hat{\mu}_k(t) \geq \hat{\mu}_{k_2}(t), N_{k_2}(t) > (\ell_k(t))^b\right) \\
&\leq \sum_{t=1}^{T} \mathbb{P}\left(L(t)=k, \hat{\mu}_k(t) > \mu_k + \delta_k, N_k(t) > \lfloor \ell_k(t)/\gamma \rfloor\right) \qquad (3) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(L(t)=k, \exists k_2 \in \mathcal{B}_{\mathcal{N}(k)}, \hat{\mu}_{k_2}(t) \leq \mu_{k_2} - \delta_k, N_{k_2}(t) > (\ell_k(t))^b\right), \qquad (4)
\end{aligned}$$

where in (3), we have used the leader exploration mechanism. (3) and (4) can be upper bounded in the same way, by introducing the sequence of stopping times $(\tau_i^k)_i$, where τ_i^k is the instant at which arm k is the leader for the i-th time (one can have $\tau_i^k > T$ or $\tau_i^k = +\infty$ if arm k would be the leader only a finite number of time when UTS is run forever).

$$\begin{aligned}
(4) &\leq \sum_{k_2 \in \mathcal{B}_{\mathcal{N}(k)}} \sum_{i=1}^{T} \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}(L(t)=k, \ell_k(t)=i, \hat{\mu}_{k_2}(t) \leq \mu_{k_2} - \delta_k, N_{k_2}(t) > i^b)] \\
&= \tilde{B} \sum_{i=1}^{T} \mathbb{P}\left(\hat{\mu}_{k_2}(\tau_i^k) \leq \mu_{k_2} - \delta_k, N_{k_2}(\tau_i^k) > i^b, \tau_i^k \leq T\right) \\
&\leq \tilde{B} \sum_{i=1}^{T} \sum_{u=i^b}^{T} \mathbb{P}\left(\hat{\mu}_{k_2,u} \leq \mu_{k_2} - \delta_k, N_{k_2}(\tau_i^k) = u\right) \leq \tilde{B} \sum_{i=1}^{\infty} \sum_{u=i^b}^{\infty} \exp(-2\delta_k^2 u) \leq \tilde{B} \sum_{i=1}^{\infty} \frac{\exp(-2\delta_k^2 i^b)}{1-\exp(-2\delta_k^2)}.
\end{aligned}$$

The notation $\hat{\mu}_{k_2,u}$ denotes the empirical mean of the first u observations from arm k_2, which are i.i.d. with mean μ_{k_2}. Thus Hoeffding's inequality can be applied to obtain the last but one inequality.

To upper bound (3) we use the same approach (with i^b replaced by $\lfloor i/\gamma \rfloor$), which yields

$$\mathcal{T}_1^k(T) \leq \sum_{i=1}^{\infty} \frac{\exp(-2\delta_k^2 i^b)}{1-\exp(-2\delta_k^2)} + \sum_{i=1}^{\infty} \frac{\exp(-2\delta_k^2 \lfloor i/\gamma \rfloor)}{1-\exp(-2\delta_k^2)} := C_k(\boldsymbol{\mu}, \gamma, b) < \infty.$$

To finish the proof, we upper bound $\mathcal{T}_2^k(T)$ for some well chosen value of $b \in (0,1)$. The upper bound given in Lemma 7 is a careful adaptation (and generalization) of the proof of Proposition 1 in Kaufmann et al. (2012), which says that for vanilla Thompson Sampling restricted to $\mathcal{N}^+(k_\star)$, the (unique) optimal arm k_2 cannot be drawn too few times. Observe that Lemma 7 permits to handle possible multiple optimal arms. Again, we emphasize that in UTS, there is an extra difficulty due to the fact that arms in $\mathcal{N}^+(k_\star)$ are not only selected when k is the leader. The proof of Lemma 7, given in Appendix C overcomes this difficulty.

Lemma 7 *When $\gamma \geq 2$, there exists $b \in (0,1)$ and a constant $D_k(\boldsymbol{\mu}, b, \gamma)$ such that*

$$\sum_{t=1}^{T} \mathbb{P}\left(L(t) = k, \forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, N_{k_2}(t) \leq (\ell_k(t))^b\right) \leq D_k(\boldsymbol{\mu}, b, \gamma).$$

Putting things together, one obtains, for all $\varepsilon > 0$, with b chosen as in Lemma 7,

$$\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) \leq (1+\varepsilon) \sum_{k \in N(k_*)} \frac{\Delta_k}{\mathrm{kl}(\mu_k, \mu_\star)} \ln(T) + \tilde{C}(\boldsymbol{\mu}, \varepsilon) + \sum_{k \neq k_\star} \left[C_k(\boldsymbol{\mu}, \gamma, b) + D_k(\boldsymbol{\mu}, b, \gamma)\right],$$

which yields the claimed upper bound.

4. Related Work on Rank-One Bandits

Multi-armed bandits are a rich class of statistical models for sequential decision making (see Lattimore and Szepesvári (2019); Bubeck et al. (2012) for two surveys). They offer a clear framework as well as computationally efficient algorithms for many practical problems such as online advertising Zoghi et al. (2017), a context in which the empirical efficiency of Thompson Sampling (Thompson, 1933) has often been noticed (Scott, 2010; Chapelle and Li, 2011). The wide success of Bayesian methods in bandit or reinforcement learning problems can no longer be ignored Russo et al. (2018); Osband and Van Roy (2017).

Stochastic rank-one bandits were introduced by Katariya et al. (2017b,a) which are indeed among the closest works related to ours. The original algorithm proposed therein, `Rank1Elim`, relies on a complex sequential elimination scheme. It operates in stages that progressively quadruple in length. At the end of each stage, the significantly worst rows and columns are eliminated; this is done using carefully tuned confidence intervals. The exploration is simple but costly: every remaining row is played with a randomly chosen remaining column, and conversely for the columns. At the end of the stage, the value of each row is computed by averaging over all columns, such that the estimate of the row parameter is scaled by some measurable constant that is *the same* for all rows. Then, `UCB` or `KL-UCB` confidence intervals are used to perform the elimination by respectively `Rank1Elim` or `Rank1ElimKL`. The advantage of this method is that the worst rows and columns disappear very early from the game. However, eliminating them requires that their confidence intervals no longer intersect, which is quite costly. Moreover, the averaging performed to compute individual estimates for each parameter may be arbitrarily bad: if all columns but one have a parameter close to zero, the scaling constant on the row estimates is close to zero and the rows become hard to distinguish. All those issues are mentioned in the according papers. Nonetheless, the advantage of a rank-one algorithm, as opposed to playing a vanilla bandit algorithm, on a large (typically 64×64) matrix remains perfectly significant, which has motivated various further work on the topic.

In particular, Kveton et al. (2017) generalizes this elimination scheme to low-rank matrices, where the objective is to discover the $d \times d$ best set of entries. Jun et al. (2019) reformulate the problem as *Bilinear bandits*, where the two chosen vector arms x_t and y_t have an expected payoff of $x_t^\top M y_t$, where M is a low-rank matrix. Kotłowski and Neu (2019) study an adversarial version of this problem, the *Bandits Online PCA*: the learner sequentially chooses vectors x_t and observes a loss $x_t x_t^\top L_t$, where the loss is arbitrarily and possibly adversarially chosen by the environment. Zimmert and Seldin (2018) considers a more general problem where matrices are replaced by rank-one tensors in dimension $d \geq 2$. Their main message is to propose a unified view of *Factored Bandits* encompassing both rank-one bandits and dueling bandits Yue and Joachims (2009).

5. Numerical Experiments

To assess the empirical efficiency of UTS against other competitors, we follow the same experimental protocol as Katariya et al. (2017a) and run the algorithm on simulated matrices of arms. We set $K = L$ for different values of K. The parameters are defined symmetrically: $\boldsymbol{u} = \boldsymbol{v} = (0.75, 0.25, \ldots, 0.25)$ such that the best entry of the matrix is always $(i^*, j^*) = (1, 1)$. In our experiments, the cumulative regret up to $T = 300000$ is estimated based on 100 independent runs. The shaded areas on our plots show the 10% percentiles.

Study of the hyperparameter γ According to the original paper, the exploration parameter of UTS should be set to $\gamma = K+L-1$ for rank-one bandits. However, in the proof we derived in Section 3, there is no need to fix γ to this value. To confirm this statement and study the influence of γ, we ran UTS on the $K = 4$ toy problem described above, with different values of $\gamma \in \{2,\ 5,\ 10,\ 20\}$. We also run the heuristic version of UTS that would use no leader exploration scheme (corresponding to $\gamma = +\infty$). In Figure 2, we show the cumulative regret in log-scale. We notice that all curves align with the optimal logarithmic rate, with a lower offset for lower values of γ. Empirically, the performance seems the best for $\gamma = 2$.

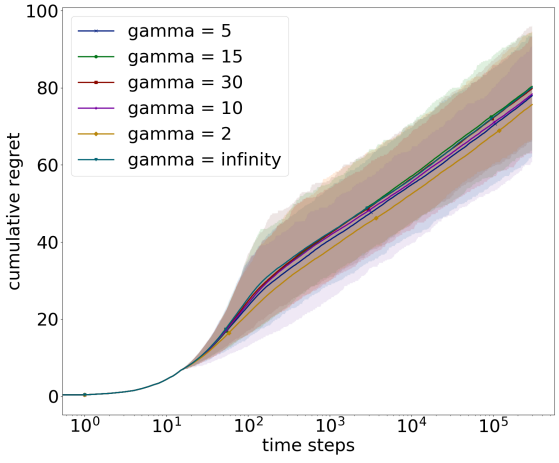

Figure 2: Cumulative regret of UTS for γ varying in $\{2, 5, 10, 15, 30, +\infty\}$ for $K = 4$.

Cumulative regret and optimality of `UTS` We now compare the regret of `UTS` run with $\gamma = 2$ to that of other algorithms on the above mentioned family of instances for $K = 4$ (Figure 3) and $K = 8, 16$ (Figure 4). Note that in Katariya et al. (2017a), the simulations are run on larger matrices, for $K = 32, 64, 128$. In those settings, `Rank1ElimKL` only outperforms `KL-UCB` for $K = 128$ but it is better than `Rank1Elim` and one can easily see that it scales better with the problem size than UCB1. However, given the much better performance of `UTS` and `OSUB`, we were able to show the same trends with much smaller problem sizes. Still, we also ran experiments with $K = 32, 64, 128$ and report the results in Appendix D.1, which confirm the superiority of `UTS`. In Appendix D.2, we also present results on a rank-one bandit with different choices of $\boldsymbol{u}$ and $\boldsymbol{v}$ leading to closer means.

In Figure 3 on the next page we compare the cumulative regret of `Rank1ElimKL` with `OSUB`, `UTS` (with $\gamma = 2$) and `KL-UCB`. One first obvious observation is that `Rank1ElimKL` has a regret an order of magnitude larger than all other policies, including `KL-UCB` on this size of problems. We also notice that the final regret, at $T = 300K$, roughly doubles for all rank-one policies while it quadruples for `KL-UCB`, as expected. To illustrate the asymptotic optimality of `OSUB` and `UTS` compared to `KL-UCB`, we show in Figure 3 (right) the results of the simulation in log-scale, and we plot the lower bound of Proposition 1. We observe that both optimal policies asymptotically align with the lower bound, while `KL-UCB` adopts a faster growth rate, that would correspond to the constant in the Lai & Robbins lower bound, which is larger than the constant in Proposition 1. Figure 4 confirms this observation for $K = 8$ and $K = 16$.

6. Conclusion

This paper proposed a new perspective on the rank-one bandit problem by showing it can be cast into the unimodal bandit framework. This led us to propose an algorithm closing the gap between existing regret upper and lower bound for Bernoulli rank-one bandits: Unimodal Thompson Sampling (`UTS`). `UTS` is easy to implement and very efficient in practice, as our experimental study reveals an improvement of a factor at least 20 with respect to the state-of-the art `Rank1ElimKL` algorithm. Our main theoretical contribution is a novel regret analysis of this algorithm in the general unimodal setting, which sheds a new light on the leader exploration parameter to use. Interestingly, we show that forcing exploration of the leader appears to help in practice in the rank-one example, and it may be interesting to investigate whether this remains the case for other structured bandit problems (Combes et al., 2017).

Acknowledgments

The authors acknowledge the French National Research Agency under projects BADASS (ANR-16-CE40-0002) and BOLD (ANR-19-CE23-0026-04).

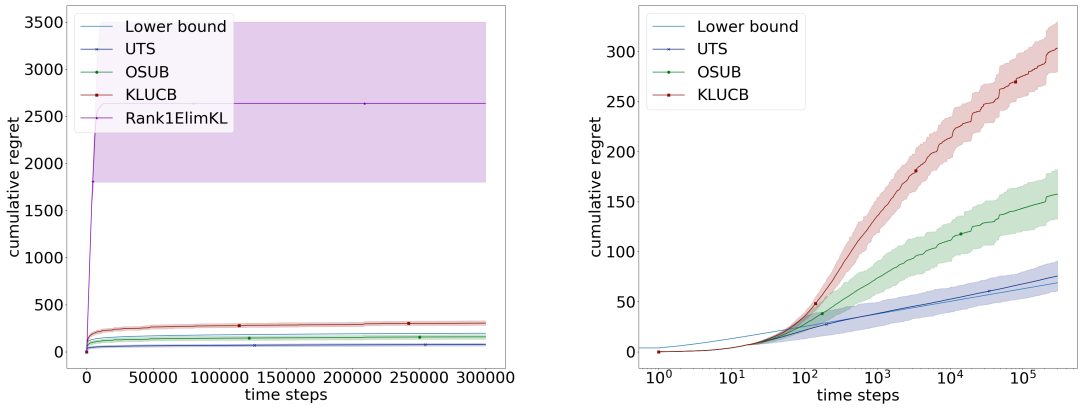

Figure 3: Cumulative regret of `Rank1ElimKL`, `OSUB`, `UTS`, `KL-UCB`, on 4×4 rank-one matrices (left). Regret in log-scale: the lower bound (in blue) shows the optimal asymptotic logarithmic growth of the regret. `UTS` and `OSUB` align with it, while `KL-UCB` has a larger slope (right).

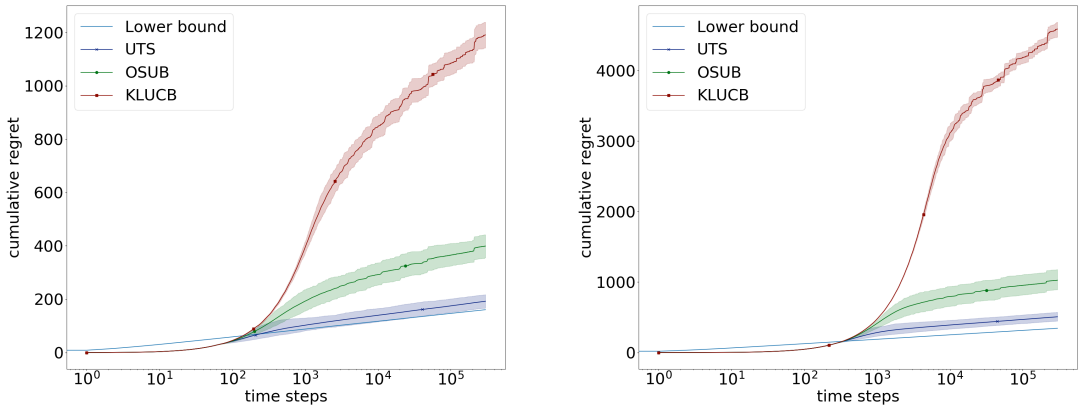

Figure 4: Cumulative regret of `OSUB`, `UTS` and `KL-UCB`, on $K \times K$ rank-one matrices with $K = 8$ (left) and $K = 16$ (right), in log scale.

References

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*, 2013.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(3):1–115, 2015.

Richard Combes and Alexandre Proutière. Unimodal bandits: Regret lower bounds and optimal algorithms. 2014.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, 2017.

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. *arXiv preprint arXiv:1901.02470*, 2019.

Sumeet Katariya, Branislav Kveton, Csaba Szepesvári, Claire Vernade, and Zheng Wen. Bernoulli rank-1 bandits for click feedback. In *IJCAI*, 2017a.

Sumeet Katariya, Branislav Kveton, Csaba Szepesvári, Claire Vernade, and Zheng Wen. Stochastic rank-1 bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017b.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis. In *Proceedings of the 23rd conference on Algorithmic Learning Theory*, 2012.

Wojciech Kotłowski and Gergely Neu. Bandit principal component analysis. *arXiv preprint arXiv:1902.03035*, 2019.

Branislav Kveton, Csaba Szepesvári, Anup Rao, Zheng Wen, Yasin Abbasi-Yadkori, and S Muthukrishnan. Stochastic low-rank bandits. *arXiv preprint arXiv:1712.04644*, 2017.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2019.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2701–2710. JMLR. org, 2017.

Stefano Paladino, Francesco Trovò, Marcello Restelli, and Nicola Gatti. Unimodal thompson sampling for graph-structured arms. *arXiv:1611.05724v2*, 2016.

Stefano Paladino, Francesco Trovò, Marcello Restelli, and Nicola Gatti. Unimodal thompson sampling for graph-structured arms. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM, 2007.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Steven L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639–658, 2010.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Jia Yuan Yu and Shie Mannor. Unimodal bandits. Citeseer, 2011.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208. ACM, 2009.

Julian Zimmert and Yevgeny Seldin. Factored bandits. In *Advances in Neural Information Processing Systems*, pages 2835–2844, 2018.

Masrour Zoghi, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. Online learning to rank in stochastic click models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4199–4208. JMLR. org, 2017.

Appendix A. Important Results

We recall two important results that are repeatedly used in our analysis.

Lemma 8 *(Hoeffding's inequality) Let $X_1, ..., X_n$ be independent bounded random variables supported in $[0, 1]$. For all $t \geq 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp(-2nt^2)$$

and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \leq -t\right) \leq \exp(-2nt^2)$$

Lemma 9 *(Beta Binomial trick) Letting $F^{\text{Beta}}_{\alpha,\beta}$ and $F^{\text{Bin}}_{n,p}$ respectively denote the cumulative distribution function of a Beta distribution with parameters α, β, and of a Binomial distribution with parameters (n, p). It holds that*

$$F^{\text{Beta}}_{\alpha,\beta}(y) = 1 - F^{\text{Bin}}_{\alpha+\beta-1,y}(\alpha - 1)$$

Appendix B. Proof of Lemma 6

In this section, we adapt the analysis of Agrawal and Goyal (2013), highlighting the steps that need extra justification.

Let k be a sub-optimal arm. We introduce two thresholds x_k and y_k such that $\mu_k < x_k < y_k < \mu_{k_\star}$, that we specify later. We define the following "good" events: E_k^μ (t) = $\{\hat{\mu}_k(t) \leq x_k\}$ and E_k^θ (t) = $\{\theta_k(t) \leq y_k\}$. The event $\{K(t) = k, L(t) = k_\star\}$ can be decomposed as follows:

$$\{K(t) = k, L(t) = k_\star\} = \{K(t) = k, L(t) = k_\star, E_k^\mu(t), E_k^\theta(t)\} \tag{5}$$

$$\cup \{K(t) = k, L(t) = k_\star, E_k^\mu(t), \overline{E_k^\theta(t)}\} \tag{6}$$

$$\cup \{K(t) = k, L(t) = k_\star, \overline{E_k^\mu(t)}\} \tag{7}$$

Observe that for $k \notin \mathcal{N}(k_\star)$, by definition of the algorithm, $\{K(t) = k, L(t) = k_\star\} = \emptyset$. For $k \in \mathcal{N}(k_\star)$, we now upper bound the probability of the three events in the decomposition.

Upper Bound on the Probability of (5) We prove the following lemma.

Lemma 10 *For all $k \in \mathcal{N}(k_\star)$, there exists a constant $\bar{C}_1(\mu_{k_\star}, y_k)$ such that*

$$\sum_{t=1}^{T} \mathbb{P}\left(K(t) = k, L(t) = k_\star, E_k^\mu(t), E_k^\theta(t)\right) \leq \bar{C}_1(\mu_{k_\star}, y_k)$$

Proof We first prove the following inequality

$$\mathbb{P}\left(K(t) = k, L(t) = k_\star, E_k^\mu(t), E_k^\theta(t) | \mathcal{F}_{t-1}\right) \leq \frac{1 - p_{kt}}{p_{kt}} \mathbb{P}\left(K(t) = k_\star, L(t) = k_\star, E_k^\mu(t), E_k^\theta(t) | \mathcal{F}_{t-1}\right) \tag{8}$$

where $p_{kt} = \mathbb{P}(\theta_1(t) > y_k | \mathcal{F}_{t-1}) = \mathbb{P}(\overline{E_k^\theta(t)} | \mathcal{F}_{t-1})$. To do so, notice that E_k^μ (t) and $\{L(t) = k_\star\}$ are $\mathcal{F}_{t-1}$-measurable, since $\hat{\mu}_k(t)$ is completely determined by the rewards and arms drawn up to time $t-1$. Therefore, one can assume that $\mathcal{F}_{t-1}$ is such that E_k^μ (t) and $\{L(t) = k_\star\}$ hold, and it suffices to show that

$$\mathbb{P}(K(t) = k, E_k^\theta(t) | \mathcal{F}_{t-1}) \leq \frac{1 - p_{kt}}{p_{kt}} \mathbb{P}(K(t) = k_\star, E_k^\theta(t) | \mathcal{F}_{t-1})$$

which can be proved as in Agrawal and Goyal (2013). With (8), we get

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}(K(t) = k, L(t) = k_\star, E_k^\mu(t), E_k^\theta(t)) \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{P}(K(t) = k, L(t) = k_\star, E_k^\mu(t), E_k^\theta(t) | \mathcal{F}_{t-1})\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{1 - p_{kt}}{p_{kt}} \mathbb{1}(K(t) = k_\star, L(t) = k_\star, E_k^\mu(t), E_k^\theta(t) | \mathcal{F}_{t-1}\right]\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{1 - p_{kt}}{p_{kt}} \mathbb{1}(K(t) = k_\star, L(t) = k_\star, E_k^\mu(t), E_k^\theta(t))\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{1 - p_{kt}}{p_{kt}} \mathbb{1}(K(t) = k_\star, E_k^\mu(t), E_k^\theta(t))\right]
\end{aligned}$$

which allows to continue with the same proof as Theorem 1 in Agrawal and Goyal (2013). ■

Upper Bound on the Probability of (6) We prove the following lemma.

Lemma 11 *For all $k \in \mathcal{N}(k_\star)$, letting $L_k(T) = \frac{\ln T}{\mathrm{kl}(x_k, y_k)}$, it holds that*

$$\sum_{t=1}^{T} \mathbb{P}\left(K(t) = k, L(t) = k_\star, \overline{E_k^\theta(t)}, E_k^\mu(t)\right) \leq L_k(T) + 1.$$

Proof We start by the following decomposition:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left(K(t) = k, L(t) = k_\star, \overline{E_k^\theta(t)}, E_k^\mu(t)\right) \\
&\quad \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(K(t) = k, L(t) = k_\star, N_k(t) \leq L_k(T), \overline{E_k^\theta(t)}, E_k^\mu(t)\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(K(t) = k, L(t) = k_\star, N_k(t) > L_k(T), \overline{E_k^\theta(t)}, E_k^\mu(t)\right)\right]
\end{aligned}$$

The first term of the sum is clearly bounded by $L_k(T)$. As for the second term, we can directly upper bound it as follows

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(K(t) = k, L(t) = k_\star, N_k(t) > L_k(T), \overline{E_k^\theta(t)}, E_k^\mu(t)\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(K(t) = k, N_k(t) > L_k(T), \overline{E_k^\theta(t)}, E_k^\mu(t)\right)\right]
\end{aligned}$$

and the conclusion follows from the same steps used in the proof of Lemma 4 of Agrawal and Goyal (2013). ■

Upper Bound on the Probability of (7) We prove the following lemma.

Lemma 12 *For $k \in \mathcal{N}(k_\star)$,*

$$\sum_{t=1}^{T} \mathbb{P}\left(K(t) = k, L(t) = k_\star, \overline{E_k^\mu(t)}\right) \leq \frac{1}{\mathrm{kl}(x_k, \mu_k)} + 1$$

Proof To prove this lemma, one can write

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}(K(t) = k, L(t) = k_\star, \overline{E_k^\mu(t)}) &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k, L(t) = k_\star, \overline{E_k^\mu(t)})\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k, \overline{E_k^\mu(t)})\right]
\end{aligned}$$

and use the same steps as in the proof of Lemma 3 of Agrawal and Goyal (2013). ■

Conclusion For $0 < \varepsilon \leq 1$, we can choose x_k and y_k in $(\mu_k, \mu_{k_\star})$ such that $\mathrm{kl}(x_k, y_k) = \frac{\mathrm{kl}(\mu_k, \mu_{k_\star})}{(1+\varepsilon)}$. Using the three above lemmas yields, for all $k \in \mathcal{N}(k_\star)$:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k, L(t) = k_\star)\right] \leq (1+\varepsilon)\frac{\Delta_k}{\mathrm{kl}(\mu_k, \mu_\star)} \ln(T) + \tilde{C}(\boldsymbol{\mu}, \varepsilon).$$

Since when the leader is $k_\star$, $\mathbb{1}(K(t) = k, L(t) = k_\star) = 0$ for all $k \notin \mathcal{N}^+(k_\star)$, we only need to sum over the arms $k \in \mathcal{N}(k_\star)$ to get the result of Lemma 6.

Appendix C. Proof of Lemma 7

Let $k \in [K] \setminus \{k_\star\}$.

Notation Recall from Section 3.3 that $\mathcal{B}_{\mathcal{N}(k)} = \text{argmax}_{\ell \in \mathcal{N}(k)} \mu_\ell$ is the set of best arms in the neighborhood of k. This set is such that $1 \leq |\mathcal{B}_{\mathcal{N}(k)}| \leq \tilde{B}$, and arms belonging to $\mathcal{B}_{\mathcal{N}(k)}$ have same mean, that we denote $\mu_{k_2} = \max_{\ell \in \mathcal{N}(k)} \mu_\ell$. We also define $N_{\mathcal{B}_{\mathcal{N}(k)}}(t) = \sum_{k_2 \in \mathcal{B}_{\mathcal{N}(k)}} N_{k_2}(t)$, the number of times arms belonging to $\mathcal{B}_{\mathcal{N}(k)}$ have been drawn up to time t. We will say that $k' \in \mathcal{N}^+(k)$ is sub-optimal if $\mu_{k'} < \mu_{k_2}$. We denote by $\tilde{M}_k = |\mathcal{N}^+(k) \setminus \mathcal{B}_{\mathcal{N}(k)}| \leq |\mathcal{N}(k)|$, the number of sub-optimal arms belonging to $\mathcal{N}(k)$.

We introduce, for every arm k',

$$\delta_{k'} = \frac{\mu_{k_2} - \mu_{k'}}{2}, \quad \text{and let} \quad \delta = \min_{k' \in \mathcal{N}^+(k) \setminus \mathcal{B}_{\mathcal{N}(k)}} \delta_{k'} \quad \text{and} \quad C := \frac{6}{\delta^2}.$$

We denote by $\tilde{k}$ any arm satisfying $\delta_{\tilde{k}} = \delta$.

Just like in Section 3, we introduce the consecutive instants in which arm k is the leader, τ_i^k. Assuming that UTS(γ) would be played forever, the instant of the i-th time arm k is the leader, τ_i^k, can be formally written as such

$$\tau_i^k = \inf\{t \in \mathbb{N} : L(t) = k, \ell_k(t) = i\},$$

with the convention that $\inf \emptyset = +\infty$.

For all $i \in \{1, \dots, T\}$, for all $b \in (0, 1)$, by definition of τ_i^k, it holds that

$$\sum_{t=1}^{T} \mathbb{1}\left(L(t) = k, \ell_k(t) = i, \forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, N_{k_2}(t) \leq (\ell_k(t))^b\right) = \mathbb{1}\left(\forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, N_{k_2}(\tau_i^k) \leq i^b\right) \mathbb{1}\left(\tau_i^k \leq T\right),$$

which permits to rewrite

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{P}\left(L(t) = k, \forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, N_{k_2}(t) \leq (\ell_k(t))^b\right) &= \sum_{i=1}^{T} \mathbb{P}\left(\forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, N_{k_2}\left(\tau_i^k\right) \leq i^b, \tau_i^k \leq T\right) \\ &\leq \sum_{i=1}^{T} \mathbb{P}\left(N_{\mathcal{B}_{\mathcal{N}(k)}}\left(\tau_i^k\right) \leq \tilde{B} i^b, \tau_i^k \leq T\right), \end{aligned} \tag{9}$$

where we recall that $N_{\mathcal{B}_{\mathcal{N}(k)}}(t)$ is the total number of pulls of all arms in $B_{\mathcal{N}}(k)$.

We now provide an upper bound on (9), for a well chosen value of b.

Our analysis bears similarity with that of Kaufmann et al. (2012): we use the fact that if arms belonging to $\mathcal{B}_{\mathcal{N}(k)}$ are not drawn much at time τ_i^k, there must exist many consecutive instants $\tau_\ell^k < \tau_i^k$ in which those arms are not selected at all. To formalize this idea, we introduce for every pair i, j the first instant preceeding τ_i^k in which arms of $\mathcal{B}_{\mathcal{N}(k)}$ have been played at least j times while arm k is the leader:

$$\nu_{i,j} = \inf\{\ell \leq i : N_{\mathcal{B}_{\mathcal{N}(k)}}(\tau_\ell^k) \geq j\},$$

with the convention $\inf \emptyset = i+1$. It holds that

$$\left(N_{\mathcal{B}_{\mathcal{N}(k)}}\left(\tau_i^k\right) \leq \tilde{B} i^b\right) = \left(\nu_{i,\lceil \tilde{B} i^b \rceil} = i+1\right) \subseteq \bigcup_{j=0}^{\lfloor \tilde{B} i^b \rfloor} \left(\nu_{i,j+1} - \nu_{i,j} \geq \frac{i^{1-b}}{\tilde{B}} - 1\right).$$

We now introduce $\mathcal{I}_{i,j} \subseteq \left(\nu_{i,j}, \nu_{i,j} + \lceil \frac{i^{1-b}}{\tilde{B}} - 2 \rceil\right]$, the subset of instants belonging to $\left(\nu_{i,j}, \nu_{i,j} + \lceil \frac{i^{1-b}}{\tilde{B}} - 2 \rceil\right]$ where no leader exploration is performed. The j-th event in the union implies that no arm belonging to $\mathcal{B}_{\mathcal{N}(k)}$ is selected in any instant τ_ℓ^k for $\ell \in \mathcal{I}_{i,j}$. More precisely, introducing

$$\mathcal{E}_{i,j} = \{\mathcal{I}_{i,j} \subseteq [i]\} \bigcap \left\{\forall \ell \in \mathcal{I}_{i,j}, K(\tau_\ell^k) \notin \mathcal{B}_{\mathcal{N}(k)}\right\}$$

one has

$$\mathbb{P}\left(N_{\mathcal{B}_{\mathcal{N}(k)}}\left(\tau_i^k\right) \leq i^b, \tau_i^k \leq T\right) \leq \sum_{j=0}^{\lfloor \tilde{B} i^b \rfloor} \mathbb{P}\left(\mathcal{E}_{i,j}, \tau_i^k \leq T\right). \tag{10}$$

Interval sub-division and saturated arms To further upper bound (10), we introduce for $m = 1, \ldots, \tilde{M}_k + 1$, the intervals $\mathcal{I}_{i,j,m}$:

$$\mathcal{I}_{i,j,m} := \left(\nu_{i,j} + (m-1)\left\lfloor \frac{i^{1-b}/\tilde{B} - 2}{\tilde{M}_k + 1} \right\rfloor, \nu_{i,j} + m\left\lfloor \frac{i^{1-b}/\tilde{B} - 2}{\tilde{M}_k + 1} \right\rfloor\right] \cap \mathcal{I}_{i,j},$$

whose length is lower bounded as follows, substracting the instant in which leader exploration is performed (that are not included in $\mathcal{I}_{i,j}$):

$$|\mathcal{I}_{i,j,m}| = \left\lfloor \frac{i^{1-b}/\tilde{B} - 2}{\tilde{M}_k + 1} \right\rfloor - \left\lceil \frac{1}{\gamma}\left(\frac{i^{1-b}/\tilde{B} - 2}{\tilde{M}_k + 1}\right) \right\rceil \geq \left\lfloor \left(1 - \frac{1}{\gamma}\right)\left(\frac{i^{1-b}/\tilde{B} - 2}{\tilde{M}_k + 1}\right) - 2 \right\rfloor := \tilde{H}_{i,b,k,\gamma}.$$

As in Kaufmann et al. (2012), we introduce the notion of saturated sub-optimal arm: we say an arm $k' \notin \mathcal{B}_{\mathcal{N}(k)}$ is saturated at ℓ if $N_{k'}(\tau_\ell^k) > C \ln(i)$. Otherwise, it is unsaturated. For an interval $\mathcal{I}_{i,j,m}$, we denote by $n_{i,j,m}$ the number of interruptions, that is, the number of times we draw an unsaturated arm during $\mathcal{I}_{i,j,m}$. We introduce $F_{i,j,m}$, the event that by the end of $\mathcal{I}_{i,j,m}$, at least m sub-optimal arms are saturated, and $\mathcal{S}_{i,j,m}$, the set of saturated arms at the end of $\mathcal{I}_{i,j,m}$.

We decompose the probability of the event $\{\mathcal{E}_{i,j}, \tau_i^k \leq T\}$ as follows

$$\mathbb{P}[\mathcal{E}_{i,j}, \tau_i^k \leq T] \leq \mathbb{P}[\mathcal{E}_{i,j}, F_{i,j,\tilde{M}_k}, \tau_i^k \leq T] \tag{11}$$

$$+ \mathbb{P}[\mathcal{E}_{i,j}, F^c_{i,j,\tilde{M}_k}, \tau_i^k \leq T] \tag{12}$$

We will prove below that

$$(11) \leq \frac{2\tilde{M}_k}{i^2(1 - \exp(-\delta^2/2))} + g_1(\boldsymbol{\mu}, j, b, i, k, \gamma) \tag{13}$$

and that for i larger than some constant $N_{\boldsymbol{\mu},b}$,

$$(12) \leq (\tilde{M}_k - 1)\left(\frac{2\tilde{M}_k}{i^2(1 - \exp(-\delta^2/2))} + g_2(\boldsymbol{\mu}, j, b, i, k, \gamma)\right) \tag{14}$$

where for a well-chosen $b \in (0,1)$ and $\gamma \geq 2$

$$\sum_{i=1}^{\infty} \sum_{j=0}^{\lfloor \tilde{B} i^b \rfloor} g_1(\boldsymbol{\mu}, j, b, i, k, \gamma) < \infty \text{ and } \sum_{i=1}^{\infty} \sum_{j=0}^{\lfloor \tilde{B} i^b \rfloor} g_2(\boldsymbol{\mu}, j, b, i, k, \gamma) < \infty.$$

Combining (10) with the upper bounds (13) and (14), we get

$$\begin{aligned}
(9) &\leq M_{\boldsymbol{\mu},b} + \sum_{i=N_{\boldsymbol{\mu},b}+1}^{T} \sum_{j=0}^{\lfloor \tilde{B} i^b \rfloor} \mathbb{P}[\mathcal{E}_{i,j}, \tau_i^k \leq T] \\
&\leq M_{\boldsymbol{\mu},b} + \sum_{i=1}^{T} \sum_{j=0}^{\lfloor \tilde{B} i^b \rfloor} \left[\frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))} + g_1(\boldsymbol{\mu}, j, b, i, k, \gamma) \right] \\
&+ \sum_{i=1}^{T} \sum_{j=0}^{\lfloor \tilde{B} i^b \rfloor} \left[(\tilde{M}_k - 1) \left(\frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))} + g_2(\boldsymbol{\mu}, j, b, i, k, \gamma) \right) \right] \\
&\leq M_{\boldsymbol{\mu},b} + \sum_{i=1}^{\infty} \frac{2\tilde{B}\tilde{M}_k^2}{i^{2-b}(1-\exp(-\delta^2/2))} + \sum_{i=1}^{\infty} \sum_{j=0}^{\lfloor \tilde{B} i^b \rfloor} [g_1(\boldsymbol{\mu}, j, b, i, k, \gamma) + g_2(\boldsymbol{\mu}, j, b, i, k, \gamma)] \\
&:= D_k(\boldsymbol{\mu}, b, \gamma),
\end{aligned}$$

which concludes the proof. We now prove the two crucial upper bounds (13) and (14).

Main ingredients We introduce two useful lemmas whose proofs are postponed to the end of this appendix. Lemma 13 establishes that it is unlikely that the Thompson sample associated to some saturated arm exceeds its true mean by too much.

Lemma 13 *Let $k \in [K]$.*

$$\mathbb{P}\left(\exists \ell \leq i, \exists k' \notin \mathcal{B}_{\mathcal{N}(k)}, \theta_{k'}(\tau_\ell^k) > \mu_{k'} + \delta, N_{k'}(\tau_\ell^k) > C \ln(i), \tau_i^k < T \right) \leq \frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))}$$

Lemma 14 shows that the Thompson samples of an arm belonging to $\mathcal{B}_{\mathcal{N}(k)}$ are unlikely to fall below $\mu_{\tilde{k}} + \delta$ during a long interval in which the posterior of this arm doesn't evolve.

Lemma 14 *Let $\tilde{\mathcal{I}}$ be a random interval such that $\forall \ell \in \tilde{\mathcal{I}}, N_{\mathcal{B}_{\mathcal{N}(k)}}(\tau_\ell^k) = j$ and $|\tilde{\mathcal{I}}| \geq x$ for some deterministic constant x. There exists $\lambda_0 = \lambda_0(\mu_{k_2}, \mu_{\tilde{k}}, \delta) > 1$ such that for $\lambda \in]1, \lambda_0[$,*

$$\mathbb{P}\left(\forall \ell \in \tilde{\mathcal{I}}, \forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta \right) \leq j\tilde{B}(\alpha_{\mu_{\tilde{k}},\delta})^x + C_{\lambda,\mu_{k_2},\mu_{\tilde{k}}} \frac{\exp(-j d_{\lambda,\mu_{k_2},\mu_{\tilde{k}}}/\tilde{B})}{x^\lambda},$$

where $C_{\lambda,\mu_{k_2},\mu_{\tilde{k}}}, d_{\lambda,\mu_{k_2},\mu_{\tilde{k}}} > 0$, and $\alpha_{\mu_{\tilde{k}},\delta} = \left(\frac{1}{2}\right)^{1-\mu_{\tilde{k}}-\delta}$.

Proof of the Upper bound (13) On the event $\mathcal{E}_{i,j} \cap F_{i,j,\tilde{M}_k}$, only saturated arms are drawn during the interval $\mathcal{I}_{i,j,\tilde{M}_k+1}$, so that one has the following decomposition:

$$\begin{aligned}
&\mathbb{P}[\mathcal{E}_{i,j} \cap F_{i,j,\tilde{M}_k} \cap \{\tau_i^k \leq T\}] \\
&\leq \mathbb{P}[\{\exists \ell \in \mathcal{I}_{i,j,\tilde{M}_k+1}, \exists k' \notin \mathcal{B}_{\mathcal{N}(k)}, \theta_{k'}(\tau_\ell^k) > \mu_{k'} + \delta\} \cap \{N_{k'}(\tau_\ell^k) > C\ln(i)\} \cap \mathcal{E}_{i,j} \cap \{\tau_i^k \leq T\}] \\
&\quad + \mathbb{P}[\{\forall \ell \in \mathcal{I}_{i,j,\tilde{M}_k+1}, \forall k' \notin \mathcal{B}_{\mathcal{N}(k)}, \theta_{k'}(\tau_\ell^k) \leq \mu_{k'} + \delta\} \cap \mathcal{E}_{i,j} \cap F_{i,j,|\mathcal{N}^+(k)|-1}] \\
&\leq \mathbb{P}\left(\exists \ell \leq i, \exists k' \notin \mathcal{B}_{\mathcal{N}(k)}, \theta_{k'}(\tau_\ell^k) > \mu_{k'} + \delta, N_{k'}(\tau_\ell^k) > C\ln(i), \tau_i^k \leq T\right) \\
&\quad + \mathbb{P}(\forall \ell \in \mathcal{I}_{i,j,\tilde{M}_k+1}, \forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta, \mathcal{E}_{i,j})
\end{aligned}$$

Using Lemma 13, we can bound the first term in this sum by

$$\frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))}.$$

On the event $\mathcal{E}_{i,j}$, $N_{\mathcal{B}_{\mathcal{N}(k)}}(\tau_\ell^k) = j$ for all $\ell \in \mathcal{I}_{i,j,\tilde{M}_k+1}$. Lemma 14 with $\tilde{\mathcal{I}} = \mathcal{I}_{i,j,\tilde{M}_k+1}$ and $x = \tilde{H}_{i,b,k,\gamma}$ yields the following upper bound for the second term

$$j\tilde{B}(\alpha_{\mu_{\tilde{k}},\delta})^{\tilde{H}_{i,b,k,\gamma}} + C_{\lambda,\mu_{k_2},\mu_{\tilde{k}}} \frac{\exp(-jd_{\lambda,\mu_{k_2},\mu_{\tilde{k}}}/\tilde{B})}{\tilde{H}_{i,b,k,\gamma}^{\lambda}} := g_1(\boldsymbol{\mu}, j, b, i, k, \gamma).$$

Summing $g_1(\boldsymbol{\mu}, j, b, i, k, \gamma)$ over $j \leq \lfloor \tilde{B}i^b \rfloor$ and expliciting $\tilde{H}_{i,b,k,\gamma}$ gives

$$\sum_{j \leq \lfloor \tilde{B}i^b \rfloor} g_1(\boldsymbol{\mu}, j, b, i, k, \gamma) = \tilde{B}\frac{\lfloor \tilde{B}i^b \rfloor(\lfloor \tilde{B}i^b \rfloor + 1)}{2}(\alpha_{\mu_{\tilde{k}},\delta})^{\left\lfloor \left(1-\frac{1}{\gamma}\right)\frac{i^{1-b}/\tilde{B}-2}{\tilde{M}_k+1} - 2 \right\rfloor} + \frac{C'_{\lambda,\mu_{k_2},\mu_{\tilde{k}}}}{\left\lfloor \left(1-\frac{1}{\gamma}\right)\frac{i^{1-b}/\tilde{B}-2}{\tilde{M}_k+1} - 2 \right\rfloor^{\lambda}}.$$

The first term of the sum is $o\left(\frac{1}{i^2}\right)$, and by choosing $b < 1 - \frac{1}{\lambda}$ for the second term, we obtain that $\sum_{i \leq \infty} \sum_{j \leq \lfloor i^b \rfloor} g_1(\boldsymbol{\mu}, j, b, i, k, \gamma)$ is finite when $\gamma > 1$.

Proof of the Upper Bound (14) Similarly to Kaufmann et al. (2012), we prove by induction that for all $2 \leq m \leq \tilde{M}_k + 1$, if i is larger than some deterministic constant $N_{\boldsymbol{\mu},b}$,

$$\mathbb{P}[\mathcal{E}_{i,j} \cap F_{i,j,m-1}^c \cap \{\tau_i^k \leq T\}] \leq (m-2)\left(\frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))} + g_2(\boldsymbol{\mu}, j, b, i, k, \gamma)\right),$$

where $N_{\boldsymbol{\mu},b}$ and $g_2(\boldsymbol{\mu}, j, b, i, k, \gamma)$ are made precise below.

Base case of the induction: On the event $\mathcal{E}_{i,j}$, only suboptimal arms are played during the interval $\mathcal{I}_{i,j,1}$, of length larger than $\tilde{H}_{i,b,k,\gamma}$. Hence at least one suboptimal arm must be played more than $\lceil \frac{\tilde{H}_{i,b,k,\gamma}}{\tilde{M}_k} \rceil$ times. Besides, there exists some deterministic constant $N_{\boldsymbol{\mu},b}$ such that for $i > N_{\boldsymbol{\mu},b}$, $\lceil \frac{\tilde{H}_{i,b,k,\gamma}}{\tilde{M}_k} \rceil \geq C\ln(i)$.

Therefore, when $i \geq N_{\boldsymbol{\mu},b}$, at least one suboptimal arm is saturated by the end of $\mathcal{I}_{i,j,1}$, so that for $i \geq N_{\boldsymbol{\mu},b}$, $\mathbb{P}(\mathcal{E}_{i,j} \cap F_{i,j,1}^c \cap \{\tau_i^k \leq T\}) = 0$. Hence, the inequality holds for $m = 2$.

Induction: Let us assume the following, for some $m \in \{2, \dots, \tilde{M}_k\}$:

$$\mathbb{P}(\mathcal{E}_{i,j} \cap F^c_{i,j,m-1} \cap \{\tau_i^k \leq T\}) \leq (m-2)\left(\frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))} + g_2(\boldsymbol{\mu}, j, b, i, k.\gamma)\right) .$$

Exploiting this inductive hypothesis, one obtains

$$\begin{aligned}
&\mathbb{P}(\mathcal{E}_{i,j} \cap F^c_{i,j,m} \cap \{\tau_i^k \leq T\}) \\
&\quad \leq \mathbb{P}(\mathcal{E}_{i,j} \cap F^c_{i,j,m-1} \cap \{\tau_i^k \leq T\}) + \mathbb{P}(\mathcal{E}_{i,j} \cap F^c_{i,j,m} \cap F_{i,j,m-1} \cap \{\tau_i^k \leq T\}) \\
&\quad \leq (m-2)\left(\frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))} + g_2(\boldsymbol{\mu}, j, b, i, k, \gamma)\right) + \mathbb{P}(\mathcal{E}_{i,j} \cap F^c_{i,j,m} \cap F_{j,m-1} \cap \{\tau_i^k \leq T\}) .
\end{aligned}$$

Let us prove that the second term of the sum is bounded by $\frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))} + g_2(\boldsymbol{\mu}, j, b, i, k)$.

On the event $(\mathcal{E}_{i,j} \cap F^c_{i,j,m} \cap F_{i,j,m-1})$, there are exactly $m-1$ saturated arms at the beginning of interval $\mathcal{I}_{i,j,m}$ and no new arm is saturated during this interval, so that $\mathcal{S}_{i,j,m-1} = \mathcal{S}_{i,j,m}$. As a result, there cannot be more than $\tilde{M}_k C \ln(i)$ interruptions during this interval, so that

$$\begin{aligned}
&\mathbb{P}(\mathcal{E}_{i,j} \cap F^c_{i,j,m} \cap F_{i,j,m-1} \cap \{\tau_i^k \leq T\}) \\
&\leq \mathbb{P}(\mathcal{E}_{i,j} \cap F_{i,j,m-1} \cap \{n_{i,j,m} \leq \tilde{M}_k C \ln(i)\} \cap \{\tau_i^k \leq T\}) \\
&\leq \mathbb{P}(\{\exists \ell \in \mathcal{I}_{i,j,m}, \exists k' \in \mathcal{S}_{i,j,m-1} \backslash \mathcal{B}_{\mathcal{N}(k)}, \theta_{k'}(\tau_\ell^k) > \mu_{k'} + \delta\} \cap \mathcal{E}_{i,j} \cap \{\tau_i^k \leq T\}) \qquad (15) \\
&+ \mathbb{P}(\{\forall \ell \in \mathcal{I}_{i,j,m}, \forall k' \in \mathcal{S}_{i,j,m-1} \backslash \mathcal{B}_{\mathcal{N}(k)}, \theta_{k'}(\tau_\ell^k) \leq \mu_{k'} + \delta\} \cap \mathcal{E}_{i,j} \cap F_{i,j,m-1} \cap \{n_{i,j,m} \leq \tilde{M}_k C \ln(i)\}) \\
&\qquad (16)
\end{aligned}$$

Lemma 13 allows us to bound the term (15):

$$(15) \leq \frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))}.$$

To deal with (16), we introduce the random intervals

$$\mathcal{J}_h = \{\ell \in \mathcal{I}_{i,j,m}, \text{between the } h\text{-th and } (h+1)\text{-th interrruptions}\}.$$

On the event in the probability of (16), there exists an interval $\mathcal{J}_h$ of length larger than $\lceil \frac{\tilde{H}_{i,b,k,\gamma}}{\tilde{M}_k C \ln(i)} \rceil$ such that there is no interruption at times τ_ℓ^k, for $\ell \in \mathcal{J}_h$. This means that, at these time steps, all Thompson samples are smaller than that of the greatest sample among the saturated arms (which are themselves smaller than $\mu_{\tilde{k}} + \delta$). In particular, in this interval, $\forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta$, and we get

$$\begin{aligned}
(16) &\leq \mathbb{P}\left(\{\exists h \in 0, ..., n_{i,j,m} - 1, |\mathcal{J}_h| \geq \left\lceil \frac{\tilde{H}_{i,b,k,\gamma}}{\tilde{M}_k C \ln(i)} \right\rceil\}\right. \\
&\qquad \left. \cap \{\forall \ell \in \mathcal{J}_h, \forall k' \in \mathcal{S}_{i,j,m-1} \backslash \mathcal{B}_{\mathcal{N}}(k), \theta_{k'}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta\} \cap \mathcal{E}_{i,j} \cap F_{i,j,m-1}\right) \\
&\leq \sum_{h=0}^{\tilde{M}_k C \ln(i)} \mathbb{P}\left(\{|\mathcal{J}_h| \geq \left\lceil \frac{\tilde{H}_{i,b,k,\gamma}}{\tilde{M}_k C \ln(i)} \right\rceil\} \cap \{\forall \ell \in \mathcal{J}_h, \forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta\} \cap \mathcal{E}_{i,j}\right) .
\end{aligned}$$

Applying Lemma 14 with $\tilde{\mathcal{I}} = \mathcal{J}_h$, we get

$$(16) \leq \tilde{M}_k C \ln(i) \left[j\tilde{B}(\alpha_{\mu_{\tilde{k}},\delta})^{\left\lceil \frac{(1-1/\gamma)(i^{1-b}/\tilde{B}-2)-2(\tilde{M}_k+1)}{\tilde{M}_k(\tilde{M}_k+1)C\ln(i)} \right\rceil} + C_{\lambda,\mu_{k_2},\mu_{\tilde{k}}} \frac{\exp(-jd_{\lambda,\mu_{k_2},\mu_{\tilde{k}}}/\tilde{B})}{\left\lceil \frac{(1-1/\gamma)(i^{1-b}/\tilde{B}-2)-2(\tilde{M}_k+1)}{\tilde{M}_k(\tilde{M}_k+1)C\ln(i)} \right\rceil^{\lambda}} \right]$$
$$:= g_2(\boldsymbol{\mu}, j, b, i, k, \gamma).$$

This proves that

$$\mathbb{P}(\mathcal{E}_{i,j} \cap F^c_{i,j,m} \cap \{\tau_i^k \leq T\}) \leq (m-1)\left(\frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))} + g_2(\boldsymbol{\mu}, j, b, i, k, \gamma) \right)$$

and the induction is verified.

As for $g_1(\boldsymbol{\mu}, j, b, i, k, \gamma)$, we observe that when $\gamma > 1$, $\sum_{i\leq\infty}\sum_{j\leq\lfloor \tilde{B}i^b \rfloor} g_2(\boldsymbol{\mu}, j, b, i, k, \gamma)$ is finite by choosing $b < 1 - \frac{1}{\lambda}$.

Proof of Lemma 13 It holds that

$$\mathbb{P}\left(\exists \ell \leq i, \exists k' \notin \mathcal{B}_{\mathcal{N}(k)}, \theta_{k'}(\tau_\ell^k) > \mu_{k'} + \delta, N_{k'}(\tau_\ell^k) > C\ln(i), \tau_i^k \leq T\right)$$
$$\leq \sum_{\ell=1}^{i} \sum_{k' \in \mathcal{N}^+(k)\setminus\mathcal{B}_{\mathcal{N}(k)}} \mathbb{P}\left(\theta_{k'}(\tau_\ell^k) > \mu_{k'} + \delta, N_{k'}(\tau_\ell^k) > C\ln(i), \tau_\ell^k \leq T\right)$$

Let $\ell \leq i$, $k' \in \mathcal{N}^+(k) \setminus \mathcal{B}_{\mathcal{N}(k)}$.

$$\mathbb{P}\left(\theta_{k'}(\tau_\ell^k) > \mu_{k'} + \delta, N_{k'}(\tau_\ell^k) > C\ln(i), \tau_\ell^k \leq T\right) \tag{17}$$
$$\leq \mathbb{P}\left(\hat{\mu}_{k'}(\tau_\ell^k) > \mu_{k'} + \delta/2, N_{k'}(\tau_\ell^k) > C\ln(i), \tau_\ell^k \leq T\right) \tag{18}$$
$$+ \mathbb{P}\left(\hat{\mu}_{k'}(\tau_\ell^k) \leq \mu_{k'} + \delta/2, \theta_{k'}(\tau_\ell^k) > \mu_{k'} + \delta, N_{k'}(\tau_\ell^k) > C\ln(i), \tau_\ell^k \leq T\right) \tag{19}$$

Using a union bound over the values of $N_{k'}(\tau_\ell^k) \geq C\ln(i)$ together with Hoeffding's inequality (Lemma 8) yields

$$(18) \leq \sum_{u=C\ln(i)}^{T} \mathbb{P}(\hat{\mu}_{k',u} > \mu_{k'} + \delta/2) \leq \sum_{u=C\ln(i)}^{\infty} \exp\left(-\frac{\delta^2 u}{2}\right) = \frac{\exp(-C\ln(i)\delta^2/2)}{1-\exp(-\delta^2/2)},$$

where we denote by $\hat{\mu}_{k',u}$ the estimated mean of the k'-th arm at the u-th draw.

We upper bound (19) by

$$\begin{aligned}
&\sum_{u=C\ln(i)}^{T} \mathbb{P}\left(\hat{\mu}_{k'}(\tau_\ell^k) \leq \mu_{k'} + \delta/2, \theta_{k'}(\tau_\ell^k) \geq \mu_{k'} + \delta, N_{k'}(\tau_\ell^k) = u, \tau_\ell^k \leq T\right) \\
&\leq \sum_{u=C\ln(i)}^{T} \mathbb{P}\left(\mu_{k'} \geq \hat{\mu}_{k',u} - \delta/2, \theta_{k'}(\tau_\ell^k) \geq \mu_{k'} + \delta, N_{k'}(\tau_\ell^k) = u, \tau_\ell^k \leq T\right) \\
&\leq \sum_{u=C\ln(i)}^{T} \mathbb{P}\left(\theta_{k'}(\tau_\ell^k) \geq \hat{\mu}_{k',u} + \delta/2, N_{k'}(\tau_\ell^k) = u, \tau_\ell^k \leq T\right) \\
&\leq \mathbb{E}\left[\sum_{u=C\ln(i)}^{T} \left(1 - F^{\text{Beta}}_{u\hat{\mu}_{k',u}+1, u-u\hat{\mu}_{k',u}+1}(\hat{\mu}_{k',u} + \delta/2)\right)\right] \\
&= \mathbb{E}\left[\sum_{u=C\ln(i)}^{T} F^{\text{Bin}}_{u+1,\hat{\mu}_{k',u}+\delta/2}(u\hat{\mu}_{k',u})\right] \\
&\leq \mathbb{E}\left[\sum_{u=C\ln(i)}^{T} F^{\text{Bin}}_{u,\hat{\mu}_{k',u}+\delta/2}(u\hat{\mu}_{k',u})\right] \\
&\leq \mathbb{E}\left[\sum_{u=C\ln(i)}^{\infty} \exp(-u\delta^2/2)\right] \\
&= \frac{\exp(-C\ln(i)\delta^2/2)}{1-\exp(-\delta^2/2)},
\end{aligned}$$

where the first equality comes from the Beta-Binomial trick (Lemma 9), and the last inequality comes from Hoeffding's inequality.

Combining (18) and (19), and recalling that $C = 6/\delta^2$, we get

$$\begin{aligned}
&\mathbb{P}\left(\exists \ell \leq i, \exists k' \notin \mathcal{B}_{\mathcal{N}(k)}, \theta_{k'}(\tau_\ell^k) > \mu_{k'} + \delta, N_{k'}(\tau_\ell^k) > C\ln(i), \tau_i^k \leq T\right) \\
&\leq \sum_{\ell=1}^{i} \sum_{k' \in \mathcal{N}^+(k) \setminus \mathcal{B}_{\mathcal{N}(k)}} 2\frac{\exp(-C\ln(i)\delta^2/2)}{1-\exp(-\delta^2/2)} \\
&\leq \frac{2\tilde{M}_k}{i^{C\delta^2/2-1}(1-\exp(-\delta^2/2))} = \frac{2\tilde{M}_k}{i^2(1-\exp(-\delta^2/2))}.
\end{aligned}$$

Proof of Lemma 14 The interval $\tilde{\mathcal{I}}$ is such that for all $\ell \in \tilde{\mathcal{I}}$, $N_{\mathcal{B}_{\mathcal{N}(k)}}(\tau_\ell^k) = j$. This implies that there exists $k_2 \in \mathcal{B}_{\mathcal{N}(k)}$ which has been drawn at least $\frac{j}{\tilde{B}}$ and is not drawn during that interval. Hence,

$$\begin{aligned}
&\mathbb{P}\left(\forall \ell \in \tilde{\mathcal{I}}, N_{\mathcal{B}_{\mathcal{N}(k)}}(\tau_\ell^k) = j, \forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta\right) \\
&\leq \mathbb{P}\left(\forall \ell \in \tilde{\mathcal{I}}, \exists k_2 \in \mathcal{B}_{\mathcal{N}(k)}, \frac{j}{\tilde{B}} \leq N_{k_2}(\tau_\ell^k) \leq j, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta\right) \\
&\leq \sum_{k_2 \in \mathcal{B}_{\mathcal{N}(k)}} \sum_{j_{k_2} = \frac{j}{\tilde{B}}}^{j} \mathbb{P}\left(\forall \ell \in \tilde{\mathcal{I}}, N_{k_2}(\tau_\ell^k) = j_{k_2}, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta\right) \qquad (20)
\end{aligned}$$

If $N_{k_2}(\tau_\ell^k) = j_{k_2}$ for all $\ell \in \tilde{\mathcal{I}}$, conditioned on $S_{k_2,j_{k_2}}$ (sum of first j_{k_2} observations from arm k_2), the Thompson samples of arm k_2 drawn during this interval are an i.i.d. sequence with distribution Beta$(S_{k_2,j_{k_2}} + 1, j_{k_2} - S_{k_2,j_{k_2}} + 1)$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(\forall \ell \in \tilde{\mathcal{I}}, N_{k_2}(\tau_\ell^k) = j_{k_2}, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta | S_{k_2,j_{k_2}}\right) &= \left(F^{\text{Beta}}_{S_{k_2,j_{k_2}}+1, j_{k_2}-S_{k_2,j_{k_2}}+1}(\mu_{\tilde{k}} + \delta)\right)^{|\tilde{\mathcal{I}}|} \\
&\leq \left(F^{\text{Beta}}_{S_{k_2,j_{k_2}}+1, j_{k_2}-S_{k_2,j_{k_2}}+1}(\mu_{\tilde{k}} + \delta)\right)^{x} \\
&= \left(1 - F^{\text{Bin}}_{j_{k_2}+1, \mu_{\tilde{k}}+\delta}(S_{k_2,j_{k_2}})\right)^{x}
\end{aligned}$$

where the inequality holds because $|\tilde{\mathcal{I}}| \geq x$, and the last equality is obtained by using the Beta-Binomial trick (Lemma 9).

It follows that

$$\begin{aligned}
\mathbb{P}\left(\forall \ell \in \tilde{\mathcal{I}}, N_{k_2}(\tau_\ell^k) = j_{k_2}, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta\right) &= \mathbb{E}\left[\mathbb{P}\left(\forall \ell \in \tilde{\mathcal{I}}, N_{k_2}(\tau_\ell^k) = j_{k_2}, \theta_{k_2}(\tau_\ell^k) \leq \mu_{\tilde{k}} + \delta | S_{k_2,j_{k_2}}\right)\right] \\
&\leq \mathbb{E}\left[\left(1 - F^{\text{Bin}}_{j_{k_2}+1, \mu_{\tilde{k}}+\delta}(S_{k_2,j_{k_2}})\right)^{x}\right]
\end{aligned}$$

where the expectation is taken with respect to $S_{k_2,j_{k_2}} \sim \text{Bin}(j_{k_2}, \mu_{k_2})$.

An upper bound on this expectation is provided by the following lemma that can be extracted from the proof of Lemma 3 in Kaufmann et al. (2012).

Lemma 15 *Let X be a random variable with Binomial distribution of parameter (j, μ_1). Let δ and μ_2 be such that $0 < \mu_2 + \delta < \mu_1$. There exists $\lambda_0 = \lambda_0(\mu_1, \mu_2, \delta) > 1$ such that for $\lambda \in (1, \lambda_0)$,*

$$\mathbb{E}\left[\left(1 - F^{Bin}_{j+1,\mu_2+\delta}(X)\right)^{x}\right] \leq (\alpha_{\mu_2,\delta})^x + C_{\lambda,\mu_1,\mu_2} \frac{\exp(-j d_{\lambda,\mu_1,\mu_2})}{x^\lambda}$$

where $C_{\lambda,\mu_1,\mu_2}, d_{\lambda,\mu_1,\mu_2} > 0$, and $\alpha_{\mu_2,\delta} = \left(\frac{1}{2}\right)^{1-\mu_2-\delta}$

Finally,

$$\begin{aligned}
(20) &\leq \tilde{B} \sum_{j_{k_2} = \frac{j}{\tilde{B}}}^{j} \left[(\alpha_{\mu_{\tilde{k}},\delta})^x + \tilde{C}_{\lambda,\mu_{k_2},\mu_{\tilde{k}}} \frac{\exp(-j_{k_2} d_{\lambda,\mu_{k_2},\mu_{\tilde{k}}})}{x^\lambda}\right] \\
&\leq j\tilde{B}(\alpha_{\mu_{\tilde{k}},\delta})^x + C_{\lambda,\mu_{k_2},\mu_{\tilde{k}}} \frac{\exp(-j d_{\lambda,\mu_{k_2},\mu_{\tilde{k}}}/\tilde{B})}{x^\lambda},
\end{aligned}$$

which concludes the proof.

Appendix D. Additional experiments

In this section, we report the results of additional experiments. Section D.1 provides results from experiments carried out in larger environments, whereas Section D.2 presents experiments in environments where the means of the arms are closer to each other.

D.1. Experiments in larger environments

For better comparison with the work of Katariya et al. (2017a), we carry out experiments in the same environments as theirs. That is, we run the algorithm on simulated matrices of arms, for $K = L = 32, 64, 128$, with $\boldsymbol{u} = \boldsymbol{v} = (0.75, 0.25, \dots, 0.25)$, and a horizon $T = 2 \times 10^6$. The shaded areas on our plots show the 10% percentiles.

Additionally, we run the `UTS` algorithm without the warm-up phase where each arm is drawn once at the beginning. Figure 5 shows that removing this warm-up phase is not harmful to the performance of `UTS`, and that in all cases, the unimodal algorithms `UTS` and `OSUB` clearly outperform `Rank1ElimKL` and `KL-UCB`.

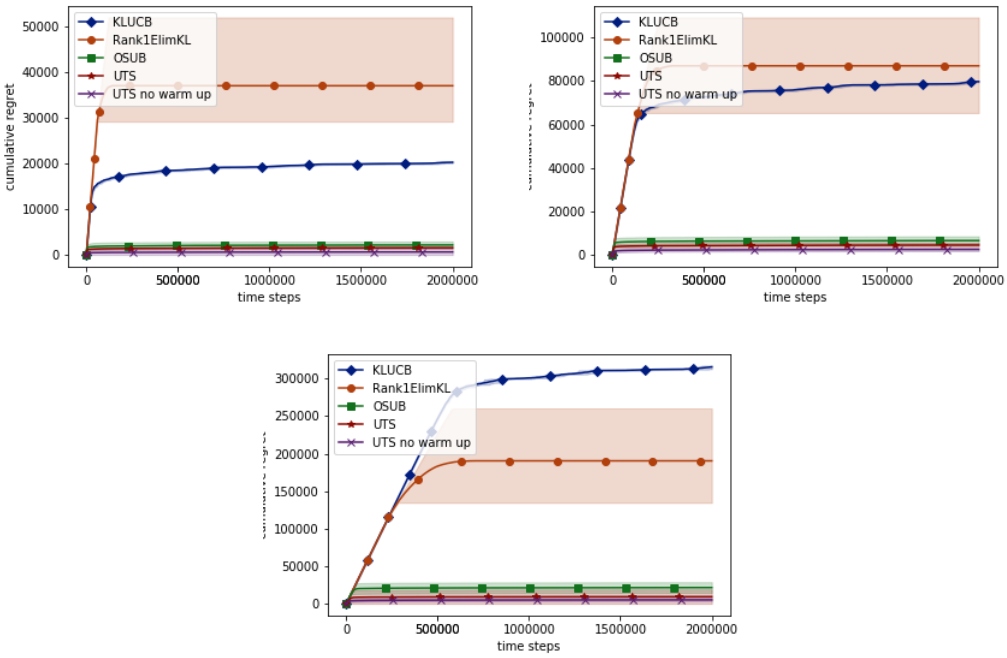

Figure 5: Cumulative regret of `Rank1ElimKL`, `OSUB`, `UTS`, `UTS` without warm-up and `KL-UCB`, on $K \times K$ rank-one matrices with $K = 32$ (top left), $K = 64$ (top right) and $K = 128$ (bottom). Regrets are averaged over 50 runs, except for KLUCB for $K = 128$ which is averaged over 20 runs.

D.2. Experiments with closer means

To conclude with our experiments, we additionally run the algorithms in an environment where the means of the arms are closer to each other. We set $\boldsymbol{u} = \boldsymbol{v} = (0.55, 0.50, \ldots, 0.50)$ and run the experiments for $K = L = 8, 16, 32$.

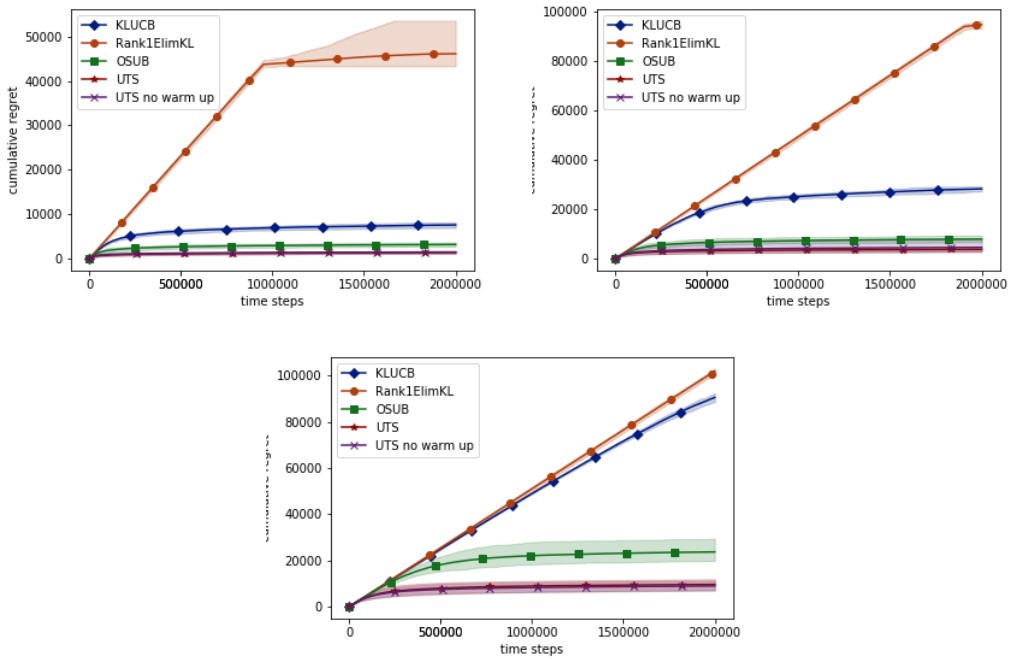

Figure 6: Cumulative regret of `Rank1ElimKL`, `OSUB`, `UTS`, `UTS` without warm-up and `KL-UCB`, on $K \times K$ rank-one matrices with $K = 8$ (top left), $K = 16$ (top right) and $K = 32$ (bottom). Regrets are averaged over 50 runs.