Routine Bandits: Minimizing Regret on Recurring Problems

Hassan Saber, Léo Saci, Odalric-Ambrym Maillard, Audrey Durand

Routine Bandits: Minimizing Regret on Recurring Problems

Hassan Saber[1], Léo Saci[2], Odalric-Ambrym Maillard[1], and Audrey Durand[3]

[1] Université de Lille, Inria, CNRS, Centrale Lille
UMR 9189 – CRIStAL, F-59000 Lille, France
[2] ENS Paris-Saclay, Gif-Sur-Yvette, France
[3] Canada CIFAR AI Chair, Université Laval, Mila

Abstract. We study a variant of the multi-armed bandit problem in which a learner faces every day one of $\mathcal{B}$ many bandit instances, and call it a routine bandit. More specifically, at each period $h \in [\![1, H]\!]$, the same bandit $b^h_\star$ is considered during $T > 1$ consecutive time steps, but the identity $b^h_\star$ is unknown to the learner. We assume all rewards distribution are Gaussian standard. Such a situation typically occurs in recommender systems when a learner may repeatedly serve the same user whose identity is unknown due to privacy issues. By combining bandit-identification tests with a `KLUCB` type strategy, we introduce the `KLUCB` for Routine Bandits (`KLUCB-RB`) algorithm. While independently running `KLUCB` algorithm at each period leads to a cumulative expected regret of $\Omega(H \log T)$ after H many periods when $T \to \infty$, `KLUCB-RB` benefits from previous periods by aggregating observations from similar identified bandits, which yields a non-trivial scaling of $\Omega(\log T)$. This is achieved without knowing which bandit instance is being faced by `KLUCB-RB` on this period, nor knowing a priori the number of possible bandit instances. We provide numerical illustration that confirm the benefit of `KLUCB-RB` while using less information about the problem compared with existing strategies for similar problems.

Keywords: Multi-armed bandits · Transfer Learning · KL-UCB

1 Introduction

The stochastic multi-armed bandit [22,16,5,18], is a popular framework to model a decision-making problem where a learning agent (*learner*) must repeatedly choose between several real-valued unknown sources of random observations (*arms*) to sample from in order to maximize the cumulative values (*rewards*) generated by these choices in expectation. This framework is commonly applied to recommender systems where arms correspond to items (e.g., ads, products) that can be recommended and rewards correspond to the success of the recommendation (e.g., click, buy). An optimal strategy to choose actions would be to always play an arm with highest expected reward. Since the distribution of rewards and in particular their mean are unknown, in practice a learner needs to trade off

exploiting arms that have shown good rewards until now with *exploring* arms to acquire information about the reward distributions. The stochastic multi-armed bandit framework has been well-studied in the literature and optimal algorithms have been proposed [15,23,6,14,13].

When a recommender system is deployed on multiple users, one does not typically assume that the best recommendation is the same for all users. The naive strategy in this situation is to consider each user as being a different bandit instance and learning from scratch for each user. When users can be recognized (e.g., characterized by features), this information can be leveraged to speed up the learning process by sharing observations across users. The resulting setting is known as contextual bandit [17,19]. In this paper, we tackle the case where users cannot be or do not want to be identified (e.g., for privacy reasons), but where we assume that there exists a (unknown) finite set of possible user profiles (bandit instances), such that information may be shared between the current user and some previously encountered users.

Outline and contributions To this end, we introduce the *routine bandit* problem (Sec. 2), together with lower bounds on the achievable cumulative regret that adapt the bound from [16] to the routine setting. We then extend the `KLUCB` [9] algorithm, known to be optimal under the classical stochastic bandit setting, into a new strategy called `KLUCB-RB` (Sec. 3) that leverages the information obtained on previously encountered bandits. We provide a theoretical analysis of `KLUCB-RB` (Sec. 4) and investigate the performance of the algorithm using extensive numerical experiments (Sec. 5). These results highlight the empirical conditions required so that past information can be efficiently leveraged to speed up the learning process. The main contributions of this work are 1) the newly proposed routine bandit setting, 2) the `KLUCB-RB` algorithm that solves this problem with asymptotically optimal regret minimization guarantees, and 3) an empirical illustration of the conditions for past information to be beneficial to the learning agent.

2 The Routine Bandit Setting

A routine bandit problem is specified by a time horizon $T \geqslant 1$ and a finite set of distributions $\nu = (\nu_b)_{b\in\mathcal{B}}$ with means $(\mu_{a,b})_{a\in\mathcal{A},b\in\mathcal{B}}$, where $\mathcal{A}$ is a finite set of arms and $\mathcal{B}$ is a finite set of bandit configurations. Each $b\in\mathcal{B}$ can be seen as a classical multi-armed bandit problem defined by $\nu_b = (\nu_{a,b})_{a\in\mathcal{A}}$. At each period $h \geqslant 1$ and for all time steps $t \in [\![1, T]\!]$, the learner deals with a bandit $b^h_\star \in \mathcal{B}$ and chooses an arm $a^h_t \in \mathcal{A}$, based only on the past. The learner then receives and observes a reward $X^h_t \sim \nu_{a^h_t,b^h_\star}$. The goal of the learner is to maximize the expected sum of rewards received over time (up to some unknown number of periods $H \geqslant 1$). The distributions are unknown, which makes the problem non-trivial. The optimal strategy therefore consists in playing repeatedly on each period h, an optimal arm $a^h_\star \in \operatorname{argmax}_{a\in\mathcal{A}} \mu_{a,b^h_\star}$, which has mean $\mu^h_\star = \mu_{a^h_\star,b^h_\star}$. The goal of the learner is equivalent to minimizing the cumulative *regret* with

respect to an optimal strategy:

$$R(\nu, H, T) = \mathbb{E}_\nu \left[\sum_{h=1}^{H} \sum_{t=1}^{T} \left(\mu_\star^h - X_t^h \right) \right] . \tag{1}$$

Related works One of the closest setting to routine bandits is the sequential transfer scenario [11], where the cardinality $|\mathcal{B}|$ and quantities H and T are known ahead of time, and the instances in $\mathcal{B}$ are either known perfectly or estimated with known confidence. Routine bandits also bear similarity with clustering bandits [10], a contextual bandit setting [17] where contexts can be clustered into finite (unknown) clusters. While both settings are recurring bandit problems, routine bandits assume no information on users (including their number) but users are recurring for several iterations of interaction, while clustering bandits assume that each user is seen only once, but is characterized by features such that they can be associated with previously seen users. Finally, latent bandits [20] consider the less structured situation when the learner faces a possibly different user at every time.

Assumptions and working conditions The configuration ν, the set of bandits $\mathcal{B}$, and the sequence of bandits $(b_\star^h)_{h \geqslant 1}$ are *unknown* (in particular $|\mathcal{B}|$ and the identity of user $b_\star^h$ are unknown to the learner at time t). The learner only knows that $\nu \in \mathcal{D}$, where $\mathcal{D}$ is a given set of bandit configurations. In order to leverage information from the bandit instances encountered, we should consider that bandits reoccur. We denote by $\beta_b^h = \sum_{h'=1}^{h} \mathbb{I}_{\{b_\star^{h'} = b\}} / h$ the frequency of bandit $b \in \mathcal{B}$ at period $h \geqslant 1$ and assume $\beta_b^H > 0$. The next two assumptions respectively allow for two bandit instances b and b' to be distinguishable from their means when $b \neq b'$ and show consistency in their optimal strategy when $b = b'$.

Assumption 1 (Separation)**.** *Let us consider* $\gamma_\nu \coloneqq \min_{b \neq b'} \min_{a \in \mathcal{A}} \{ |\mu_{a,b} - \mu_{a,b'}| , 1 \}$. *We assume* $\gamma_\nu > 0$.

Assumption 2 (Unique optimal arm)**.** *Each bandit* $b \in \mathcal{B}$ *has a unique optimal arm* $a_b^\star$.

Assumption 2 is standard. Finally, we consider normally-distributed rewards. Although most of our analysis (e.g., concentration) would extend to exponential families of dimension 1, Assumption 3 increases readability of the statements.

Assumption 3 (Gaussian arms)**.** *The set* $\mathcal{D}$ *is the set of bandit configurations such that for all bandit* $b \in \mathcal{B}$*, for all arm* $a \in \mathcal{A}$*,* $\nu_{a,b}$ *is a one-dimensional Gaussian distribution with mean* $\mu_{a,b} \in \mathbb{R}$ *and variance* $\sigma^2 = 1$.

For $\nu \in \mathcal{D}$, we define for an arm $a \in \mathcal{A}$ and a bandit $b \in \mathcal{B}$ their gap $\Delta_{a,b} = \mu_b^\star - \mu_{a,b}$ and their total number of pulls over H periods $N_{a,b}(H,T) = \sum_{h=1}^{H} \sum_{t=1}^{T} \mathbb{I}_{\{a_t^h = a, b_\star^h = b\}}$. An arm is optimal for a bandit if their gap is equal to zero and sub-optimal if it is positive. Thanks to the chain rule, the regret rewrites as

$$R(\nu, H, T) = \sum_{b \in \mathcal{B}} \sum_{a \neq a_b^\star} \mathbb{E}_\nu [N_{a,b}(H,T)] \, \Delta_{a,b} \, . \tag{2}$$

Remark 1 (Fixed horizon time). We assume the time horizon T to be the same for all periods $h \in [\![1\,,H]\!]$ out of clarity of exposure of the results and simplified definition of consistency (Definition 1). Considering a different time T_h for each h would indeed require a substantial rewriting of the statements (e.g. think of the regret lower bound), which we believe hinders readability and comparison to classical bandits.

We conclude this section by adapting for completeness the known lower bound on the regret [16,2,12] for *consistent* strategies to the routine bandit setting. We defer the proof to Appendix A.

Definition 1 (Consistent strategy). *A strategy is H-consistent on $\mathcal{D}$ if for all configuration $\nu \in \mathcal{D}$, for all bandit $b \in \mathcal{B}$, for all sub-optimal arm $a \neq a^b_\star$, for all $\alpha > 0$,*

$$\lim_{T\to\infty} \mathbb{E}_\nu\left[\frac{N_{a,b}(H,T)}{N_b(H,T)^\alpha}\right] = 0\,,$$

where $N_b(H,T) = \beta_b^H HT$ is the number of time steps the learner has dealt with bandit b.

Proposition 1 (Lower bounds on the regret). *Let us consider a consistent strategy. Then, for all configuration $\nu \in \mathcal{D}$, it must be that*

$$\liminf_{T\to\infty} \frac{R(\nu,H,T)}{\log(T)} \geqslant c^\star_\nu \coloneqq \sum_{b\in\mathcal{B}} \sum_{a\neq a^\star_b} \frac{\Delta_{a,b}}{\mathrm{KL}(\mu_{a,b}|\,\mu^\star_b)}\,,$$

where $\mathrm{KL}(\mu|\,\mu') = (\mu'-\mu)^2/2\sigma^2$ denotes the Kullback-Leibler divergence between one-dimensional Gaussian distributions with means $\mu, \mu' \in \mathbb{R}$ and variance $\sigma^2 = 1$.

This lower bound differs (it is larger) from structured lower bound that can exclude some set of arms, as in [2,20] using prior knowledge on $\mathcal{B}$, which here is not available. On the other hand, we remark that the right hand side of the bound does not depend on H, which suggests that one at least asymptotically, one can learn from the recurring bandits. In the classical bandit setting, lower bounds on the regret [16] have inspired the design of the well-known `KLUCB` [9] algorithm. In the next section, we build on this optimal strategy to propose a variant for the routine bandit.

3 The `KLUCB-RB` Strategy

Given the current period h, the general idea of this optimistic strategy consists in aggregating observations acquired in previous periods $1 \ldots h-1$ where bandit instances are tested to be the same as the current bandit $b^h_\star$. To achieve this, `KLUCB-RB` relies both on concentration of observations gathered in previous periods and the consistency of the allocation strategy between different periods.

Notations The number of pulls, the sum of the rewards and the empirical mean of the rewards from the arm a in period $h \geqslant 1$ at time $t \geqslant 1$, are respectively denoted by $N_a^h(t) = \sum_{s=1}^t \mathbb{I}_{\{a_s^h = a\}}$, $S_a^h(t) = \sum_{s=1}^t \mathbb{I}_{\{a_s^h = a\}} X_s^h$ and $\widehat{\mu}_a^h(t) = S_a^h(t)/N_a^h(t)$ if $N_a^h(t) > 0$, 0 otherwise.

Strategy For each period $h \geqslant 1$ we compute an empirical best arm for bandit $b_\star^h$ as the arm with maximum number of pulls in this period: $\overline{a}_\star^h \in \operatorname*{argmax}_{a \in \mathcal{A}} N_a^h(T)$.[4] Similarly, in the current period $h \geqslant 1$, for each time step $t \in [\![1, T]\!]$, we consider an arm with maximum number of pulls: $\overline{a}_t^h \in \operatorname*{argmax}_{a \in \mathcal{A}} N_a^h(t)$ (arbitrarily chosen). At each period $h \in [\![2, H]\!]$ each arm is pulled once. Then at each time step $t \geqslant |\mathcal{A}| + 1$, in order to possibly identify the current bandit $b_\star^h$ with some bandits $b_\star^k$ from a previous period $k \in [\![1, h-1]\!]$, we introduce for all arm $a \in \mathcal{A}$, the test statistics

$$Z_a^{k,h}(t) = \infty \cdot \mathbb{I}_{\left\{\overline{a}_t^h \neq \overline{a}_\star^k\right\}} + \left|\widehat{\mu}_a^h(t) - \widehat{\mu}_a^k(T)\right| - \mathrm{d}\big(N_a^h(t), \delta^h(t)\big) - \mathrm{d}\big(N_a^k(T), \delta^h(t)\big)\,, \quad (3)$$

where the deviation for $n \geqslant 1$ pulls with probability $1-\delta$, for $\delta > 0$, and probability $\delta^h(t)$ are, respectively,

$$\mathrm{d}(n, \delta) = \sqrt{2\left(1 + \frac{1}{n}\right)\frac{\log\left(\sqrt{n+1}/\delta\right)}{n}} \qquad \delta^h(t) = \frac{1}{4\,|\mathcal{A}|} \times \frac{1}{h-1} \times \frac{1}{t(t+1)}\,.$$

The algorithm finally computes the test

$$\mathrm{T}^{k,h}(t) \coloneqq \max_{a \in \mathcal{A}} Z_a^{k,h}(t) \leqslant 0\,. \quad (4)$$

After t rounds in current period h, the previous bandit $b_\star^k$ is suspected of being the same as $b_\star^h$ if the test $\mathrm{T}^{k,h}(t)$ is true. From Eq. 3, we note that this requires the current mostly played arm to be the same as the arm that was mostly played in period k, which happens if there is consistency in the allocation strategy for both periods under Assumption 2. We then define aggregated numbers of pulls and averaged means: For all arm $a \in \mathcal{A}$, for all period $h \geqslant 1$, for all time step $t \geqslant 1$,

$$\overline{N}_a^h(t) \coloneqq N_a^h(t) + \sum_{k=1}^{h-1} \mathbb{I}_{\left\{\mathrm{T}^{k,h}(t)\right\}} N_a^k(T), \quad \overline{K}_t^h \coloneqq \sum_{k=1}^{h-1} \mathbb{I}_{\left\{\mathrm{T}^{k,h}(t)\right\}},$$
$$\overline{S}_a^h(t) \coloneqq S_a^h(t) + \sum_{k=1}^{h-1} \mathbb{I}_{\left\{\mathrm{T}^{k,h}(t)\right\}} S_a^k(T), \quad \overline{\mu}_a^h(t) = \overline{S}_a^h(t)/\overline{N}_a^h(t)\,.$$

and follow a `KLUCB` strategy by defining the index of arm $a \in \mathcal{A}$ in period $h \geqslant 1$ at time step $t \geqslant 1$ as

$$u_a^h(t) = \min\left\{U_a^h(t)\,, \overline{U}_a^h(t)\right\}, \quad (5)$$

[4] ties are broken arbitrarily

where

$$U_a^h(t) \coloneqq \widehat{\mu}_a^h(t) + \sqrt{\frac{2f(t)}{N_a^h(t)}}\,, \tag{6}$$

$$\overline{U}_a^h(t) \coloneqq \overline{\mu}_a^h(t) + \sqrt{\frac{2f\left(\overline{K}_t^h T + t\right)}{\overline{N}_a^h(t)}}\,, \tag{7}$$

with the function f being chosen, following [6] for classical bandits, as

$$f(x) \coloneqq \log(x) + 3\log\log(\max\{e, x\})\,, \forall x \geqslant 1\,.$$

One recognizes that Eq. 6 corresponds to the typical `KLUCB` upper bound for Gaussian distributions. The resulting `KLUCB-RB` strategy is summarized in Algorithm 1.

Algorithm 1 `KLUCB-RB`

Initialization (period $h=1$): follow a `KLUCB` strategy for bandit $b_\star^1$.
for period $h \geqslant 2$ **do**
  Pull each arm once
  for time step $t \in [\![|\mathcal{A}|\,, T-1]\!]$ **do**
    Compute for each previous period $k \in [\![1, h-1]\!]$ the test $\mathrm{T}^{k,h}(t) \coloneqq \max_{a\in\mathcal{A}} Z_a^{k,h}(t) \leqslant 0$
    Aggregate data from periods with positive test and compute for each arm $a \in \mathcal{A}$ the index $u_a^h(t)$ according to equations (5)-(6)-(7).
    Pull an arm with maximum index $a_{t+1}^h \in \operatorname{argmax}_{a\in\mathcal{A}} u_a^h(t)$
  end for
end for

Theoretical guarantees The next result shows that the number of sub-optimal pulls done by `KLUCB-RB` is upper-bounded in a near-optimal way.

Theorem 1 (Upper bounds). *Let us consider a routine bandit problem specified by a set of Gaussian distributions $\nu \in \mathcal{D}$ and a number of periods $H \geqslant 1$. Then under `KLUCB-RB` strategy, for all $0 < \varepsilon < \varepsilon_\nu$, for all bandit $b \in \mathcal{B}$, for all sub-optimal arm $a \neq a_b^\star$,*

$$\mathbb{E}_\nu[N_{a,b}(H,T)] \leqslant \frac{f(\beta_b^H HT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon\,|\,\mu_b^\star)} + \sum_{h=1}^{H} \mathbb{I}_{\{b_\star^h = b\}}\left[\tau_\nu^h + 4\,|\mathcal{A}|\left(\frac{1}{\varepsilon^2}+1\right)\left(5+\frac{8h\,f(hT)}{T\,\mathrm{KL}(\mu_{a,b}+\varepsilon\,|\,\mu_b^\star)}\right)\right],$$

where, for all period $h \geqslant 2$, $\tau_\nu^h \coloneqq 2\varphi\big(8|\mathcal{A}|\big[\varepsilon_\nu^{-2} + 65\gamma_\nu^{-2}\log\big(128|\mathcal{A}|(4h)^{1/3}\gamma_\nu^{-2}\big)\big]\big)$, $\varphi\colon x \geqslant 1 \mapsto x\log(x)$, $\varepsilon_\nu = \min_{b\in\mathcal{B}} \min_{a\neq a_b^\star} \Delta_{a,b}/2$ and $\gamma_\nu = \min_{b\neq b'} \min_{a\in\mathcal{A}} \{|\mu_{a,b} - \mu_{a,b'}|\,, 1\}$.

This implies that the dependency on the time horizon T in these upper bounds is asymptotically optimal with regard to the lower bound on the regret given in Proposition 1. From Eq. 2, by considering the case when the time horizon T tends to infinity, we deduce that KLUCB-RB achieves asymptotic optimality.

Corollary 1 (Asymptotic optimality). *With the same notations and under the assumptions as in Theorem 1, KLUCB-RB achieves*

$$\limsup_{T\to\infty} \frac{R(\nu, H, T)}{\log(T)} \leqslant c^\star_\nu \,,$$

where $c^\star_\nu$ is defined as in Proposition 1.

For comparison, let us remark that under the strategy that runs a separate KLUCB type strategy for each period, the regret normalized by $\log(T)$ asymptotically scales as $H \sum_{b\in\mathcal{B}} \beta_b^H \sum_{a\neq a_b^\star} \Delta_{a,b}/\mathrm{KL}(\mu_{a,b}|\,\mu_b^\star)$. KLUCB-RB strategy then performs better than this naive strategy by a factor of the order of $H/|\mathcal{B}|$. Also, up to our knowledge, this result is the first showing provably asymptotic optimal regret guarantee in a setting when an agent attempts at transferring information from past to current bandits without contextual information. In the related but different settings considered in [11,10,20], only logarithmic regret was shown, however asymptotic optimality was not proved for the considered strategies. Also, let us remind that $|\mathcal{B}|$ does not need to be known ahead of time by the KLUCB-RB algorithm.

4 Sketch of Proof

This section contains a sketch of proof for Theorem 1. We refer to Appendix B for more insights and detailed derivations. The first preoccupation is to ensure that KLUCB-RB is a consistent strategy. This is achieved by showing that KLUCB-RB aggregates observations that indeed come from the same bandits with high probability. In other words, we want to control the number of previously encountered bandits falsely identified as similar to the current one.

Definition 2 (False positive). *At period $h\geqslant 2$ and step $t\geqslant 1$, a previous period $k\in[\![1,h-1]\!]$ is called a false positive if the test $\mathrm{T}^{k,h}(t)$ is true while previous bandit $b^k_\star$ differs from current bandit $b^h_\star$.*

Combining the triangle inequality and time-uniform Gaussian concentration inequalities (see e.g., [1]), we prove necessary condition for having $Z^{k,h}_a(t)\leqslant 0$ for some arm $a\in\mathcal{A}$ at current period h and time step t, while having $b^k_\star \neq b^h_\star$.

Lemma 1 (Condition for false positives). *If there exists a* false positive *at period $h\geqslant 2$ and time step $t>|\mathcal{A}|$, then with probability $1-1/t(t+1)$, it must be that*

$$\min_{k\in[\![1,h-1]\!]:b^k_\star\neq b^h_\star} \min_{a\in\mathcal{A}} \left|\mu_{a,b^h_\star} - \mu_{a,b^k_\star}\right| \leqslant 4\ \mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right) .$$

The proof of this key result is provided in Appendix B.1. It relies on time-uniform concentration inequalities. We now introduce a few quantities.

Let us first consider at period $h \geqslant 2$ the time step

$$t_\nu^h \coloneqq \max\left\{t \geqslant |\mathcal{A}| :\ \gamma_\nu \leqslant 4\ \mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right)\right\} + 1\,, \tag{8}$$

beyond which there is no false positives with high probability. We define for all $a \neq a_\star^h$, for all $0 < \varepsilon < \varepsilon_\nu \coloneqq \min_{b\in\mathcal{B}} \min_{a\neq a_b^\star}\{\Delta_{a,b}, 1\}/2$ the subsets of times when there is a false positive

$$\mathcal{T}_a^h \coloneqq \left\{t \geqslant t_\nu^h :\ a_{t+1}^h = a \text{ and } \mathcal{K}_+^h(t) \neq \mathcal{K}_\star^h(t)\right\} \qquad \mathcal{T}^h \coloneqq \bigcup_{a\neq a_\star^h} \mathcal{T}_a^h\,, \tag{9}$$

where we introduced for convenience the sets $\mathcal{K}_+^h \coloneqq \left\{k \in [\![1, h-1]\!] : \mathrm{T}^{k,h}(t) \text{ is true}\right\}$ and $\mathcal{K}_\star^h(t) \coloneqq \left\{k \in [\![1, h-1]\!] : b_\star^k = b_\star^h \text{ and } \overline{a}_t^h = \overline{a}_\star^k\right\}$. We also consider the times when the mean of the current pulled arm is poorly estimated or the best arm $a_\star^h$ is below its mean (either for the current period or by aggregation) and define

$$\mathcal{C}_{a,\varepsilon}^h \coloneqq \left\{t \geqslant 1 : a_{t+1}^h = a \text{ and } \left(\left|\widehat{\mu}_a^h(t) - \mu_a^h\right| > \varepsilon \text{ or } u_{a_\star^h}^h(t) = U_{a_\star^h}^h(t) < \mu_\star^h\right)\right\}$$

$$\mathcal{C}_\varepsilon^h \coloneqq \bigcup_{a\neq a_\star^h} \mathcal{C}_{a,\varepsilon}^h \tag{10}$$

$$\overline{\mathcal{C}}_{a,\varepsilon}^h \coloneqq \mathcal{T}_a^h \cup \left\{t \geqslant t_\nu^h : t \notin \mathcal{T}^h, a_{t+1}^h = a \text{ and } \left(\left|\overline{\mu}_a^h(t) - \mu_a^h\right| > \varepsilon \text{ or } u_{a_\star^h}^h(t) = \overline{U}_{a_\star^h}^h(t) < \mu_\star^h\right)\right\}$$

$$\overline{\mathcal{C}}_\varepsilon^h \coloneqq \bigcup_{a\neq a_\star^h} \overline{\mathcal{C}}_{a,\varepsilon}^h\,. \tag{11}$$

The size of these (bad events) sets can be controlled by resorting to concentration arguments. The next lemma borrows elements of proof from [7] for the estimation of the mean of current pulled arm and [6] for the effectiveness of the upper confidence bounds on the empirical means of optimal arms. We adapt these arguments to the routine-bandit setup, and provide additional details in the appendix.

Lemma 2 (Bounded subsets of times). *For all period $h \geqslant 2$, for all arm $a \in \mathcal{A}$, for all $0 < \varepsilon < \varepsilon_\nu$,*

$$\mathbb{E}_\nu\left[\left|\mathcal{T}^h\right|\right] \leqslant 1 \qquad \mathbb{E}_\nu\left[\left|\mathcal{C}_{a,\varepsilon}^h\right|\right] \leqslant 4\varepsilon^{-2} + 2 \qquad \mathbb{E}_\nu\left[\left|\overline{\mathcal{C}}_{a,\varepsilon}^h\right|\right] \leqslant 4\varepsilon^{-2} + 3\,.$$

By definition of the index (Eq. 7), we have

$$\begin{aligned} \forall t > |\mathcal{A}|\,, \quad & N_a^h(t)\mathrm{KL}\left(\widehat{\mu}_a^h(t)\middle|\, U_a^h(t)\right) = f(t) \\ & \overline{N}_a^h(t)\mathrm{KL}\left(\overline{\mu}_a^h(t)\middle|\, \overline{U}_a^h(t)\right) = f\left(\overline{K}_t^h T + t\right). \end{aligned}$$

We then provide logarithmic upper bounds on the aggregated number of pulls $\overline{N}_a^h(t)$ to deduce the consistency of `KLUCB-RB` strategy. The following non-trivial result combines standard techniques with the key mechanism of the algorithm.

Lemma 3 (Consistency). *Under* `KLUCB-RB` *strategy for all period* $h \geqslant 2$*, for all* $0 < \varepsilon < \varepsilon_\nu$*, for all sub-optimal arm* $a \neq a^h_\star$*, for all* $t > |\mathcal{A}|$ *such that* $a^h_{t+1} = a$*,*

$$\text{if } t \notin \mathcal{C}^h_{a,\varepsilon},\ N^h_a(t) \leqslant \frac{f(t)}{\mathrm{KL}(\mu^h_a + \varepsilon | \, \mu^h_\star)}, \quad \text{if } t \geqslant t^h_\nu \text{ and } t \notin \overline{\mathcal{C}}^h_{a,\varepsilon}, \overline{N}^h_a(t) \leqslant \frac{f\Big(\underline{K}^h_t T + t\Big)}{\mathrm{KL}(\mu^h_a + \varepsilon | \, \mu^h_\star)},$$

where $\underline{K}^h_t \coloneqq \min\Big\{\overline{K}^h_t, \beta^{h-1}_{b^h_\star}(h-1)\Big\}$*. In particular this implies*

$$\forall t \geqslant 1, \forall a \neq a^h_\star, \quad N^h_a(t) \leqslant \frac{f(t)}{\mathrm{KL}(\mu^h_a + \varepsilon | \, \mu^h_\star)} + \big|\mathcal{C}^h_{a,\varepsilon}\big| + N^h_a(|\mathcal{A}|+1)\ ,$$

where $N^h_a(|\mathcal{A}|+1) \leqslant 2$ *and* $\mathbb{E}_\nu\big[\big|\mathcal{C}^h_{a,\varepsilon}\big|\big] \leqslant 4\varepsilon^{-2} + 2$*.*

Thanks to Eq. 5 that involves the minimum of the aggregated index $\overline{U}^h_a(t)$ on past episodes and (not aggregated) indexes $U^h_a(t)$ for the current epoch, the proof proceeds by considering the appropriate sets of time, namely $t \notin \mathcal{C}^h_{a,\varepsilon}$ or $t \notin \overline{\mathcal{C}}^h_{a,\varepsilon}$ depending on the situation. In particular, we get for the considered a that the maximum index $u^h_a(t)$ is either greater than $u^h_{a^h_\star}(t) = U^h_{a^h_\star}(t)$ or $u^h_{a^h_\star}(t) = \overline{U}^h_{a^h_\star}(t)$, which in turns enable to have a control either on $N^h_a(t)$ or $\overline{N}^h_a(t)$. In order to obtain the last statement, it essentially remains to consider the maximum time $t' \in [\![|\mathcal{A}|+1; t]\!]$ such that $a^h_{t'+1} = a$ and $t' \notin \mathcal{C}^h_{a,\varepsilon}$.

In order to be asymptotically optimal (in the sense of Corollary 1), the second preoccupation is to ensure with high probability that we aggregate all of the observations coming from current bandit $b^h_\star$ when computing the indexes. From the definition of $\mathcal{T}^h$ (Eq. 9) and Lemma 2, this amounts to ensure that the current most pulled arm and the most pulled arms of previous periods are the optimal arms of the corresponding periods with high probability. By using the consistency of `KLUCB-RB`, we prove necessary conditions for the most pulled arms being different from the optimal ones.

Lemma 4 (Most pulled arms). *For all period* $h \geqslant 2$*, for all* $0 < \varepsilon < \varepsilon_\nu$*, for all* $t \geqslant t^h_\nu$ *such that* $t \notin \mathcal{T}^h$ *and* $\overline{a}^h_t \neq a^h_\star$ *,*

$$\frac{t + |\mathcal{K}^h_\star(t)|T}{2} - \Big(f(t) + \Big|\mathcal{K}^h_\star(t)\Big| f(T)\Big) \sum_{a \neq a^h_\star} \frac{1}{\mathrm{KL}\big(\mu^h_a + \varepsilon | \, \mu^h_\star\big)} - \Big(1 + \Big|\mathcal{K}^h_\star(t)\Big|\Big)|\mathcal{A}| \leqslant \sum_{k \in \mathcal{K}^h_\star(t) \cup \{h\}} \Big|\mathcal{C}^k_\varepsilon\Big|\,.$$

Let us remind that $\mathcal{K}^h_\star(t)$, defined after Lemma 1, counts the previous phases before h facing the same bandit as the current one, and for which the most-played arm until then agree. Then, by combining Lemma 3 and Lemma 4 we obtain randomized upper bounds on the number of pulls of sub-optimal arms.

Proposition 2 (Randomized upper bounds). *Under* `KLUCB-RB` *strategy, for all bandit* $b \in \mathcal{B}$*, for all sub-optimal arm* $a \neq a^\star_b$*, for all* $0 < \varepsilon < \varepsilon_\nu$*,*

$$N_{a,b}(H,T) \leqslant \frac{f(\beta^H_b HT)}{\mathrm{KL}(\mu_{a,b} + \varepsilon | \, \mu^\star_b)} + \sum_{h=1}^{H} \mathbb{I}_{\{b^h_\star = b\}} \left[T^h_{\nu,\varepsilon} + 4\big|\mathcal{C}^h_\varepsilon\big| + \Big|\overline{\mathcal{C}}^h_\varepsilon\Big| + \frac{f(hT)}{\mathrm{KL}(\mu_{a,b} + \varepsilon | \, \mu^\star_b)} \sum_{k=1}^{h} \frac{8\big|\mathcal{C}^k_\varepsilon\big|}{T} + \mathbb{I}_{\{T \in \mathcal{T}^k\}} \right],$$

where $T^h_{\nu,\varepsilon} \coloneqq \max\left\{t \geqslant t^h_\nu \colon \frac{t}{4} - \sum_{a \neq a^h_\star} \frac{f(t)}{\mathrm{KL}(\mu^h_a + \varepsilon | \, \mu^h_\star)} \leqslant |\mathcal{A}|\right\} + 1$ *for* $h \geqslant 2$, *with* t^h_ν *defined in Eq.* (8).

We prove Theorem 1 by averaging the randomized upper bounds from Proposition 2.

5 Numerical Experiments

We now perform experiments to illustrate the performance of the proposed `KLUCB-RB` under different empirical conditions. We compare `KLUCB-RB` with a baseline strategy which consists in using a `KLUCB` that restarts from scratch at every new period, that is the default strategy when no information (features) is provided to share information across periods. We also include a comparison with the sequential transfer algorithm `tUCB` [11] which constitutes interesting baseline to compare with, since it transfers the knowledge of past periods to minimize the regret in a very similar context. Through the periods $h \in [\![1, H]\!]$, `tUCB` incrementally estimates the mean vectors by the Robust Tensor Power method [3,4], then yielding a deviation of rate $\mathcal{O}(1/\sqrt{h})$ over the empirical means. Thus, it needs to know in advance the total number of instances $|\mathcal{B}|$. Besides the RTP method requires the mean vectors to be linearly independent mutually, which forces the number of arms $|\mathcal{A}|$ to be larger than $|\mathcal{B}|$, while `KLUCB-RB` can tackle this kind of distributions. The next comparisons between `KLUCB-RB` and `tUCB` will mainly illustrate the ability of the former to make large profits from the very first periods, while the later needs to get a sufficiently high confidence over the models estimates before beginning to use knowledge from the previous periods.

All experiments are repeated 100 times. Sequence $(b^h)_{1 \leqslant h \leqslant H}$ is chosen randomly each time. All the different strategies are compared based on their cumulative regret (Eq. 1). Additional experiments are provided in Appendix C.

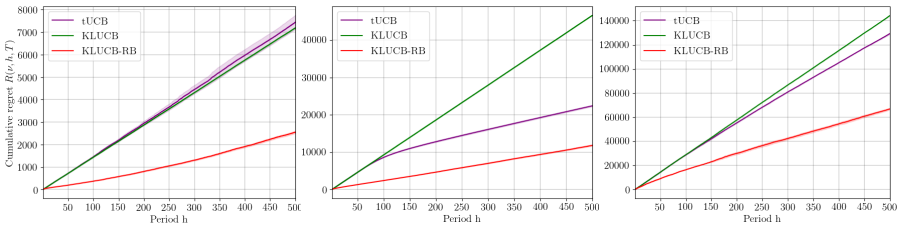
5.1 More Arms than Bandits: A Beneficial Case

We first investigate how Assumption 1 can be relaxed in practice. Indeed `KLUCB-RB` is designed such that only data from previous periods $k < h$ for which the most pulled arm $\bar{a}^k_\star$ is the same as the current most pulled arm $\bar{a}^h_t$ may be aggregated. Consequently, let us define $\gamma^\star_\nu \coloneqq \min_{b \neq b'} \min_{a \in \mathcal{A}^\star} |\mu_{a,b} - \mu_{a,b'}|$ with $\mathcal{A}^\star$ being the set of arms optimal on at least one instance $b \in \mathcal{B}$. Assuming that `KLUCB-RB` converges to the optimal action in a given period, it is natural in practice to relax Assumption 1 from $\gamma_\nu > 0$ to $\gamma^\star_\nu > 0$. Let us consider a routine two-bandit setting $\mathcal{B} = \{b_1, b_2\}$ with actions $\mathcal{A}$ such that

$$b_1 : (\mu_{1,b_1}, \mu_{2,b_1}) = (\frac{\Delta}{2}, -\frac{\Delta}{2}) \qquad \text{and} \qquad \forall a \geqslant 3,\ \mu_{a,b_1} = \mu \tag{12}$$

$$b_2 : (\mu_{1,b_2}, \mu_{2,b_2}) = (\frac{\Delta}{2} - \gamma, -\frac{\Delta}{2} + \gamma) \qquad \text{and} \qquad \forall a \geqslant 3,\ \mu_{a,b_2} = \mu, \tag{13}$$

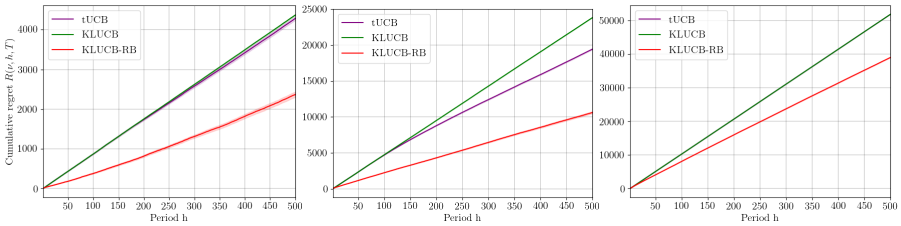
with $\mu = -\frac{\Delta}{2}$, and $\gamma = 0.85\Delta$, and where $\Delta = 10\sqrt{\frac{\log(HT)}{T}}$ is set to accomodate the convergence of KLUCB in the experiment. Note that Assumption 1 is not satisfied anymore since $\gamma_\nu = 0$, but that $\gamma_\nu^\star = \gamma$. Fig. 1 shows the average cumulative regret with one standard deviation after $H = 500$ periods of $T = 10^3$ rounds on settings where $|\mathcal{A}^\star| = 2$ and $|\mathcal{A}| \geqslant 2$.



(a) $|\mathcal{A}| = 2$ (b) $|\mathcal{A}| = 10$ (c) $|\mathcal{A}| = 50$

Fig. 1: Cumulative regret of KLUCB, KLUCB-RB and tUCB along $H = 500$ periods of $\boldsymbol{T = 10^3}$ **rounds**, for different action sets.

We observe that KLUCB-RB can largely benefit from relying on previous periods when the number of arms exceeds the number of optimal arms, which naturally happens when $|\mathcal{A}| > |\mathcal{B}|$. This can also happen for $|\mathcal{A}| \leqslant |\mathcal{B}|$ if several bandits $b \in \mathcal{B}$ share the same optimal arm. Besides, Fig. 2 shows a remake of the same experiment, that is $\Delta = 10\sqrt{\frac{\log(H \times 10^3)}{10^3}}$, where the number of rounds per period is decreased from 10^3 to $T = 100$. We can see that KLUCB-RB still yields good satisfying performances, although T is not large enough to enable a sure identification at each period of the current instance.



(a) $|\mathcal{A}| = 2$ (b) $|\mathcal{A}| = 10$ (c) $|\mathcal{A}| = 50$

Fig. 2: Cumulative regret of KLUCB, KLUCB-RB and tUCB along $H = 500$ periods of $\boldsymbol{T = 100}$ **rounds**, for different action sets.

5.2 Increasing the Number of Bandit Instances

We now consider experiments where we switch among $|\mathcal{B}| = 5$ four-armed bandits. This highlights the kind of settings which may cause more difficulties to KLUCB-RB in distinguishing the different instances: the lesser is the number of arms $|\mathcal{A}|$ compared to the number of bandits $|\mathcal{B}|$, the harder it should be for KLUCB-RB to distinguish efficiently the different instances, in particular when the separation gaps are tight. Let us precise that tUCB cannot be tested on such settings, where the number of models $|\mathcal{B}|$ exceeds the number of arms $|\mathcal{A}|$, since it requires that the mean vectors $(\mu_{a,b})_{a\in\mathcal{A}}$ for all b in $\mathcal{B}$ to be linearly independent.

Generating specific settings is far more complicated here than in cases where $|\mathcal{B}| = 2$ because of the intrinsic dependency between regret gaps $(\Delta_{a,b})_{a\in\mathcal{A},b\in\mathcal{B}}$ and separation gaps $(|\mu_{a,b} - \mu_{a,b'}|)_{a\in\mathcal{A},b\neq b'}$. Thus, distributions of bandits $\nu \in \mathcal{D}$ used in the next experiments are generated randomly so that some conditions are satisfied (see Eq. 14, 15). Recall that $\nu : (\nu_{b_1}, \ldots, \nu_{b_{|\mathcal{B}|}})$ is the set of bandit configurations in the bandit set $\mathcal{B}$. We consider two different distributions $\nu^{(1)}$ and $\nu^{(2)}$, resulting in associated sets of bandits $\mathcal{B}_1$ and $\mathcal{B}_2$, satisfying the condition $C(\nu)$ in order to ensure the convergence of algorithms at each period:

$$C(\nu) : \forall b \in \mathcal{B}, \quad 8\sqrt{\frac{\log(HT)}{T}} \leqslant \min_{a\neq a_b^\star} \Delta_{a,b} \leqslant 12\sqrt{\frac{\log(HT)}{T}}. \tag{14}$$

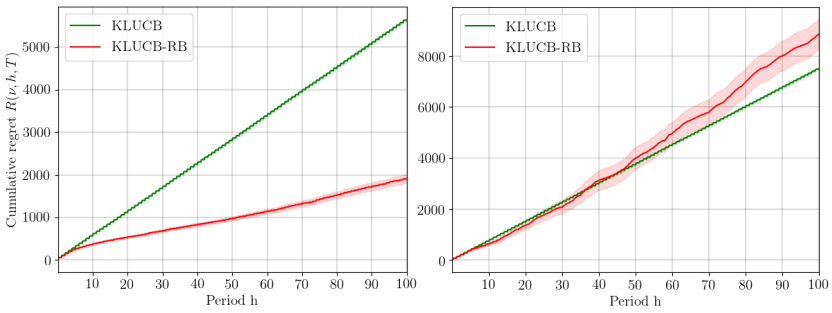
Let $\gamma(\alpha) \coloneqq \alpha\sqrt{\frac{\log(HT)}{T}}$. We generate two sets of bandits $\mathcal{B}_1$ and $\mathcal{B}_2$ such as to ensure that $\nu^{(1)}$ and $\nu^{(2)}$ satisfy

$$\gamma(12) \leqslant \gamma^\star_{\nu^{(1)}} \leqslant \gamma(16) \qquad \gamma(4) \leqslant \gamma^\star_{\nu^{(2)}} \leqslant \gamma(8). \tag{15}$$

Fig. 7 (Appendix C.3) shows the bandit instances in the two generated bandit sets.

All experiments are conducted under the fair frequency $\beta = 1/|\mathcal{B}|$. More precisely, once a period $h \geqslant 1$ ends, $b_\star^{h+1}$ is sampled uniformly in $\mathcal{B}$ and independently of the past sequence $(b_\star^k)_{1\leqslant k\leqslant h}$. Fig. 3 shows the average cumulative regret with one standard deviation after $H = 100$ periods of $T = 5000$ rounds for the two settings.

We observe that the performance of KLUCB-RB is tied to the smallest sub-optimal gap for all bandit instances. Fig. 3a highlights that KLUCB-RB outperforms KLUCB if the minimal sub-optimal gap of each bandit is less than the characteristic smaller separation gap $\gamma^\star_\nu$. This supports the observation from Sec. 5.1 that separation on optimal arms is sufficient. When arms are easier to separate than bandits, one might as well restart a classical KLUCB from scratch on each period (Fig. 3b). Note that situations where $0 < \gamma_\nu \ll \min_{b\in\mathcal{B}} \min_{a\neq a_b^\star} \Delta_{a,b}$ may not result in a catastrophic loss in learning performances if the arms in $\mathcal{A}^\star$ are *close enough* not to distort estimates computed on aggregated samples of from false positive models (see Appendix C).

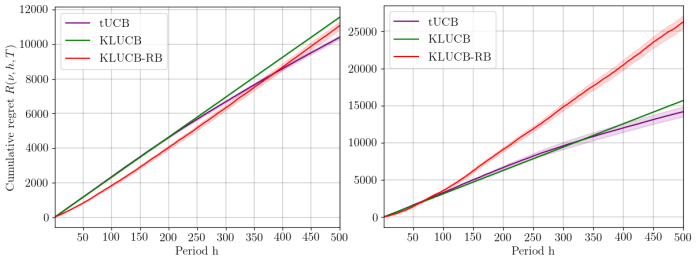


(a) Setting $\nu^{(1)}$ (b) Setting $\nu^{(2)}$

Fig. 3: Cumulative regret of `KLUCB` and `KLUCB-RB` along $H = 100$ periods of $T = 5000$ rounds over three generated settings of $|\mathcal{B}| = 5$ bandit instances with $|\mathcal{A}| = 4$ arms per instance.

5.3 Critical Settings

We saw previously that settings where bandit instances are difficult to distinguish may yield poor performance (see Section 5.2, Fig. 3b). Indeed, to determine if two estimated bandit models might result from the same bandit, both `KLUCB-RB` and `tUCB` rely on a compatibility over each arm, i.e. the intersection of confidence intervals. Therefore, it is generally harder to distinguish rollouts from many different distributions (that is the cardinal of $|\mathcal{B}|$ is high) when $|\mathcal{A}|$ is low and differences between arms are tight. To illustrate that, we consider an experiment on the setting described in Figure 8 (Appendix C.3), composed of 4-armed bandits. We recall that `tUCB` requires in particular $|\mathcal{A}| \geqslant |\mathcal{B}|$. Thus we choose a set $|\mathcal{B}|$ of cardinal 4 in order to include a comparison of our algorithm with `tUCB`.



(a) $T = 100$ (b) $T = 500$

Fig. 4: Cumulative regret of `KLUCB`, `KLUCB-RB` and `tUCB` along $H = 500$ period for different numbers of rounds.

Here we have $|\mathcal{A}^\star| = \{0, 1, 3\}$ and $\gamma_\nu^\star := \min_{b \neq b'} \min_{a \in \mathcal{A}^\star} |\mu_{a,b} - \mu_{a,b'}| = 0.15$, while the minimal regret gaps of each instances are $(\min_{a \neq a_b^\star} \Delta_{a,b})_{b \in \mathcal{B}} = (0.74, 0.80, 0.81, 0.89)$. Consequently, finding the optimal arm at each period independently is here far less difficult than separating the different instances. Such a setting is clearly unfavorable for `KLUCB-RB` and we expect `KLUCB` to perform better.

Fig. 4a and Fig. 4b the cumulative regret for the three strategies, along $H = 500$ periods of $T = 100$ and $T = 500$ rounds respectively. As expected, `KLUCB` outperforms `KLUCB-RB` under this critical setting. On the other hand `tUCB` seems more robust and displays a cumulative regret trend that would be improving compared with `KLUCB` in the long run. One should still recall that `tUCB` requires knowing the cardinality of $|\mathcal{B}|$, while `KLUCB-RB` does not.

We may notice (Fig. 4a) that if the number of rounds T is sufficiently small, that is `KLUCB` does not have enough time to converge for each bandit, then `KLUCB-RB` does not perform significantly worse than `KLUCB` for the first periods. Then, as T rises (Fig. 4b), `KLUCB` begins to converge while `KLUCB-RB` still aggregate samples from confusing instances, which yields an explosion of the cumulative regret curve. We then expect for such setting that `KLUCB` will need far more longer periods ($T \to \infty$) to reach a regime in which it will discard all false positive rollouts and takes advantage over `KLUCB`. On the contrary, `tUCB` takes advantage of the knowledge of $|\mathcal{B}|$ and then waits to have enough confidence over the mean vectors of the 4 models to exploit them.

6 Conclusion

In this paper we introduced the new routine bandits framework, for which we provided lower bounds on the regret (Proposition 1). This setting applies well to problems where, for example, customers anonymously return to interact with a system. These dynamics are known to be of interest to the community, as evidenced by the existing literature [11,10,20]. Routine bandits complement well these existing settings.

We then proposed the `KLUCB-RB` strategy (Alg. 1) to tackle the routine bandit setting by building on the seminal `KLUCB` algorithm for classical bandits. We proved upper bounds on the number of sub-optimal plays by `KLUCB-RB` (Theorem 1), which were used to prove asymptotic upper bounds on the regret (Corollary 1). This result shows the asymptotic optimality of the strategy and thanks to the proof technique that we considered, which is of independent interest, we further obtained finite-time regret guarantees with explicit quantities. We indeed believe the proof technique may be useful to handle other structured setups beyond routine bandits.

We finally provided extensive numerical experiments to highlight the situations where `KLUCB-RB` can efficiently leverage information from previously encountered bandit instances to improve over a classical `KLUCB`. More importantly, we highlighted the cost to pay for re-using observations from previous periods, and showed that easy tasks may be better tackled independently. This is akin to an agent that

would behave badly by relying on a wrong inductive bias. Fortunately, there are many situations where one can leverage knowledge from bandit instances faced in the past. This would notably be the case if the agent has to select products to recommend from a large set ($\mathcal{A}$) and it turns out that there exists a much smaller set of products ($\mathcal{A}^\star$) that is preferred by users (Sec. 5.1).

Our results notably show that transferring information from previously encountered bandits can be highly beneficial (e.g., see Fig. 1 and 3a). However, the lack of prior knowledge about previous instances (including the cardinality of the set of instances) introduces many challenges in transfer learning. For example, attempting to leverage knowledge from previous instances could result in negative transfer if bandits cannot be distinguished properly (e.g., see Fig. 4).

Therefore, reducing the cost incurred for separating bandit instances should constitute a relevant angle to tackle as future work. Another natural line of other future work could investigate extensions of `KLUCB-RB` to the recurring occurrence of other bandit instances, e.g., linear bandits, contextual bandits, and others.

Acknowledgments

This work has been supported by CPER Nord-Pas-de-Calais/FEDER DATA Advanced data science and technologies 2015-2020, the French Ministry of Higher Education and Research, Inria, the French Agence Nationale de la Recherche (ANR) under grant ANR-16-CE40-0002 (the BADASS project), the MEL, the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF, and CIFAR.

References

1. Abbasi-Yadkori, Y., Pál, D., Szepesvári, C.: Improved algorithms for linear stochastic bandits. In: Advances in Neural Information Processing Systems. pp. 2312–2320 (2011)
2. Agrawal, R., Teneketzis, D., Anantharam, V.: Asymptotically efficient adaptive allocation schemes for controlled iid processes: Finite parameter space. IEEE Transactions on Automatic Control **34**(3) (1989)
3. Anandkumar, A., Ge, R., Hsu, D., Kakade, S.: A tensor spectral approach to learning mixed membership community models. In: Conference on Learning Theory. pp. 867–881. PMLR (2013)
4. Anandkumar, A., Ge, R., Hsu, D.J., Kakade, S.M., Telgarsky, M.: Tensor decompositions for learning latent variable models. Journal of Machine Learning Research **15**(1), 2773–2832 (2014)
5. Bubeck, S., Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Founadtions and Trends in Machine Learning **abs/1204.5721** (2012), `http://arxiv.org/abs/1204.5721`
6. Cappé, O., Garivier, A., Maillard, O.A., Munos, R., Stoltz, G.: Kullback–Leibler upper confidence bounds for optimal sequential allocation. The Annals of Statistics **41**(3), 1516–1541 (2013)
7. Combes, R., Proutiere, A.: Unimodal bandits: Regret lower bounds and optimal algorithms. In: Proceedings of the 31st International Conference on Machine Learning (ICML). pp. 521–529 (2014)

8. Garivier, A.: Informational confidence bounds for self-normalized averages and applications. In: Information Theory Workshop (ITW), 2013 IEEE. pp. 1–5. IEEE (2013)
9. Garivier, A., Cappé, O.: The KL-UCB algorithm for bounded stochastic bandits and beyond. In: Proceedings of the 24th Annual Conference on Learning Theory (COLT). pp. 359–376 (2011)
10. Gentile, C., Li, S., Zappella, G.: Online clustering of bandits. In: Proc. ICML. pp. 757–765 (2014)
11. Gheshlaghi Azar, M., Lazaric, A., Brunskill, E.: Sequential transfer in multi-armed bandit with finite set of models. In: Proc. NIPS, pp. 2220–2228. Curran Associates, Inc. (2013)
12. Graves, T.L., Lai, T.L.: Asymptotically efficient adaptive choice of control laws incontrolled markov chains. SIAM journal on control and optimization **35**(3), 715–743 (1997)
13. Honda, J., Takemura, A.: An asymptotically optimal bandit algorithm for bounded support models. In: Proceedings of the 23rd Annual Conference on Learning Theory. Haifa, Israel (2010)
14. Korda, N., Kaufmann, E., Munos, R.: Thompson Sampling for 1-dimensional exponential family bandits. In: Advances in Neural Information Processing Systems (NIPS). pp. 1448–1456 (2013)
15. Lai, T.L.: Adaptive treatment allocation and the multi-armed bandit problem. The Annals of Statistics pp. 1091–1114 (1987)
16. Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics **6**(1), 4–22 (1985)
17. Langford, J., Zhang, T.: The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S.T., Platt, J.C., Koller, D., Singer, Y., Roweis, S.T. (eds.) NIPS. MIT Press (2007)
18. Lattimore, T., Szepesvári, C.: Bandit Algorithms. Cambridge University Press (2020)
19. Lu, T., Pál, D., Pál, M.: Contextual multi-armed bandits. In: Teh, Y.W., Titterington, M. (eds.) Proceedings of the 13th international conference on Artificial Intelligence and Statistics. vol. 9, pp. 485–492 (2010)
20. Maillard, O.A., Mannor, S.: Latent bandits. In: International Conference on Machine Learning (ICML) (2014)
21. Peña, V.H., Lai, T.L., Shao, Q.M.: Self-normalized processes: Limit theory and Statistical Applications. Springer Science & Business Media (2008)
22. Robbins, H.: Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society **58**(5), 527–535 (1952)
23. Thompson, W.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**, 285–294 (1933)

A Proof of Proposition 1

Let us denote by $\mathcal{S}$ the routine bandit setting and by $\mathcal{S}_0$ the setting resulting from the routine bandit setting and the *additional assumption* that now the sequence of bandits $(b_\star^h)_{h\in[\![1,H]\!]}$ is known to the learner. Then, since a consistent strategy for $\mathcal{S}$ is also consistent for $\mathcal{S}_0$ (in the sense of Definition 1), we deduce Proposition 1 from Lemma 5.

Lemma 5 (Lower bounds on the regret for $\mathcal{S}_0$). *Let us consider a consistent strategy for the setting $\mathcal{S}_0$. Then, for all configuration $\nu\in\mathcal{D}$, it must be that*

$$\liminf_{T\to\infty}\frac{R(\nu,H,T)}{\log(T)}\geqslant c_\nu^\star\coloneqq\sum_{b\in\mathcal{B}}\sum_{a\neq a_b^\star}\frac{\Delta_{a,b}}{\mathrm{KL}(\mu_{a,b}|\,\mu_b^\star)}\,.$$

Proof. Since the sequence of bandits $(b_\star^h)_{h\in[\![1,H]\!]}$ is known to the learner and since there is no shared information between the bandits at first glance, the setting $\mathcal{S}_0$ amounts to consider each of the $|\mathcal{B}|$ bandits $(\nu_b)_{b\in\mathcal{B}}$ as a separate problem, where $\nu_b\coloneqq(\nu_{a,b})_{a\in\mathcal{A}}$ for $b\in\mathcal{B}$. Then, from the known lower bound on the regret for the classical multi-armed bandit problem [16], we get under the assumption of consistency for all bandit $b\in\mathcal{B}$,

$$\liminf_{T\to\infty}\frac{1}{\log(N_b(H,T))}\sum_{a\neq a_b^\star}N_{a,b}(T)\Delta_{a,b}\geqslant\sum_{a\neq a_b^\star}\frac{\Delta_{a,b}}{\mathrm{KL}(\mu_{a,b}|\,\mu_b^\star)},\ \text{where } N_b(H,T)=\beta_b^H HT\,.$$

From previous inequalities and Eq. 2, we conclude that

$$\begin{aligned}\liminf_{T\to\infty}\frac{R(\nu,H,T)}{\log(T)}&\geqslant\sum_{b\in\mathcal{B}}\liminf_{T\to\infty}\frac{\log\big(\beta_b^HHT\big)}{\log(T)}\liminf_{T\to\infty}\frac{1}{\log\big(\beta_b^HHT\big)}\sum_{a\neq a_b^\star}N_{a,b}(T)\Delta_{a,b}\\&\geqslant\sum_{b\in\mathcal{B}}\sum_{a\neq a_b^\star}\frac{\Delta_{a,b}}{\mathrm{KL}(\mu_{a,b}|\,\mu_b^\star)}\,,\end{aligned}$$

by Fatou's Lemma and since we have $\liminf_n u_nv_n\geqslant\liminf_n u_n\liminf_n v_n$ for all positive real-valued sequences u,v. □

B Proof of Theorem 1

From Proposition 2, we have the following inequality

$$\begin{aligned}N_{a,b}(H,T)\leqslant{}&\frac{f(\beta_b^HHT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\,\mu_b^\star)}\\&+\sum_{h=1}^H\mathbb{I}_{\{b_\star^h=b\}}\left[T_{\nu,\varepsilon}^h+4\big|\mathcal{C}_\varepsilon^h\big|+\Big|\overline{\mathcal{C}}_\varepsilon^h\Big|+\frac{f(hT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\,\mu_b^\star)}\sum_{k=1}^h\frac{8\,\big|\mathcal{C}_\varepsilon^k\big|}{T}+\mathbb{I}_{\{T\in\mathcal{T}^k\}}\right],\end{aligned}\tag{16}$$

where for all $h\geqslant1$, $\mathbb{P}_\nu\big(T\in\mathcal{T}^h\big)\leqslant1/T(T+1)$, $\mathbb{E}_\nu\big[\big|\mathcal{C}_\varepsilon^h\big|\big],\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{C}}_\varepsilon^h\Big|\Big]\leqslant4\,|\mathcal{A}|\,\varepsilon^{-2}+3$ according to Lemma 10 and $T_{\nu,\varepsilon}^h\leqslant\tau_\nu^h$ according to Lemma 6 stated below.

By taking the expectation in Eq. 16 then it comes

$$\begin{aligned}
&\mathbb{E}_\nu[N_{a,b}(H,T)] \\
&\leqslant \frac{f(\beta_b^H HT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\,\mu_b^\star)} \\
&\quad + \sum_{h=1}^{H} \mathbb{I}_{\{b_\star^h=b\}}\left[\tau_\nu^h + 5\times\left(4\,|\mathcal{A}|\,\varepsilon^{-2}+3\right) + \frac{f(hT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\,\mu_b^\star)}\sum_{k=1}^{h}\frac{32\,|\mathcal{A}|\,\varepsilon^{-2}+24}{T} + 1/T(T+1)\right].
\end{aligned}$$

We conclude the proof of Theorem 1 by using the two following inequalities

$$\begin{aligned}
&4\,|\mathcal{A}|\,\varepsilon^{-2}+3 \leqslant 4\,|\mathcal{A}|\,(\varepsilon^{-2}+1) \\
&\frac{32\,|\mathcal{A}|\,\varepsilon^{-2}+24}{T} + \frac{1}{T(T+1)} \leqslant \frac{32\,|\mathcal{A}|\,(\varepsilon^{-2}+1)}{T}\,.
\end{aligned}$$

Lemma 6 (Upper bound on $T_{\nu,\varepsilon}^h$). *With the same notations as Proposition 2, for all $0<\varepsilon<\varepsilon_\nu$,*

$$T_{\nu,\varepsilon}^h \leqslant \tau_\nu^h := 2\varphi\Big(8|\mathcal{A}|\left[\varepsilon_\nu^{-2} + 65\gamma_\nu^{-2}\log\Big(128\,|\mathcal{A}|\,(4h)^{1/3}\gamma_\nu^{-2}\Big)\right]\Big)\,,$$

where $\varphi\colon x\geqslant 1\mapsto x\log(x)$.

Proof. We first show that

$$t_\nu^h < 130\,|\mathcal{A}|\,\gamma_\nu^{-2}\log\Big(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\Big)\,. \tag{17}$$

Let us consider $t\geqslant 3\,|\mathcal{A}|$. We have

$$\mathrm{d}\Big(\frac{t}{|\mathcal{A}|},\delta^h(t)\Big) = \sqrt{2\Big(1+\frac{|\mathcal{A}|}{t}\Big)\,\frac{\log\Big(4\,|\mathcal{A}|^3\,(h-1)\sqrt{t/\,|\mathcal{A}|+1}(t/\,|\mathcal{A}|)(t/\,|\mathcal{A}|+1/\,|\mathcal{A}|)\Big)}{t/\,|\mathcal{A}|}}\,.$$

Since $1/|\mathcal{A}|<1$ and $t/\,|\mathcal{A}|\geqslant 3$, we have

$$\sqrt{t/\,|\mathcal{A}|+1}(t/\,|\mathcal{A}|)(t/\,|\mathcal{A}|+1/\,|\mathcal{A}|) \leqslant \sqrt{t/\,|\mathcal{A}|+1}(t/\,|\mathcal{A}|)(t/\,|\mathcal{A}|+1) \leqslant (t/\,|\mathcal{A}|)^3\,.$$

Then, since $1+|\mathcal{A}|/t<1+1/3$ and $h-1\leqslant h$, we get

$$\mathrm{d}\Big(\frac{t}{|\mathcal{A}|},\delta^h(t)\Big) \leqslant \sqrt{\frac{8\,|\mathcal{A}|\,(4h)^{1/3}}{\Phi\big((4h)^{1/3}t\big)}}$$

where $\Phi : x\geqslant 3\mapsto x/\log(x)\geqslant\Phi(3)$. $\Phi(\cdot)$ is a one-to-one function and $\forall y\geqslant\Phi(3), \Phi^{-1}(y)\leqslant y\log(y)+2\log(y)$. Thus we have

$$\begin{aligned}
\gamma_\nu \leqslant 4\,\mathrm{d}\Big(\frac{t}{|\mathcal{A}|},\delta^h(t)\Big) \quad\Rightarrow\quad t &\leqslant (4h)^{-1/3}\Phi^{-1}\Big(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\Big) \\
&\leqslant 128\,|\mathcal{A}|\,\gamma_\nu^{-2}\log\Big(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\Big) \\
&\quad + 2(4h)^{-1/3}\log\Big(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\Big)\,.
\end{aligned}$$

In particular, we get the following implication

$$\gamma_\nu \leqslant 4\,\mathrm{d}\Big(\frac{t}{|\mathcal{A}|}, \delta^h(t)\Big) \quad \Rightarrow \quad t < 130\,|\mathcal{A}|\,\gamma_\nu^{-2}\log\Big(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\Big) - 1$$

and $t_\nu^h < 130\,|\mathcal{A}|\,\gamma_\nu^{-2}\log\big(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\big)$.

Furthermore, we have

$$\sum_{a \neq a_\star^h} \frac{f(t)}{\mathrm{KL}(\mu_a^h + \varepsilon|\,\mu_\star^h)} < 2\,|\mathcal{A}|\,\varepsilon_\nu^{-2} f(t)\,. \tag{18}$$

By combining Eq. 17 and Eq. 18, from the definition of $T_{\nu,\varepsilon}^h$ (see Proposition 2) we get

$$T_{\nu,\varepsilon}^h \leqslant \max\Big\{t \geqslant 10:\ t - 8\,|\mathcal{A}|\,\Big(\varepsilon_\nu^{-2} f(t) - 65\gamma_\nu^{-2}\log\Big(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\Big)\Big) \leqslant 0\Big\}\,. \tag{19}$$

We finally prove Lemma 6 by applying Lemma 7 with $c = 8\,|\mathcal{A}|\,\varepsilon_\nu^{-2}$ and $c' = 130\,|\mathcal{A}|\,\gamma_\nu^{-2}\log\big(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\big)$. □

Lemma 7. *For all $c, c' > 10$, it holds*

$$\max\{t \geqslant 10:\ t - cf(t) - c' \leqslant 0\} \leqslant 2\varphi(c + c')\,,$$

where $\varphi\colon x \geqslant 1 \mapsto x\log(x)$.

Proof. It can be shown that $(t - cf(t) - c')_{t \geqslant 2\varphi(c+c')}$ is non-decreasing by standard derivative analysis and that $2\varphi(c+c') - cf(2\varphi(c+c')) - c \geqslant 0$. □

In the following we prove the results stated in Section 4.

B.1 Proof of Lemma 1

In this subsection we control the number previously encountered bandits falsely identified as different to the current one (see Definition 3) in addition to false positives and prove Lemma 8, an extension of Lemma 1.

Definition 3 (False negative). *At period $h \geqslant 2$ and step $t \geqslant 1$, a previous period $k \in [\![1, h-1]\!]$ is called a false negative if the test $\mathrm{T}^{k,h}(t)$ is false while previous bandit $b_\star^k$ corresponds to current bandit $b_\star^h$.*

We prove necessary conditions for having false positives or false negatives.

Lemma 8 (Condition for false positives/negatives). *At period $h \geqslant 2$ and time step $t > |\mathcal{A}|$, for all period $k \in [\![1, h-1]\!]$, , with probability $1 - 1/(h-1)t(t+1)$,*

$$k \text{ is a false positive } \implies b_\star^k \neq b_\star^h \text{ and } \min_{a \in \mathcal{A}} \big|\mu_{a,b_\star^h} - \mu_{a,b_\star^k}\big| \leqslant 4\,\mathrm{d}\Big(\frac{t}{|\mathcal{A}|}, \delta^h(t)\Big)$$
$$k \text{ is a false negative} \iff b_\star^k = b_\star^h \text{ and } \overline{a}_t^k \neq \overline{a}_t^h\,.$$

Proof. From Lemma 14, with probability $1-4\,|\mathcal{A}|\,\delta^h(t)=1-1/(h-1)t(t+1)$, it holds,

$$\forall a \in \mathcal{A}, \quad \left|\widehat{\mu}_a^h(t) - \mu_a^h\right| \leqslant \mathrm{d}\big(N_a^h(t), \delta^h(t)\big) \quad \text{and} \quad \left|\widehat{\mu}_a^k(T) - \mu_a^k\right| \leqslant \mathrm{d}\big(N_a^k(T), \delta^h(t)\big)\,. \tag{20}$$

False negative: Here we assume that $b_\star^k = b_\star^h$. By the triangle inequality, this implies

$$\forall a \in \mathcal{A}, \quad \left|\widehat{\mu}_a^h(t) - \widehat{\mu}_a^k(T)\right| = \left|\big(\widehat{\mu}_a^h(t) - \mu_a^h\big) - \big(\widehat{\mu}_a^k(T) - \mu_a^k\big)\right| \leqslant \left|\widehat{\mu}_a^h(t) - \mu_a^h\right| + \left|\widehat{\mu}_a^k(T) - \mu_a^T\right|\,. \tag{21}$$

By combining Eq.20 and 21, with probability $1-1/(h-1)t(t+1)$, we have

$$\forall a \in \mathcal{A}, \quad \left|\widehat{\mu}_a^h(t) - \widehat{\mu}_a^k(T)\right| - \mathrm{d}\big(N_a^h(t), \delta^h(t)\big) - \mathrm{d}\big(N_a^k(T), \delta^h(t)\big) \leqslant 0\,.$$

Then, from the definitions of the random variables $(Z_a^{k,h}(t))_{a\in\mathcal{A}}$ (Eq. 3) and the test $\mathrm{T}^{k,h}(t)$ (Eq. 4), this implies with probability $1-1/(h-1)t(t+1)$,

$$\max_{a\in\mathcal{A}} Z_a^{k,h}(t) \leqslant \infty \cdot \mathbb{I}_{\left\{\overline{a}_t^h \neq \overline{a}_\star^k\right\}}, \qquad \mathrm{T}^{k,h}(t) = \left(\overline{a}_t^h = \overline{a}_\star^k\right).$$

Thus, with probability $1-1/(h-1)t(t+1)$, period k is a false negative if, and only if, $\overline{a}_t^h \neq \overline{a}_\star^k$.

False positive: Here we assume that period k is a false positive. In particular, we have $b_\star^k \neq b_\star^h$. By the triangle inequality, this implies

$$\forall a \in \mathcal{A}, \quad \left|\widehat{\mu}_a^h(t) - \widehat{\mu}_a^k(T)\right| \geqslant \left|\mu_a^h - \mu_a^k\right| - \left|\widehat{\mu}_a^h(t) - \mu_a^h\right| - \left|\widehat{\mu}_a^k(T) - \mu_a^T\right|\,. \tag{22}$$

By combining Eq.20 and 22, with probability $1-1/(h-1)t(t+1)$, we have

$$\forall a \in \mathcal{A}, \quad Z_a^{k,h}(t) \geqslant \infty \cdot \mathbb{I}_{\left\{\overline{a}_t^h \neq \overline{a}_\star^k\right\}} + \min_{a\in\mathcal{A}} \left|\mu_a^h - \mu_a^k\right| - 2\,\mathrm{d}\big(N_a^h(t), \delta^h(t)\big) - 2\,\mathrm{d}\big(N_a^k(T), \delta^h(t)\big)\,. \tag{23}$$

Since period is assumed to be a false positive, we have $\max_{a\in\mathcal{A}} Z^{k,h}(t) \leqslant 0$ and Eq. 23 implies that, with probability $1-1/(h-1)t(t+1)$,

$$\overline{a}_t^h = \overline{a}_\star^k, \qquad \min_{a\in\mathcal{A}} \left|\mu_a^h - \mu_a^k\right| \leqslant 2\,\mathrm{d}\Big(N_{\overline{a}_t^h}^h(t), \delta^h(t)\Big) + 2\,\mathrm{d}\Big(N_{\overline{a}_\star^k}^k(T), \delta^h(t)\Big)\,. \tag{24}$$

Since $N_{\overline{a}_t^h}^h(t) \geqslant t/\left|\mathcal{A}\right|$, $N_{\overline{a}_\star^k}^k(T) \geqslant T/\left|\mathcal{A}\right|$ ($\overline{a}_t^h$ and $\overline{a}_\star^k$ are most pulled arms) and $\delta^h(T) \leqslant \delta^h(t)$, the monotonic properties of $\mathrm{d}(\cdot,\cdot)$ and Eq. 24, imply that, with probability $1-1/(h-1)t(t+1)$,

$$\min_{a\in\mathcal{A}} \left|\mu_a^h - \mu_a^k\right| \leqslant 2\,\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right) + 2\,\mathrm{d}\left(\frac{T}{|\mathcal{A}|}, \delta^h(T)\right)\,.$$

We conclude the proof of Lemma 8 by using Lemma 9 stated below. □

Lemma 9 (Monotonic properties of $\mathrm{d}(\cdot,\cdot)$). *For all period $h \geqslant 2$, $\left(\mathrm{d}\big(t/|\mathcal{A}|\,, \delta^h(t)\big)\right)_{t>|\mathcal{A}|}$ is non-increasing.*

Proof. For all time step $t>|\mathcal{A}|$, a direct calculation gives

$$\mathrm{d}\Big(\frac{t}{|\mathcal{A}|},\delta^h(t)\Big)=\sqrt{2\,|\mathcal{A}|\Big(1+\frac{|\mathcal{A}|}{t}\Big)\Big(\frac{1}{2}\frac{\log(t/\,|\mathcal{A}|+1)}{t}+\frac{\log(4\,|\mathcal{A}|\,(h-1))}{t}+\frac{\log(t+1)}{t}+\frac{\log(t)}{t}\Big)}.$$

Then, in order to prove Lemma 9, it is sufficient to note that $(\log(t/|\mathcal{A}|+1)/t)_{t\geqslant 1}$, $(\log(t+1)/t)_{t\geqslant 2}$ and $(\log(t)/t)_{t\geqslant 3}$ are non-increasing. □

B.2 Proof of Lemma 2

Let us consider the subsets of times when the mean of the current pulled arm is poorly estimated

$$\mathcal{E}^h_{a,\varepsilon}\coloneqq\big\{t>|\mathcal{A}|:\ a^h_{t+1}=a \text{ and } \big|\widehat{\mu}^h_a(t)-\mu^h_a\big|>\varepsilon\big\}\qquad \mathcal{E}^h_{\varepsilon}\coloneqq\bigcup_{a\neq a^h_\star}\mathcal{E}^h_{a,\varepsilon}$$

$$\overline{\mathcal{E}}^h_{a,\varepsilon}\coloneqq\big\{t\geqslant t^h_\nu:\ t\notin\mathcal{T}^h,\ a^h_{t+1}=a \text{ and } \big|\overline{\mu}^h_a(t)-\mu^h_a\big|>\varepsilon\big\}\qquad \overline{\mathcal{E}}^h_{\varepsilon}\coloneqq\bigcup_{a\neq a^h_\star}\overline{\mathcal{E}}^h_{a,\varepsilon}$$

and the subsets of times when the best arm $a^h_\star$ is below its mean

$$\mathcal{U}^h_a\coloneqq\Big\{t>|\mathcal{A}|:\ a^h_{t+1}=a \text{ and } u^h_{a^h_\star}(t)=U^h_{a^h_\star}(t)<\mu^h_\star\Big\}\qquad \mathcal{U}^h\coloneqq\bigcup_{a\neq a^h_\star}\mathcal{U}^h_a\,.$$

$$\overline{\mathcal{U}}^h_a\coloneqq\Big\{t\geqslant t^h_\nu:\ t\notin\mathcal{T}^h,\ a^h_{t+1}=a \text{ and } u^h_{a^h_\star}(t)=\overline{U}^h_{a^h_\star}(t)<\mu^h_\star\Big\}\qquad \overline{\mathcal{U}}^h\coloneqq\bigcup_{a\neq a^h_\star}\overline{\mathcal{U}}^h_a\,.$$

Then we have

$$\mathcal{C}^h_{a,\varepsilon}=\mathcal{T}^h_a\cup\mathcal{E}^h_{a,\varepsilon}\cup\mathcal{U}^h_a\qquad \mathcal{C}^h_{\varepsilon}=\mathcal{T}^h\cup\mathcal{E}^h_{\varepsilon}\cup\mathcal{U}^h$$

$$\overline{\mathcal{C}}^h_{a,\varepsilon}=\mathcal{T}^h_a\cup\mathcal{E}^h_{a,\varepsilon}\cup\overline{\mathcal{U}}^h_a\qquad \overline{\mathcal{C}}^h_{\varepsilon}=\mathcal{T}^h\cup\mathcal{E}^h_{\varepsilon}\cup\overline{\mathcal{U}}^h$$

and deduce Lemma 2 from the extended Lemma 10.

Lemma 10 (Bounded subsets of times). *For all period $h\geqslant 2$, for all arm $a\in\mathcal{A}$, for all $0<\varepsilon<\varepsilon_\nu$,*

$$\forall t\in[\![1,T]\!],\ \mathbb{P}_\nu\big(t\in\mathcal{T}^h\big)\leqslant\frac{1}{t(t+1)},\quad \mathbb{E}_\nu\big[\big|\mathcal{E}^h_{a,\varepsilon}\big|\big],\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{E}}^h_{a,\varepsilon}\Big|\Big]\leqslant 4\varepsilon^{-2},\quad \mathbb{E}_\nu\big[\big|\mathcal{U}^h\big|\big],\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{U}}^h\Big|\Big]\leqslant 2\,.$$

This implies

$$\mathbb{E}_\nu\big[\big|\mathcal{T}^h\big|\big]\leqslant 1,\qquad \mathbb{E}_\nu\big[\big|\mathcal{C}^h_{a,\varepsilon}\big|\big],\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{C}}^h_{a,\varepsilon}\Big|\Big]\leqslant 4\varepsilon^{-2}+3,\qquad \mathbb{E}_\nu\big[\big|\mathcal{C}^h_{\varepsilon}\big|\big],\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{C}}^h_{\varepsilon}\Big|\Big]\leqslant 4\,|\mathcal{A}|\,\varepsilon^{-2}+3\,.$$

Proof. Subset $\mathcal{T}^h$: From Lemma 10 and the definition of t_ν^h (see Eq. 8), for all $t \geqslant t_\nu^h$, with probability $1-1/t(t+1)$, there is no false positive and if a previous period $k \in [\![1, h-1]\!]$ is a false negative then $b_\star^k = b_\star^h$ and $\overline{a}_\star^k \neq \overline{a}_t^h$ (the most pulled arms are different). From the definition of $\mathcal{T}^h$ (see Eq. 9) this implies that for all $t \geqslant t_\nu^h$, with probability $1-1/t(t+1)$, $t \notin \mathcal{T}^h$. That is $\forall t \geqslant t_\nu^h$, $\mathbb{P}_\nu\big(t \in \mathcal{T}^h\big) \leqslant 1/t(t+1)$. Since on the other hand, we have

$$\big|\mathcal{T}^h\big| = \sum_{t=t_\nu^h}^{T} \mathbb{I}_{\{t \in \mathcal{T}^h\}}\,,$$

by taking expectation on both sides, it comes

$$\mathbb{E}_\nu\big[\big|\mathcal{T}^h\big|\big] = \sum_{t=t_\nu^h}^{T} \mathbb{P}_\nu\big(t \in \mathcal{T}^h\big) \leqslant \sum_{t=t_\nu^h}^{T} \frac{1}{t(t+1)} \leqslant 1\,.$$

We note that for $1 \leqslant t < t_\nu^h$, it holds that $t \notin \mathcal{T}^h$ and $\mathbb{P}_\nu\big(t \in \mathcal{T}^h\big) = 0 \leqslant 1/t(t+1)$.

Subset $\mathcal{E}_{a,\varepsilon}^h$:
Since we have

$$\big|\mathcal{E}_{a,\varepsilon}^h\big| = \sum_{t>|\mathcal{A}|}^{T} \mathbb{I}_{\left\{a_{t+1}^h = a,\ |\widehat{\mu}_a^h(t) - \mu_a^h| > \varepsilon\right\}}\,,$$

by taking the expectation on both sides, it comes

$$\mathbb{E}_\nu\big[\big|\mathcal{E}_{a,\varepsilon}^h\big|\big] \leqslant \sum_{t=1}^{T} \mathbb{P}_\nu\big(a_{t+1}^h = a,\ \big|\widehat{\mu}_a^h(t) - \mu_a^h\big| > \varepsilon\big)\,. \tag{25}$$

Then, by combining Eq. 25 and Lemma 13, we prove $\mathbb{E}_\nu\big[\big|\mathcal{E}_{a,\varepsilon}^h\big|\big] \leqslant 4\varepsilon^{-2}$.

Subset $\overline{\mathcal{E}}_{a,\varepsilon}^h$: From the definitions of t_ν^h and $\mathcal{T}^h$ (Eq. 8 and 9), we get the following inclusion

$$\Big\{t \geqslant t_\nu^h\colon\ t \notin \mathcal{T}^h,\ a_{t+1}^h = a,\ \big|\overline{\mu}_a^h(t) - \mu_a^h\big| > \varepsilon\Big\} \subset \Big\{t \geqslant t_\nu^h\colon a_{t+1}^h = a,\ \Big|\widehat{\mu}_a^{\mathcal{K}_\star^h(t),h}(t) - \mu_a^h\Big| > \varepsilon\Big\}, \tag{26}$$

where $\mathcal{K}_\star^h(t) \coloneqq \big\{k \in [\![1, h-1]\!]\colon b_\star^k = b_\star^h \text{ and } \overline{a}_\star^k = \overline{a}_t^h\big\}$ and $N_a^{\mathcal{K},h}(t) = \sum_{k \in \mathcal{K}} N_a^k(T) + N_a^h(t)$,
$S_a^{\mathcal{K},h}(t) = \sum_{k \in \mathcal{K}} S_a^k(T) + S_a^h(t)$, $\widehat{\mu}_a^{\mathcal{K},h}(t) = S_a^{\mathcal{K},h}(t)/N_a^{\mathcal{K},h}(t)$, $\forall \mathcal{K} \subset \mathcal{K}^h \coloneqq \big\{k \in [\![1, h-1]\!]\colon b_\star^k = b_\star^h\big\}$.

Thus, by defining $\mathcal{K}_t \coloneqq \mathcal{K}_\star^h(t)$ if $t \geqslant t_\nu^h$ and $t \notin \mathcal{T}^h$, $\emptyset$ otherwise, Eq. 26 implies

$$\forall t \geqslant t_\nu^h, \mathbb{P}_\nu\big(t \notin \mathcal{T}^h, a_{t+1}^h = a, \big|\overline{\mu}_a^h(t) - \mu_a^h\big| > \varepsilon\big) \leqslant \mathbb{P}_\nu\big(a_{t+1}^h = a, \big|\widehat{\mu}_a^{\mathcal{K}_t,h}(t) - \mu_a^h\big| > \varepsilon\big). \tag{27}$$

Since we have

$$\Big|\overline{\mathcal{E}}_{a,\varepsilon}^h\Big| = \sum_{t=t_\nu^h}^{T} \mathbb{I}_{\left\{t \notin \mathcal{T}^h,\ a_{t+1}^h = a,\ |\overline{\mu}_a^h(t) - \mu_a^h| > \varepsilon\right\}}\,,$$

by taking the expectation on both sides and using inequalities from Eq.27, it comes

$$\begin{aligned}\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{E}}^h_{a,\varepsilon}\Big|\Big] &= \sum_{t=t^h_\nu}^{T} \mathbb{P}_\nu\big(t\notin\mathcal{T}^h,\ a^h_{t+1}=a,\ \big|\overline{\mu}^h_a(t)-\mu^h_a\big|>\varepsilon\big)\\ &\leqslant \sum_{t=t^h_\nu}^{T} \mathbb{P}_\nu\big(a^h_{t+1}=a, \big|\widehat{\mu}^{\mathcal{K}_t,h}_a(t)-\mu^h_a\big|>\varepsilon\big)\,. \end{aligned} \tag{28}$$

Then, by combining Eq. 28 and Lemma 13, we prove $\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{E}}^h_{a,\varepsilon}\Big|\Big]\leqslant 4\varepsilon^{-2}$.

Subset $\mathcal{U}^h$:
By definition of the index (Eq. 6), we have

$$\forall t>|\mathcal{A}|\,,\quad N^h_{a^h_\star}(t)\mathrm{KL}\Big(\widehat{\mu}^h_{a^h_\star}(t)\Big|\,U^h_{a^h_\star}(t)\Big)=f(t)\,. \tag{29}$$

Since $\widehat{\mu}^h_{a^h_\star}(t)\leqslant U^h_{a^h_\star}(t)$ for all $t>|\mathcal{A}|$, from the monotony of $\mathrm{KL}(x|\,\cdot)$ on $[x,+\infty)$, it comes

$$\forall t>|\mathcal{A}|\ \text{ such that } U^h_{a^h_\star}(t)\leqslant\mu^h_{a^h_\star},\quad \mathrm{KL}\Big(\widehat{\mu}^h_{a^h_\star}\Big|\,\mu^h_{a^h_\star}\Big)\geqslant\mathrm{KL}\Big(\widehat{\mu}^h_{a^h_\star}(t)\Big|\,U^h_{a^h_\star}(t)\Big)\,. \tag{30}$$

From Eq. 29 and 30 we deduce that

$$\mathcal{U}^h\subset\Big\{t>|\mathcal{A}|:\ N^h_{a^h_\star}(t)\mathrm{KL}\Big(\widehat{\mu}^h_{a^h_\star}(t)\big|\,\mu_{a^h_\star,b^h_\star}\Big)\geqslant f(t)\Big\}\,. \tag{31}$$

From Eq. 31 plus the union bound, it comes

$$\big|\mathcal{U}^h\big|\leqslant\sum_{t>|\mathcal{A}|}^{T}\mathbb{I}_{\left\{N^h_{a^h_\star}(t)\mathrm{KL}\left(\widehat{\mu}^h_{a^h_\star}(t)\middle|\,\mu_{a^h_\star,b^h_\star}\right)\geqslant f(t)\right\}}\,. \tag{32}$$

By taking the expectation on both sides in previous inequality (Eq. 32), we have

$$\mathbb{E}_\nu\big[\big|\mathcal{U}^h\big|\big]\leqslant\sum_{t>|\mathcal{A}|}^{T}\mathbb{P}_\nu\Big(N^h_{a^h_\star}(t)\mathrm{KL}\Big(\widehat{\mu}^h_{a^h_\star}(t)\big|\,\mu_{a^h_\star,b^h_\star}\Big)\geqslant f(t)\Big)\,. \tag{33}$$

Combining Eq. 33 and Lemma 14, it comes

$$\mathbb{E}_\nu\big[\big|\mathcal{U}^h\big|\big]\leqslant\sum_{t>|\mathcal{A}|}t^{-1}\log(t)^{-2}\,.$$

This implies $\mathbb{E}_\nu\big[\big|\mathcal{U}^h\big|\big]\leqslant 2$ since it can be shown that

$$\sum_{t>|\mathcal{A}|}t^{-1}\log(t)^{-2}\leqslant\int_{t\geqslant|\mathcal{A}|}^{\infty}t^{-1}\log(t)^{-2}\,\mathrm{d}t=\frac{1}{\log(|\mathcal{A}|)}\leqslant\frac{1}{\log(2)}\leqslant 2\,.$$

Subset $\overline{\mathcal{U}}^h$: From the definition of subset $\mathcal{T}^h$ (see Eq. 9), we have

$$\left\{t \geqslant t_\nu^h : t \notin \mathcal{T}^h\right\} \subset \left\{t \geqslant t_\nu^h :\ \overline{N}_{a_\star^h}^h(t) = N_{a_\star^h}^{\mathcal{K}_\star^h(t),h}(t),\ \overline{\mu}_{a_\star^h}^h(t) = \widehat{\mu}_{a_\star^h}^{\mathcal{K}_\star^h(t),h}(t),\ \overline{K}_t^h = \left|\mathcal{K}_\star^h(t)\right|\right\}, \tag{34}$$

where $\mathcal{K}_\star^h(t) \coloneqq \left\{k \in [\![1, h-1]\!] : b_\star^k = b_\star^h \text{ and } \overline{a}_\star^k = \overline{a}_t^h\right\} \subset \mathcal{K}^h$, $\mathcal{K}^h \coloneqq \left\{k \in [\![1, h-1]\!] : b_\star^k = b_\star^h\right\}$, $N_a^{\mathcal{K},h}(t) = \sum_{k\in\mathcal{K}} N_a^k(T) + N_a^h(t)$, $S_a^{\mathcal{K},h}(t) = \sum_{k\in\mathcal{K}} S_a^k(T) + S_a^h(t)$ and $\widehat{\mu}_a^{\mathcal{K},h}(t) = S_a^{\mathcal{K},h}(t)/N_a^{\mathcal{K},h}(t)$ for all $\mathcal{K} \subset \mathcal{K}^h$ and $a \in \mathcal{A}$.
By definition of the index (Eq. 7), we have

$$\forall t \geqslant t_\nu^h, \quad \overline{N}_{a_\star^h}^h(t) \mathrm{KL}\left(\overline{\mu}_{a_\star^h}^h(t) \middle| \overline{U}_{a_\star^h}^h(t)\right) = f\left(\overline{K}_t^h T + t\right). \tag{35}$$

Since $\overline{\mu}_{a_\star^h}^h(t) \leqslant \overline{U}_{a_\star^h}^h(t)$ for all $t \geqslant t_\nu^h$, from the monotony of $\mathrm{KL}(x|\cdot)$ on $[x, +\infty)$, it comes

$$\forall t \geqslant t_\nu^h \text{ such that } \overline{U}_{a_\star^h}^h(t) \leqslant \mu_{a_\star^h}^h, \quad \mathrm{KL}\left(\overline{\mu}_{a_\star^h}^h(t) \middle| \mu_{a_\star^h}^h\right) \geqslant \mathrm{KL}\left(\overline{\mu}_{a_\star^h}^h(t) \middle| \overline{U}_{a_\star^h}^h(t)\right). \tag{36}$$

By defining $\mathcal{K}_t \coloneqq \mathcal{K}_\star^h(t)$ if $t \geqslant t_\nu^h$ and $t \notin \mathcal{T}^h$, $\emptyset$ otherwise, from Eq. 34, 35 and 36 we deduce that

$$\overline{\mathcal{U}}^h \subset \left\{t \geqslant t_\nu^h :\ N_{a_\star^h}^{\mathcal{K}_t,h}(t) \mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t) \middle| \mu_{a_\star^h, b_\star^h}\right) \geqslant f(|\mathcal{K}_t|\, T + t)\right\}. \tag{37}$$

Since we have

$$\begin{aligned}
&\left\{t \geqslant t_\nu^h :\ N_{a_\star^h}^{\mathcal{K}_t,h}(t) \mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t) \middle| \mu_{a_\star^h, b_\star^h}\right) \geqslant f(|\mathcal{K}_t|\, T + t)\right\} \\
&= \bigcup_{K=0}^{h-1} \left\{t \geqslant t_\nu^h :\ |\mathcal{K}_t| = K,\ N_{a_\star^h}^{\mathcal{K}_t,h}(t) \mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t) \middle| \mu_{a_\star^h, b_\star^h}\right) \geqslant f(KT + t)\right\}
\end{aligned}$$

by using the inclusion from Eq. 37 plus the union bound, it comes

$$\left|\overline{\mathcal{U}}^h\right| \leqslant \sum_{K=0}^{h-1} \sum_{t=t_\nu^h}^{T} \mathbb{I}_{\left\{|\mathcal{K}_t| = K,\ N_{a_\star^h}^{\mathcal{K}_t,h}(t) \mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t) \middle| \mu_{a_\star^h, b_\star^h}\right) \geqslant f(KT+t)\right\}}. \tag{38}$$

By taking the expectation on both sides in previous inequality (Eq. 38), we have

$$\mathbb{E}_\nu\left[\left|\overline{\mathcal{U}}^h\right|\right] \leqslant \sum_{K=0}^{h-1} \sum_{t=t_\nu^h}^{T} \mathbb{P}_\nu\left(|\mathcal{K}_t| = K,\ N_{a_\star^h}^{\mathcal{K}_t,h}(t) \mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t) \middle| \mu_{a_\star^h, b_\star^h}\right) \geqslant f(KT+t)\right). \tag{39}$$

Combining Eq. 39 and Lemma 14, it comes

$$\begin{aligned}
\mathbb{E}_\nu\left[\left|\overline{\mathcal{U}}^h\right|\right] &\leqslant \sum_{K=0}^{h-1} \sum_{t=t_\nu^h}^{T} (KT+t)^{-1} \log(KT+t)^{-2} \\
&\leqslant \sum_{t \geqslant t_\nu^h} t^{-1} \log(t)^{-2}.
\end{aligned}$$

This implies $\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{U}}^h\Big|\Big] \leqslant 2$ since it can be shown that

$$\sum_{t \geqslant t_\nu^h} t^{-1}\log(t)^{-2} \leqslant \int_{t=t_\nu^h-1}^{\infty} t^{-1}\log(t)^{-2}\,\mathrm{d}t = \frac{1}{\log(t_\nu^h-1)} \leqslant \frac{1}{\log(2)} \leqslant 2\,.$$

Subsets $\mathcal{C}_{a,\varepsilon}^h$, $\mathcal{C}_\varepsilon^h$, $\overline{\mathcal{C}}_{a,\varepsilon}^h$ and $\overline{\mathcal{C}}_\varepsilon^h$: We conclude the proof of Lemma 10 by taking the expectation on both sides in the following inequalities and by using the bounds on subsets $\mathcal{T}^h$, $\mathcal{U}^h$, $\overline{\mathcal{U}}^h$, $\mathcal{E}_{a,\varepsilon}^h$ and $\overline{\mathcal{E}}_{a,\varepsilon}^h$.

$$\begin{aligned}
\big|\mathcal{C}_{a,\varepsilon}^h\big| &\leqslant \big|\mathcal{U}^h\big| + \big|\mathcal{E}_{a,\varepsilon}^h\big| \\
\big|\mathcal{C}_\varepsilon^h\big| &\leqslant \big|\mathcal{U}^h\big| + \sum_{a\neq a_\star^h} \big|\mathcal{E}_{a,\varepsilon}^h\big| \\
\Big|\overline{\mathcal{C}}_{a,\varepsilon}^h\Big| &\leqslant \big|\mathcal{T}^h\big| + \Big|\overline{\mathcal{U}}^h\Big| + \Big|\overline{\mathcal{E}}_{a,\varepsilon}^h\Big| \\
\Big|\overline{\mathcal{C}}_\varepsilon^h\Big| &\leqslant \big|\mathcal{T}^h\big| + \Big|\overline{\mathcal{U}}^h\Big| + \sum_{a\neq a_\star^h} \Big|\overline{\mathcal{E}}_{a,\varepsilon}^h\Big|\,.
\end{aligned}$$

□

B.3 Proof of Lemma 3

From Lemma 10, we have the bound $\mathbb{E}_\nu\big[\big|\mathcal{C}_{a,\varepsilon}^h\big|\big] \leqslant 4\varepsilon^{-2}+2$ and $\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{C}}_{a,\varepsilon}^h\Big|\Big] \leqslant 4\varepsilon^{-2}+3$.

Let us consider $t > |\mathcal{A}|$ such that $t \notin \mathcal{C}_{a,\varepsilon}^h = \mathcal{E}_{a,\varepsilon}^h \cup \mathcal{U}_a^h$ and $a_{t+1}^h = a$. By definition of the index (Eq. 6), we have

$$N_a^h(t)\mathrm{KL}\big(\widehat{\mu}_a^h(t)\big|\,U_a^h(t)\big) = f(t)\,. \tag{40}$$

Since $a_{t+1}^h = a$, it follows from the `KLUCB-RB` strategy that

$$u_{a_\star^h}^h(t) \leqslant u_a^h(t) \leqslant U_a^h(t)\,. \tag{41}$$

Since $a_{t+1}^h = a$, we have $t \notin \mathcal{U}^h$ and

$$\mu_\star^h \leqslant u_{a_\star^h}^h(t) = U_{a_\star^h}^h(t)\,. \tag{42}$$

Since $\varepsilon < \varepsilon_\nu$ and since a is a sub-optimal arm, we have

$$\mu_a^h + \varepsilon < \mu_\star^h\,. \tag{43}$$

Since $a_{t+1}^h = a$, we have $t \notin \mathcal{E}_{a,\varepsilon}^h$ and

$$\widehat{\mu}_a^h(t) \leqslant \mu_a^h + \varepsilon\,. \tag{44}$$

Then Eq. 41, 42, 43 and 44 imply

$$\widehat{\mu}_a^h(t) \leqslant \mu_a^h + \varepsilon < \mu_\star^h \leqslant U_a^h(t)\,. \tag{45}$$

Combining Eq. 40 and Eq. 45, it holds

$$\mathrm{KL}\big(\mu_a^h + \varepsilon \big| \, \mu_\star^h\big) \leqslant \mathrm{KL}\big(\widehat{\mu}_a^h(t) \big| \, U_a^h(t)\big) \quad \text{and} \quad N_a^h(t)\mathrm{KL}\big(\mu_a^h + \varepsilon \big| \, \mu_\star^h\big) \leqslant f(t)\,.$$

Let us consider $t \geqslant t_\nu^h$ such that $t \notin \overline{\mathcal{C}}_{a,\varepsilon}^h = \mathcal{T} \cup \overline{\mathcal{E}}_{a,\varepsilon}^h \cup \overline{\mathcal{U}}_a^h$ and $a_{t+1}^h = a$. By definition of the index (Eq. 7), we have

$$\overline{N}_a^h(t)\mathrm{KL}\Big(\overline{\mu}_a^h(t) \Big| \, \overline{U}_a^h(t)\Big) = f\Big(\overline{K}_t^h T + t\Big)\,. \tag{46}$$

Since $a_{t+1}^h = a$, it follows from the `KLUCB-RB` strategy that

$$u_{a_\star^h}^h(t) \leqslant u_a^h(t) \leqslant \overline{U}_a^h(t)\,. \tag{47}$$

Since $a_{t+1}^h = a$, we have $t \notin \mathcal{T}^h \cup \overline{\mathcal{U}}^h$ and

$$\mu_\star^h \leqslant u_{a_\star^h}^h(t) = \overline{U}_{a_\star^h}^h(t)\,. \tag{48}$$

Since $\varepsilon < \varepsilon_\nu$ and since a is a sub-optimal arm, we have

$$\mu_a^h + \varepsilon < \mu_\star^h\,. \tag{49}$$

Since $a_{t+1}^h = a$, we have $t \notin \mathcal{T}^h \cup \overline{\mathcal{E}}_{a,\varepsilon}^h$ and

$$\overline{\mu}_a^h(t) \leqslant \mu_a^h + \varepsilon\,. \tag{50}$$

Then Eq. 47, 48, 49 and 50 imply

$$\overline{\mu}_a^h(t) \leqslant \mu_a^h + \varepsilon < \mu_\star^h \leqslant \overline{U}_a^h(t)\,. \tag{51}$$

Combining Eq. 46 and Eq. 51, it holds

$$\mathrm{KL}\big(\mu_a^h + \varepsilon \big| \, \mu_\star^h\big) \leqslant \mathrm{KL}\Big(\overline{\mu}_a^h(t) \Big| \, \overline{U}_a^h(t)\Big) \quad \text{and} \quad \overline{N}_a^h(t)\mathrm{KL}\big(\mu_a^h + \varepsilon \big| \, \mu_\star^h\big) \leqslant f\Big(\overline{K}_t^h T + t\Big)\,.$$

In order to conclude the proof it remains to show that $\overline{K}_t^h \leqslant \beta_{b_\star^h}^{h-1}(h-1)$. Since $a_{t+1}^h = a$, we have $t \notin \mathcal{T}^h$ and we deduce from the definition of $\mathcal{T}^h$ (see Eq. 9) that

$$\overline{K}_t^h = \big|\mathcal{K}_\star^h(t)\big| \leqslant \big|\big\{k \in [\![1, h-1]\!]:\ b_\star^k = b_\star^h\big\}\big| = \ \beta_{b_\star^h}^{h-1}(h-1)\,,$$

where $\mathcal{K}_\star^h(t) \coloneqq \big\{k \in [\![1, h-1]\!] \colon b_\star^k = b_\star^h \text{ and } \overline{a}_\star^k = \overline{a}_t^h\big\}$.

Finally, we prove the last statement of Lemma 3. For all sub-optimal arm $a \in \mathcal{A}$, for all period $h \geqslant 1$, for all time step $t > |\mathcal{A}|$, we denote by

$$\tau_a^h(t) = \max\left\{t' \in [\![|\mathcal{A}|+1\,;t]\!] :\ a_{t'+1}^h = a \quad \text{and} \quad t' \notin \mathcal{C}_{a,\varepsilon}^h\right\} \tag{52}$$

the last time step before time step t that does not belong to $\mathcal{C}_{a,\varepsilon}^h$ such that we pull arm a in period h. In particular, we have

$$N_a^h\big(\tau_a^h(t)\big) \leqslant \frac{f\big(\tau_a^h(t)\big)}{\mathrm{KL}(\mu_a^h+\varepsilon|\,\mu_\star^h)} \leqslant \frac{f(t)}{\mathrm{KL}(\mu_a^h+\varepsilon|\,\mu_\star^h)}\,. \tag{53}$$

Then, from Eq. 52 and Eq. 53 we have

$$\begin{aligned}
N_a^h(t) &= N_a^h(|\mathcal{A}|+1) + \sum_{t'>|\mathcal{A}|}^{t-1} \mathbb{I}_{\left\{a_{t'+1}^h = a\right\}} \\
&= N_a^h(|\mathcal{A}|+1) + \sum_{t'>|\mathcal{A}|}^{t-1} \mathbb{I}_{\left\{a_{t'+1}^h = a,\ t' \in \mathcal{C}_{a,\varepsilon}^h\right\}} + \sum_{t'>|\mathcal{A}|}^{t-1} \mathbb{I}_{\left\{a_{t'+1}^h = a,\ t' \notin \mathcal{C}_{a,\varepsilon}^h\right\}} \\
&\leqslant N_a^h(|\mathcal{A}|+1) + \left|\mathcal{C}_{a,\varepsilon}^h\right| + N_a^h\big(\tau_a^h(t)\big) \\
&\leqslant N_a^h(|\mathcal{A}|+1) + \left|\mathcal{C}_{a,\varepsilon}^h\right| + \frac{f(t)}{\mathrm{KL}(\mu_a^h+\varepsilon|\,\mu_\star^h)}\,.
\end{aligned}$$

B.4 Proof of Lemma 4

Let us consider $t \geqslant t_\nu^h$ such that $t \notin \mathcal{T}^h$ and $\overline{a}_t^h \neq a_\star^h$. Since $t \notin \mathcal{T}^h$ (see Eq. 9),

$$\forall a \in \mathcal{A}, \quad \overline{N}_a^h(t) = N_a^h(t) + \sum_{k \in \mathcal{K}_\star^h(t)} N_a^k(T)\,, \tag{54}$$

where $\mathcal{K}_\star^h(t) \coloneqq \left\{k \in [\![1, h-1]\!] : \overline{a}_\star^k = \overline{a}_t^h\right\}$. Since for all $k \in \mathcal{K}_\star^h$, $\overline{a}_\star^k \in \operatorname{argmax}_{a\in\mathcal{A}} N_a^k(T)$, from Eq. 54 we deduce that $\overline{a}_t^h \in \operatorname{argmax}_{a\in\mathcal{A}} \overline{N}_a^h(t)$. Since $\overline{a}_t^h \neq a_\star^h$, this implies

$$\overline{N}_{a_\star^h}^h(t) \leqslant \overline{N}_{\overline{a}_t^h}^h(t) \quad \text{and} \quad \overline{N}_{\overline{a}_t^h}^h(t) \leqslant \sum_{a \neq a_\star^h} \overline{N}_a^h(t)\,. \tag{55}$$

Furthermore, since $t \notin \mathcal{T}^h$ (see Eq. 9), we have

$$\overline{K}_t^h \coloneqq \left|\mathcal{K}_+^h(t)\right| = \left|\mathcal{K}_\star^h(t)\right|\,. \tag{56}$$

Then it comes

$$\overline{N}_{a_\star^h}^h(t) = \left|\mathcal{K}_\star^h(t)\right| T + t - \sum_{a \neq a_\star^h} \overline{N}_a^h(t)\,. \tag{57}$$

Then Eq. 54, 55 and 57 imply

$$\frac{\left|\mathcal{K}_\star^h(t)\right| T}{2} + \frac{t}{2} \leqslant \sum_{a \neq a_\star^h} N_a^h(t) + \sum_{k \in \mathcal{K}_\star^h(t)} N_a^k(T)\,. \tag{58}$$

For $a \neq a_\star^h$ and for $k \in \mathcal{K}_\star^h(t)$, the arm a is sub-optimal for bandit $b_\star^k = b_\star^h$. Thus, from Lemma 3, we have

$$\forall a \neq a_\star^h, \forall k \in \mathcal{K}_\star^h(t),\ N_a^h(t) \leqslant \frac{f(t)}{\mathrm{KL}(\mu_a^h + \varepsilon | \, \mu_\star^h)} + \left|\mathcal{C}_{a,\varepsilon}^h\right| + N_a^h(|\mathcal{A}|+1) \quad (59)$$
$$N_a^k(T) \leqslant \frac{f(T)}{\mathrm{KL}(\mu_a^k + \varepsilon | \, \mu_\star^k)} + \left|\mathcal{C}_{a,\varepsilon}^k\right| + N_a^k(|\mathcal{A}|+1)\ .$$

Then, by combining Eq. 58 and Eq. 59, we get

$$\begin{aligned} &\frac{t}{2} - \left(\sum_{a \neq a_\star^h} \frac{f(t)}{\mathrm{KL}(\mu_a^h + \varepsilon | \, \mu_\star^h)} + N_a^h(|\mathcal{A}|+1) \right) \\ &+ \frac{\left|\mathcal{K}_\star^h(t)\right| T}{2} - \left(\sum_{k \in \mathcal{K}_\star^h(t)} \sum_{a \neq a_\star^h} \frac{f(T)}{\mathrm{KL}(\mu_a^k + \varepsilon | \, \mu_\star^k)} + N_a^k(|\mathcal{A}|+1) \right) \\ &\leqslant \sum_{k \in \mathcal{K}_\star^h(t) \cup \{h\}} \sum_{a \neq a_\star^h} \left|\mathcal{C}_{a,\varepsilon}^k\right| \end{aligned} \quad (60)$$

We finally prove Lemma 4 from Eq. 60 and the following inequalities

$$\begin{aligned} &\forall k \in \mathcal{K}_\star^h(t) \cup \{h\}\,, \quad a_\star^k = a_\star^h\,, \\ &\forall k \in \mathcal{K}_\star^h(t) \cup \{h\}\,, \quad \sum_{a \neq a_\star^h} N_a^k(|\mathcal{A}|+1) = \sum_{a \neq a_\star^k} N_a^k(|\mathcal{A}|+1) \leqslant |\mathcal{A}|\ , \\ &\forall k \in \mathcal{K}_\star^h(t) \cup \{h\}\,, \quad \sum_{a \neq a_\star^h} \left|\mathcal{C}_{a,\varepsilon}^k\right| = \sum_{a \neq a_\star^k} \left|\mathcal{C}_{a,\varepsilon}^k\right| = \left|\mathcal{C}_\varepsilon^k\right|\ . \end{aligned}$$

B.5 Proof of Proposition 2

We first deduce Lemma 11 from Lemma 4.

Lemma 11 (Conditions for misidentifying the best arms). *For all period $h \geqslant 1$, for all $0 < \varepsilon < \varepsilon_\nu$, for all $t \geqslant T_{\nu,\varepsilon}^h$,*

$$\left(\ t \notin \mathcal{T}^h \ \text{ and } \ \overline{a}_t^h \neq a_\star^h \ \right) \quad \Longleftrightarrow \quad \left(\ t < 4\left|\mathcal{C}_\varepsilon^h\right| \ \text{ or } \ \exists k \in [\![1,h]\!],\, T < 8\left|\mathcal{C}_\varepsilon^k\right| \ \right).$$

This implies

$$\left(\ T \notin \mathcal{T}^h \ \text{ and } \ \overline{a}_\star^h \neq a_\star^h \ \right) \quad \Longleftrightarrow \quad \exists k \in [\![1,h]\!],\, T < 8\left|\mathcal{C}_\varepsilon^k\right|\ .$$

We respectively refer to Proposition 2, Eq. 9 and Eq. 11 for the definitions of $T_{\nu,\varepsilon}^h$, $\mathcal{T}^h$ and $\mathcal{C}_\varepsilon^h$.

The proof of Lemma 11 is deferred to the Section B.5. We prove Proposition 2 in the following.

Let us introduce the subset $\mathcal{P}$ of pairs period-time when there is false positives

or false negatives, or when the mean of the current pulled arm is underestimated, or when the index of the best arm is below its mean, or when the most pulled arms are different from the best arms. More formally,

$$\mathcal{P} \coloneqq \left\{(h,t)\in[\![1,h]\!]\times[\![1,T]\!]:\ \left(t\geqslant T^h_{\nu,\varepsilon}, t\in\overline{\mathcal{C}}^h_\varepsilon\cup\mathcal{M}^h_\varepsilon\right)\ \text{ or }\ \left(\exists k\in[\![1,h-1]\!], T\in\mathcal{T}^k\cup\mathcal{M}^k_\varepsilon\right)\right\}, \tag{61}$$

where $\mathcal{M}^h_\varepsilon \coloneqq \left\{t\geqslant T^h_{\nu,\varepsilon} : t\notin\mathcal{T}^h \text{ and } \overline{a}^h_t\neq a^h_\star\right\}$, for all period $h\geqslant 1$.

Then, for a bandit $b\in\mathcal{B}$ and a sub-optimal arm $a\in\mathcal{A}$, from Lemma 3 we have

$$N_{a,b}(H,T) \leqslant \frac{f\left(\beta_b^H HT\right)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\,\mu_b^\star)} + \sum_{h=1}^{H}\sum_{t=0}^{T-1}\mathbb{I}_{\left\{b^h_\star=b,\ a^h_{t+1}=a,\ t<T^h_{\nu,\varepsilon}\text{ or }(h,t)\in\mathcal{P}\right\}}. \tag{62}$$

From the definitions of $\mathcal{P}$ (Eq. 61), $T^h_{\nu,\varepsilon}$ (Proposition 2) and $\overline{\mathcal{C}}^h_\varepsilon$ (Eq. 11) for $h\geqslant 1$, this implies

$$\begin{aligned}
N_{a,b}(H,T) \leqslant{}& \frac{f\left(\beta_b^H HT\right)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\,\mu_b^\star)} \\
&+ \sum_{h=1}^{H}\sum_{t=0}^{T-1}\mathbb{I}_{\left\{b^h_\star=b,\ a^h_{t+1}=a,\ t<T^h_{\nu,\varepsilon}\right\}} \\
&+ \sum_{h=1}^{H}\sum_{t=0}^{T-1}\mathbb{I}_{\left\{b^h_\star=b,\ a^h_{t+1}=a,\ t\in\overline{\mathcal{C}}^h_\varepsilon\right\}} \\
&+ \sum_{h=1}^{H}\sum_{t=0}^{T-1}\mathbb{I}_{\left\{b^h_\star=b,\ a^h_{t+1}=a,\ t\notin\overline{\mathcal{C}}^h_{a,\varepsilon}\right\}}^{\mathbb{I}_{\{\exists k\in[\![1,h-1]\!],T\in\mathcal{T}^k\}}} \\
&+ \sum_{h=1}^{H}\sum_{t=0}^{T-1}\mathbb{I}_{\left\{b^h_\star=b,\ a^h_{t+1}=a,\ t\notin\overline{\mathcal{C}}^h_{a,\varepsilon}\right\}}^{\mathbb{I}_{\{t\in\mathcal{M}^h_\varepsilon\text{ or }\exists k\in[\![1,h-1]\!],T\in\mathcal{M}^k_\varepsilon\}}}.
\end{aligned} \tag{63}$$

Furthermore, from Lemma 11 we have for all period $h\geqslant 1$,

$$\mathbb{I}_{\{t\in\mathcal{M}^h_\varepsilon\text{ or }\exists k\in[\![1,h-1]\!],T\in\mathcal{M}^k_\varepsilon\}} \leqslant \mathbb{I}_{\{t<4|\mathcal{C}^h_\varepsilon|\}} + \sum_{k=1}^{h}\mathbb{I}_{\{T<8|\mathcal{C}^k_\varepsilon|\}}. \tag{64}$$

By combining Eq.63 and Eq.64, we get

$$\begin{aligned}
N_{a,b}(H,T) \leqslant{}& \frac{f\left(\beta_b^H HT\right)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\,\mu_b^\star)} \\
&+\sum_{h=1}^{H}\mathbb{I}_{\{b^h_\star=b\}}\left[T^h_{\nu,\varepsilon}+4\left|\mathcal{C}^h_\varepsilon\right|+\left|\overline{\mathcal{C}}^h_\varepsilon\right|+\left(\sum_{t=0}^{T-1}\mathbb{I}_{\left\{b^h_\star=b,\ a^h_{t+1}=a,\ t\notin\overline{\mathcal{C}}^h_{a,\varepsilon}\right\}}\right)\left(\sum_{k=1}^{h}\mathbb{I}_{\{T<8|\mathcal{C}^k_\varepsilon|\}}+\mathbb{I}_{\{T\in\mathcal{T}^k\}}\right)\right].
\end{aligned} \tag{65}$$

Since the arm a is sub-optimal for the bandit b, the consistency (Lemma 3) implies

$$\forall h\geqslant 1,\quad \sum_{t=0}^{T-1}\mathbb{I}_{\left\{b^h_\star=b,\ a^h_{t+1}=a,\ t\notin\overline{\mathcal{C}}^h_{a,\varepsilon}\right\}} \leqslant \frac{f(hT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\,\mu_b^\star)}. \tag{66}$$

In addition, the following Markov's type inequalities are satisfied

$$\forall k \geqslant 1, \quad \mathbb{I}_{\{T<8|\mathcal{C}_\varepsilon^k|\}} \leqslant \frac{8\left|\mathcal{C}_\varepsilon^k\right|}{T}\,. \tag{67}$$

By combining Eq. 65, 66 and 67, we prove Proposition 2, that is

$$\begin{aligned} &N_{a,b}(H,T) \\ &\leqslant \frac{f(\beta_b^H HT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)} + \sum_{h=1}^{H} \mathbb{I}_{\{b_\star^h=b\}}\left[T_{\nu,\varepsilon}^h + 4\left|\mathcal{C}_\varepsilon^h\right| + \left|\overline{\mathcal{C}}_\varepsilon^h\right| + \frac{f(hT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\,\mu_b^\star)}\sum_{k=1}^{h}\frac{8|\mathcal{C}_\varepsilon^k|}{T} + \mathbb{I}_{\{T\in\mathcal{T}^k\}}\right]. \end{aligned}$$

Proof of Lemma 11 Let us consider a period $h \geqslant 1$, $0<\varepsilon<\varepsilon_\nu$, and a time step all $t \geqslant T_{\nu,\varepsilon}^h$ such that $t \notin \mathcal{T}^h$ and $\overline{a}_t^h \neq a_\star^h$. Then, since $T_{\nu,\varepsilon}^h \geqslant t_\nu^h$, from Lemma 4 we have

$$\frac{t+\left|\mathcal{K}_\star^h(t)\right|T}{2} - \left(1+\left|\mathcal{K}_\star^h(t)\right|\right)|\mathcal{A}| - \left(f(t)+\left|\mathcal{K}_\star^h(t)\right|f(T)\right)\sum_{a\neq a_\star^h}\frac{1}{\mathrm{KL}\left(\mu_a^h+\varepsilon|\,\mu_\star^h\right)} \leqslant \sum_{k\in\mathcal{K}_\star^h(t)\cup\{h\}}\left|\mathcal{C}_\varepsilon^k\right|. \tag{68}$$

Furthermore, by definition of $T_{\nu,\varepsilon}^h$, since $t \geqslant T_{\nu,\varepsilon}^h$, we have

$$\begin{aligned} \frac{t}{2} - \textstyle\sum_{a\neq a_\star^h}\dfrac{f(t)}{\mathrm{KL}(\mu_a^h+\varepsilon|\,\mu_\star^h)} - |\mathcal{A}| &> \frac{t}{4} \\ \frac{T}{2} - \textstyle\sum_{a\neq a_\star^h}\dfrac{f(T)}{\mathrm{KL}(\mu_a^h+\varepsilon|\,\mu_\star^h)} - |\mathcal{A}| &> \frac{T}{4}\,. \end{aligned} \tag{69}$$

By respectively combining Eq. 68 and Eq. 69, we thus deduce

$$\begin{aligned} \left|\mathcal{K}_\star^h(t)\right| = 0 \quad &\Rightarrow \quad t \leqslant 4\left|\mathcal{C}_\varepsilon^h\right| \\ \left|\mathcal{K}_\star^h(t)\right| \geqslant 1 \quad &\Rightarrow \quad \exists k \in [\![1,h]\!], T \leqslant 8\left|\mathcal{C}_\varepsilon^k\right| \end{aligned}$$

which implies Lemma 11.

B.6 Tools from Concentration of Measure

This subsection gathers useful concentration lemmas that do not depend on the considered strategy.

Notations For all period $h \geqslant 2$, for all time step $t > |\mathcal{A}|$, for each (possible random) subset of past periods $\mathcal{K} \subset \mathcal{K}^h := \left\{k \in [\![1,h-1]\!] : b_\star^k = b_\star^h\right\}$, for all arm $a \in \mathcal{A}$, we define $N_a^{\mathcal{K},h}(t) := \sum_{k\in\mathcal{K}} N_a^k(T) + N_a^h(t)$, $S_a^{\mathcal{K},h}(t) := \sum_{k\in\mathcal{K}} S_a^k(T) + S_a^h(t)$ and $\widehat{\mu}_a^{\mathcal{K},h}(t) := S_a^{\mathcal{K},h}(t)/N_a^{\mathcal{K},h}(t)$.

In particular, for $\mathcal{K} = \mathcal{K}_\star^h(t) := \left\{k \in [\![1,h-1]\!] : b_\star^k = b_\star^h \text{ and } \overline{a}_\star^k = \overline{a}_t^h\right\}$, we have $N_a^{\mathcal{K}_\star^h(t),h}(t) = \overline{N}_a^h(t)$ and $\widehat{\mu}_a^{\mathcal{K}_\star^h(t),h}(t) = \overline{\mu}_a^h(t)$ when $t \geqslant t_\nu^h$ and $t \notin \mathcal{T}^h$ (see Eq. 8 and 9).

Uniform bounds based on the Laplace method (method of mixtures for sub-Gaussian random variables, see [21]) are given in Lemma 12.

Lemma 12 (Uniform sub-Gaussian concentration). *For all period $h \geqslant 2$, for all time step $t > |\mathcal{A}|$, for all arm $a \in \mathcal{A}$, for all $\delta \in (0,1)$, it holds*

$$\begin{aligned} \mathbb{P}_\nu\big(\widehat{\mu}_a^h(t) - \mu_{a,b_\star^h} \geqslant \ \mathrm{d}\big(N_a^h(t),\delta\big)\big) &\leqslant \delta \\ \mathbb{P}_\nu\big(\mu_{a,b_\star^h} - \widehat{\mu}_a^h(t) \geqslant \ \mathrm{d}\big(N_a^h(t),\delta\big)\big) &\leqslant \delta\,, \end{aligned}$$

where $\mathrm{d}(n,\delta) = \sqrt{2(1+1/n)\log\big(\sqrt{n+1}/\delta\big)/n}$, *for all* $n \geqslant 1$.

Lemma 13 reformulates Lemma B.1 from [7].

Lemma 13 (Concentration inequalities). *For all period $h \geqslant 2$, for all arm $a \in \mathcal{A}$, for all $\varepsilon \in (0,1/2)$, and all possibly random subset of periods $\mathcal{K}_t$ such that the random variable $N_a^{\mathcal{K}_t,h}(t)$ is a random stopping time, it holds*

$$\sum_{t\geqslant 1} \mathbb{P}_\nu\big(a_{t+1}^h = a,\ \big|\widehat{\mu}_a^{\mathcal{K}_t,h}(t) - \mu_a^h\big| \geqslant \varepsilon\big) \leqslant 4\varepsilon^{-2}\,.$$

Lemma 14 reformulates Theorem 1 from [8].

Lemma 14 (Self-normalized inequalities). *For all period $h \geqslant 2$, for all time step $t > |\mathcal{A}|$, for all arm $a \in \mathcal{A}$, for all $K \in [\![0, h-1]\!]$, for all $\delta > 0$ and all possibly random subset of periods $\mathcal{K}_t$ such that the random variable $N_a^{\mathcal{K}_t,h}(t)$ is a random stopping time, it holds*

$$\mathbb{P}_\nu\big(|\mathcal{K}_t| = K,\ N_a^{\mathcal{K}_t,h}(t)\mathrm{KL}\big(\widehat{\mu}_a^{\mathcal{K}_t,h}(t)\big|\,\mu_{a,b_\star^h}\big) \geqslant \delta\big) \leqslant 2e\lceil \delta\log(KT+t)\rceil \exp(-\delta)\,.$$

In particular, this implies for $\delta = f(KT+t)$,

$$\mathbb{P}_\nu\big(|\mathcal{K}_t| = K,\ N_a^{\mathcal{K}_t,h}(t)\mathrm{KL}\big(\widehat{\mu}_a^{\mathcal{K}_t,h}(t)\big|\,\mu_{a,b_\star^h}\big) \geqslant f(KT+t)\big) \leqslant (KT+t)^{-1}\log(KT+t)^{-2}\,.$$

C Additional Experiments: Ideal Cases for which Bandits are Close Enough on the Subset of Optimal Arms

This section provides additional experiments where we investigate some favorable distributions ν where it is hard to separate the different bandits from each other. All experiments are repeated 100 times.

C.1 A Single Instance

Let us first make a remark in the trivial limit case of a unique bandit, that is to say $\mathcal{B} = \{b\}$. In such cases, playing `KLUCB-RB` is obviously equivalent to playing for $T_{total} := HT$ rounds a `KLUCB` strategy on the bandit instance b, with an additional term $(H-1)\sum_{a\in\mathcal{A}} \Delta_{a,b}$ in the final cumulative regret due to the initialization at each period. Figure 5 highlights this fact for the two-armed

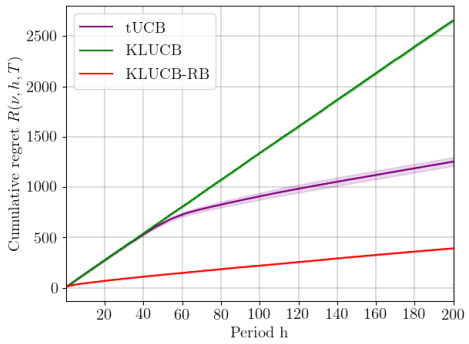


Fig. 5: Cumulative regret of `KLUCB`, `KLUCB-RB` and `tUCB` along $H = 200$ periods of $T = 10^3$ rounds for the bandit set $\mathcal{B} = \{b\}$.

bandit b defined in Eq. 70, over $H = 200$ periods of $T = 10^3$ rounds.

$$b : (\mu_{1,b}, \mu_{2,b}) = (-\frac{\Delta}{2}, \frac{\Delta}{2}) \qquad \text{where} \qquad \Delta = 10\sqrt{\frac{\log(HT)}{T}}. \tag{70}$$

Although the case $|\mathcal{B}| = 1$ is not an interesting one since there is no switches between different bandits instances, it enables to understand what happens when $|\mathcal{B}| > 1$ and bandits are similar, that is $\max_{a \in \mathcal{A}^\star} \max_{b \neq b'} |\mu_{a,b} - \mu_{a,b'}|$ approaches 0. Besides, it highlights the need for `tUCB` to see a sufficient number of periods before exploiting the estimated models of the bandits.

C.2 Similarity of Different Instances on the Optimal Subset $\mathcal{A}^\star$

Let us consider routines over two bandits b_1 and b_2 composed of two arms such that $(\mu_{1,b_2}, \mu_{2,b_2}) = (\mu_{1,b_1} + \gamma, \mu_{2,b_1} - \gamma)$ and $a^\star_{b_1} = 2$. If $\gamma > \Delta_{2,b_1}/2$, arms arrangements are different in both instances and these cases are studied in subsection 5.1. Otherwise we have $a^\star_{b_1} = a^\star_{b_2} = 2$ if ever $0 < \gamma < \Delta_{2,b_1}/2$. Although separation of instances is particularly hard in such cases, samples aggregation from false positive periods does not perturb the empirical means arrangement, and thus yields great performances for `KLUCB-RB`. To explain how this kind of distribution generalizes to settings composed of arbitrary numbers of bandits and arms, we present in Fig. 6b a distribution ν such that $|\mathcal{B}| = 5$ and $|\mathcal{A}| = 4$. In this setting, we have $\mathcal{A}^\star = \{1, 4\}$. Considering distributions of bandits restricted to $\mathcal{A}^\star$, $\mathcal{B}$ naturally decomposes into 3 clusters $\mathcal{C}^{(1)} := \{b_1, b_4\}$, $\mathcal{C}^{(2)} := \{b_2, b_3\}$ and $\mathcal{C}^{(3)} := \{b_5\}$ so that

$$\forall i \in \{1,2,3\}, \forall b, b' \in \mathcal{C}^{(i)}, \forall a \in \mathcal{A}^\star, |\mu_{a,b} - \mu_{a,b'}| < \frac{1}{2} \min_{y \in \mathcal{C}^{(i)}} \min_{x \in \mathcal{A}, x \neq a^\star_y} \Delta_{x,y} \tag{71}$$

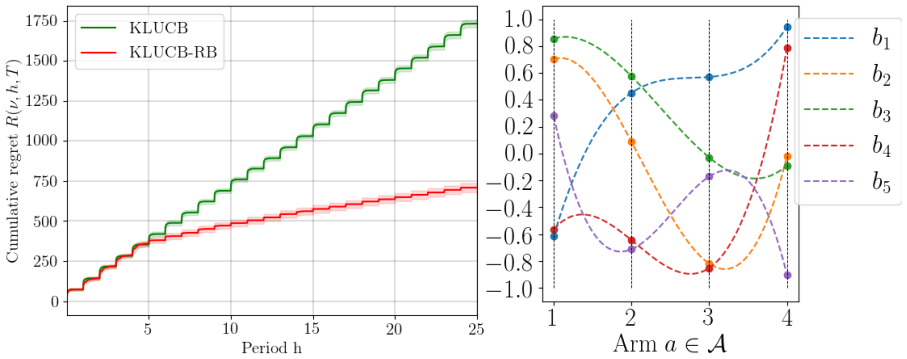
which entails in particular

$$\forall i \in \{1,2,3\}, \exists a^{(i)} \in \mathcal{A}^\star, \forall b \in \mathcal{C}^{(i)}, a^\star_b = a^{(i)} \quad ,$$

and

$$\forall i \in \{1,2,3\}, \forall b \in \mathcal{C}^{(i)}, \forall b' \notin \mathcal{C}^{(i)}, |\mu_{a^{(i)},b} - \mu_{a^{(i)},b'}| > \min_{x \in \mathcal{C}^{(i)}} \min_{a \neq a^\star_x} \Delta_{a,x}. \quad (72)$$

On the one hand, Eq. 71 sums up that different bandits from a same cluster are hard to distinguish, in comparison with the difficulty of learning each instance independently. On the other hand, Eq. 72 implies that clusters are easy to separate from each other. Besides Eq. 71 also implies that the permutation of $\mathcal{A}$ sorting arms according to an increasing order is the same for all instances from a same cluster. Thus `KLUCB-RB` is expected to perform well for this kind of arms distributions. Fig. 6a shows the cumulative regret curves with one standard deviation obtained on this setting, along $H = 25$ periods of $T = 2 \times 10^4$ rounds. As expected, it highlights that a positive cluster effect causes an improvement in regret minimization. In practice, `KLUCB-RB` naturally clusterizes the previously seen periods while the current period index h increases. More specifically, noting $\mathcal{C}(h)$ the cluster containing $b^h_\star$, `KLUCB-RB` makes all the different bandits from $\mathcal{C}(h)$ share their samples with $b^h_\star$ for a large amount of rounds, which enables to boost the minimization of regret across period h.

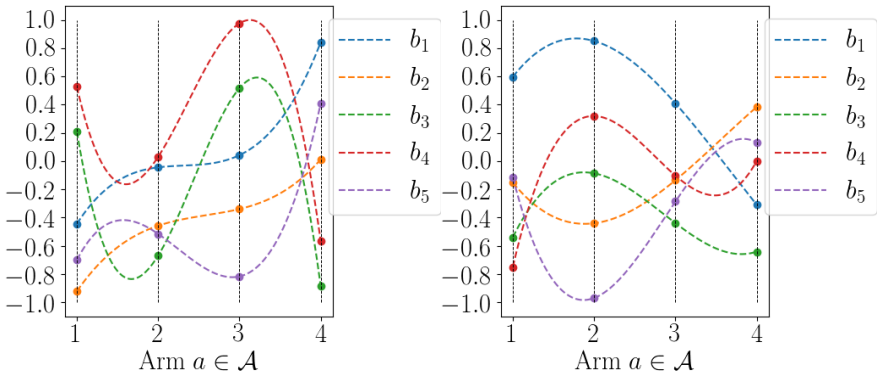


(a) Cumulative regret (b) $\nu : \{\nu_{b_1}, \ldots, \nu_{b_5}\}$

Fig. 6: `KLUCB-RB` and `KLUCB` performances on a clusterized distribution according to $\mathcal{A}^\star$, along $H = 25$ periods of 2×10^4 rounds.

C.3 Complement of Sections 5.2 and 5.3

Figure 7 shows the generated settings used in experiments of Section 5.2, and Figure 8 the setting used in Section 5.3. More specifically, each sub-figure displays the expected reward for each of the four arms, in each of the bandits, for the three considered bandit sets.



(a) $\nu^{(1)} : \{\nu^{(1)}_{b_1}, \ldots, \nu^{(1)}_{b_5}\}$ (b) $\nu^{(2)} : \{\nu^{(2)}_{b_1}, \ldots, \nu^{(2)}_{b_5}\}$

Fig. 7: Distribution ν for each bandit in sets $\mathcal{B}_1$ and $\mathcal{B}_2$.

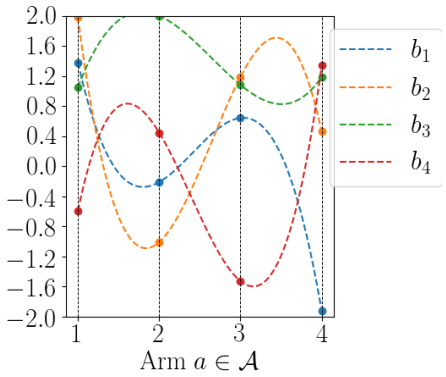


Fig. 8: Distribution ν for each bandit in the critical setting.