The Generalized Likelihood Ratio Test meets klUCB: an Improved Algorithm for Piece-Wise Non-Stationary Bandits

Lilian Besson, Emilie Kaufmann

Proceedings of Machine Learning Research vol XX:1–35, 2019

The Generalized Likelihood Ratio Test meets klUCB: an Improved Algorithm for Piece-Wise Non-Stationary Bandits

Lilian Besson LILIAN.BESSON@CENTRALESUPELEC.FR
CentraleSupélec (campus of Rennes), IETR, SCEE Team,
Avenue de la Boulaie – CS 47601*, F-*35576 *Cesson-Sévigné, France*

Emilie Kaufmann EMILIE.KAUFMANN@UNIV-LILLE.FR
CNRS & Université de Lille, Inria SequeL team
*UMR 9189 – CRIStAL, F-*59000 *Lille, France*

Abstract

We propose a new algorithm for the piece-wise *i.i.d.* non-stationary bandit problem with bounded rewards. Our proposal, GLR-klUCB, combines an efficient bandit algorithm, klUCB, with an efficient, *parameter-free*, change-point detector, the Bernoulli Generalized Likelihood Ratio Test, for which we provide new theoretical guarantees of independent interest. We analyze two variants of our strategy, based on *local restarts* and *global restarts*, and show that their regret is upper-bounded by $\mathcal{O}(\Upsilon_T \sqrt{T \ln(T)})$ if the number of change-points Υ_T is unknown, and by $\mathcal{O}(\sqrt{\Upsilon_T T \ln(T)})$ if Υ_T is known. This improves the state-of-the-art bounds, as our algorithm needs no tuning based on knowledge of the problem complexity other than Υ_T. We present numerical experiments showing that GLR-klUCB outperforms passively and actively adaptive algorithms from the literature, and highlight the benefit of using local restarts.

Keywords: Multi-Armed Bandits; Change Point Detection; Non-Stationary Bandits.

1. Introduction

Multi-Armed Bandit (MAB) problems form a well-studied class of sequential decision making problems, in which an agent repeatedly chooses an action $I_t \in \{1, \dots, K\}$ or "arm" –in reference to the arm of a one-armed bandit– among a set of K arms. In the most standard version of the stochastic bandit model, each arm i is associated with an *i.i.d.* sequence of rewards $(r_{i,t})$ that follow some distribution of mean μ_i. Upon selecting arm I_t, the agent receives the reward $r_{I_t,t}$ associated to the chosen arm, and her goal is to adopt a sequential sampling strategy that maximize the expected sum of these rewards. This is equivalent to minimizing the *regret*, defined as the difference between the total reward of the oracle strategy always selecting the arm with largest mean, μ^*, and that of our strategy: $R_T = \mathbb{E}[\sum_{t=1}^T (\mu^* - \mu_{I_t})]$.

Regret minimization in stochastic bandits has been extensively studied since the works of Robbins (1952) and Lai and Robbins (1985), and several algorithms with a $\mathcal{O}(\ln(T))$ *problem-dependent* regret upper bound have been proposed (see, *e.g.*, Lattimore and Szepesvári (2019) for a survey). Among those, the klUCB algorithm (Cappé et al., 2013) has been shown to be asymptotically optimal for Bernoulli distributions (in that it exactly matches the lower bound given by Lai and Robbins (1985)) and can also be employed when the rewards are assumed to be bounded in $[0, 1]$. *Problem-independent* upper bounds of the form $R_T = \Omega(\sqrt{KT})$ (with no hidden constant depending on the arms distributions) have also been established for stochastic algorithms, like MOSS, or klUCB-Switch by Garivier et al. (2018), while klUCB is known to enjoy a sub-optimal $\mathcal{O}(\sqrt{KT \ln(T)})$ problem-independent regret.

Stochastic bandits were historically introduced as a simple model for clinical trials, where arms correspond to some treatments with unknown efficacy (Thompson, 1933). More recently, MAB models have been proved useful for different applications, like cognitive radio, where arms can model the vacancy of radio channels, or parameters of a dynamically configurable radio hardware (Maghsudi and Hossain, 2016; Bonnefoi et al., 2017; Kerkouche et al., 2018). Another application is the design of recommender systems, where arms model the popularity of different items (*e.g.*, news recommendation, Li et al. (2010)).

For both cognitive radio and recommender systems, the assumption that the arms distribution *do not evolve over time* may be a big limitation. Indeed, in cognitive radio new devices can enter or leave the network, which impacts the availability of the radio channel they use to communicate; whereas in online recommendation, the popularity of items is also subject to trends. Hence, there has been some interest on how to take those *non-stationary* aspects into account within a multi-armed bandit model.

A first possibility to cope with non-stationary is to model the decision making problem as an *adversarial bandit problem* (Auer et al., 2002b). Under this model, rewards are completely arbitrary and are not assumed to follow any probability distribution. For adversarial environments, the pseudo-regret, which compares the accumulated reward of a given strategy with that of the best fixed-arm policy, is often studied. The pseudo regret of the EXP3 algorithm has been shown to be $\mathcal{O}(\sqrt{KT})$, which matches the lower bound given by Auer et al. (2002b). However, this model is a bit too general for the considered applications, where reward distributions do not necessarily vary at every round. For these reasons, an intermediate model, called the *piece-wise stationary MAB*, has been introduced by Kocsis and Szepesvári (2006) and Yu and Mannor (2009). In this model, described in full details in Section 2, the (random) reward of arm i at round t has some mean $\mu_i(t)$, that is constant on intervals between two *breakpoints*, and the regret is measured with respect to the *current* best arm $i_t^* = \arg\max_i \mu_i(t)$.

In this paper, we propose a new algorithm for the piece-wise stationary bandit problem with bounded rewards, called GLR-klUCB. Like previous approaches – CUSUM (Liu et al., 2018) and $\mathrm{M\text{-}UCB}$ (Cao et al., 2019) – our algorithm relies on combining a standard multi-armed bandit algorithm with a change-point detector. For the bandit component, we propose the use of the klUCB algorithm, that is known to outperform $\mathrm{UCB1}$ (Auer et al., 2002a) used in previous works. For the change-point detector, we propose the Bernoulli Generalized Likelihood Ratio Test (GLRT), for which we provide new non-asymptotic properties that are of independent interest. This choice is particularly appealing because unlike previous approaches, the Bernoulli GLRT is *parameter-free*: it does not need the tuning of a window size (w in $\mathrm{M\text{-}UCB}$), or the knowledge of a lower bound on the magnitude of the smallest change (ε in CUSUM).

In this work we jointly investigate, both in theory and in practice, two possible combinations of the bandit algorithm with a change-point detector, namely the use of *local restarts* (resetting the history of an arm each time a change-point is detected on that arm) and *global restarts* (resetting the history of *all* arms once a change-point is detected on one of them). We provide a regret upper bound scaling in $\mathcal{O}(\sqrt{\Upsilon_T T \ln(T)})$ for both versions of GLR-klUCB, matching existing results (when Υ_T is known). Our numerical simulations reveal that using local restart leads to better empirical performance, and show that our approach often outperforms existing competitors.

The article is structured as follows. We introduce the model and review related works in Section 2. In Section 3, we study the Generalized Likelihood Ratio test (GLRT) as a Change-Point Detector (CPD) algorithm. We introduce the two variants of $\mathrm{GLR\text{-}klUCB}$ algorithm in Section 4, where we also present upper bound on the regret of each variant. The unified regret analysis for these two algorithms is sketched in Section 5. Numerical experiments are presented in Section 6, with more details in the Appendix.

2. The Piece-Wise Stationary Bandit Setup and Related Works

A *piece-wise stationary bandit model* is characterized by a set of K arms. A (random) stream of rewards $(X_{i,t})_{t\in\mathbb{N}^*}$ is associated to each arm $i \in \{1,\dots,K\}$. We assume that the rewards are bounded, and without loss of generality we assume that $X_{i,t} \in [0,1]$. We denote by $\mu_i(t) := \mathbb{E}[X_{i,t}]$ the mean reward of arm i at round t. At each round t, a decision maker has to select an arm $I_t \in \{1,\dots,K\}$, based on past observation and receives the corresponding reward $r(t) = X_{I_t,t}$. At time t, we denote by i_t^* an arm with maximal expected reward, *i.e.*, $\mu_{i_t^*}(t) = \max_i \mu_i(t)$, called an optimal arm (possibly not unique).

A policy π chooses the next arm to play based on the sequence of past plays and obtained rewards. The performance of π is measured by its (piece-wise stationary) *regret*, the difference between the expected reward obtained by an oracle policy playing an optimal arm i_t^* at time t, and that of the policy π:

$$R_T^\pi = \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_{i_t^*}(t) - \mu_{I_t}(t)\right)\right]. \tag{1}$$

In the piece-wise *i.i.d.* model, we furthermore assume that there is a (relatively small) number of *breakpoints*, denoted by $\Upsilon_T := \sum_{t=1}^{T-1} \mathbb{1}\left(\exists i \in \{1,\dots,K\} : \mu_t(i) \neq \mu_{t+1}(i)\right)$. We define the k-th breakpoint by $\tau^{(k)} = \inf\{t > \tau^{(k-1)} : \exists i : \mu_i(t) \neq \mu_i(t+1)\}$. Hence for $t \in [\tau^{(k)}+1, \tau^{(k+1)}]$, the rewards $(X_{i,t})$ associated to each arms are *i.i.d.*. Note than when a breakpoint occurs, we do not assume that all the arms means change, but that *there exists* an arm whose mean has changed. Depending on the application, many scenario can be meaningful: changes occurring on all arms simultaneously (due to some exogenous event), or only a few arms change at each breakpoint. Introducing the number of change-points on arm i, defined as $\mathrm{NC}_i := \sum_{t=1}^{T-1} \mathbb{1}\left(\mu_t(i) \neq \mu_{t+1}(i)\right)$, it clearly holds that $\mathrm{NC}_i \leq \Upsilon_T$, but there can be an arbitrary difference between these two quantities for some arms. Letting $C_T := \sum_{i=1}^{K} \mathrm{NC}_i$ be the total number of change-points on the arms, one can have $C_T \in \{\Upsilon_T, \dots, K\Upsilon_T\}$.

The piece-wise stationary bandit model can be viewed as an interpolation between stationary and adversarial models, as the stationary model corresponds to $\Upsilon_T = 0$, while the adversarial model can be considered as a special (worst) case, with $\Upsilon_T = T$. However, analyzing an algorithm for the piece-wise stationary bandit model requires to assume a small number of changes, typically $\Upsilon_T = o(\sqrt{T})$.

Related work. The piece-wise stationary bandit model was first studied by Kocsis and Szepesvári (2006); Yu and Mannor (2009); Garivier and Moulines (2011). It is also known as *switching* (Mellor and Shapiro, 2013) or *abruptly changing stationary* (Wei and Srivastava, 2018) environment. To our knowledge, all the previous approaches combine a standard bandit algorithm, like UCB, Thompson Sampling or EXP3, with a strategy to account for changes in the arms distributions. This strategy often consists in *forgetting old rewards*, to efficiently focus on the most recent ones, more likely to be similar to future rewards. We make the distinction between *passively* and *actively* adaptive strategies.

The first proposed mechanisms to forget the past consist in either discounting rewards (at each round, when getting a new reward on an arm, past rewards are multiplied by γ^n if that arm was not seen since $n > 0$ times, for a discount factor $\gamma \in (0,1)$), or using a sliding window (only the rewards gathered in the τ last observations of an arm are taken into account, for a window size τ). Those strategies are passively adaptive as the discount factor or the window size are fixed, and can be tuned as a function of T and Υ_T to achieve a certain regret bound. Discounted UCB (D-UCB) was proposed by Kocsis and Szepesvári (2006) and analyzed by Garivier and Moulines (2011), who prove a $\mathcal{O}(\sqrt{\Upsilon_T T}\ln(T))$ regret bound, if $\gamma = 1 - \sqrt{\Upsilon_T/T}/4$. The same authors proposed the Sliding-Window UCB (SW-UCB) and prove a $\mathcal{O}(\sqrt{\Upsilon_T T\ln(T)})$ regret bound, if $\tau = 2\sqrt{T\ln(T)/\Upsilon_T}$.

More recently, Raj and Kalyani (2017) proposed the Discounted Thompson Sampling (DTS) algorithm, which performs well in practice with $\gamma = 0.75$. However, no theoretical guarantees are given for this strategy, and our experiments did not really confirm the robustness to γ. The RExp3 algorithm (Besbes et al., 2014) can also be qualified as passively adaptive: it is based on (non-adaptive) restarts of the EXP3 algorithm. Note that this algorithm is introduced for a different setting, where the quantity of interest is not Υ_T but a quantity V_T called the total variational budget (satisfying $\Delta^{\text{change}}\Upsilon_T \leq V_T \leq \Upsilon_T$ with Δ^{change} the minimum magnitude of a change-point). A $\mathcal{O}(V_T^{1/3}T^{2/3})$ regret bound is proved, which is weaker than existing results in our setting. Hence we do no include this algorithm in our experiments.

The first *actively adaptive* strategy is Windowed-Mean Shift (Yu and Mannor, 2009), which combines any bandit policy with a change point detector which performs *adaptive restarts* of the bandit algorithm. However, this approach is not applicable to our setting as it takes into account side observations. Another line of research on actively adaptive algorithms uses a Bayesian point of view. A Bayesian Change-Point Detection (CPD) algorithm is combined with Thompson Sampling by Mellor and Shapiro (2013), and more recently in the Memory Bandit algorithm of Alami and Féraud (2017). Both algorithms do not have theoretical guarantees and their implementation is very costly, hence we do not include them in our experiments. Our closest competitors rather use frequentist CPD algorithms (see, e.g. Basseville et al. (1993)) combined with a bandit algorithm. The first algorithm of this flavor, Adapt-EVE algorithm (Hartland et al., 2006) uses a Page-Hinkley test and the UCB policy, but no theoretical guarantee are given. EXP3.R (Allesiardo and Féraud, 2015; Allesiardo et al., 2017) combines a CPD with EXP3, and the history of all arms are reset as soon as a a sub-optimal arm is detecting to become optimal and it achieve a $\mathcal{O}(\Upsilon_T\sqrt{T\ln(T)})$ regret. This is weaker than the $\mathcal{O}(\sqrt{\Upsilon_T T\ln(T)})$ regret achieved by two recent algorithms, CUSUM-UCB (Liu et al., 2018) and Monitored UCB (M-UCB, Cao et al. (2019)).

CUSUM-UCB is based on a rather complicated two-sided CUSUM test, that uses the first M samples from one arm to compute an initial average, and then detects whether a drift of size larger than ε occurred from this value by checking whether a random walk based on the remaining observations crosses a threshold h. Thus it requires the tuning of three parameters, M, ε and h. CUSUM-UCB performs *local restarts* using this test, to reset the history of *one arm* for which the test detects a change. M-UCB uses a much simpler test, based on the w most recent observations from an arm: a change is detected if the absolute difference between the empirical means of the first and second halves of those w observations exceeds a threshold h. So it requires the tuning of two parameters, w and h. M-UCB performs *global restarts* using this test, to reset the history of *all arms* whenever the test detects a change on one of the arms. Compared to CUSUM-UCB, note that M-UCB is numerically much simpler as it only uses a bounded memory, of order $\mathcal{O}(Kw)$ for K arms.

Advantages of our approach. CUSUM-UCB and M-UCB are both analyzed under some reasonable assumptions on the problem parameters –the means $(\mu_i(t))$– mostly saying that the breakpoints are sufficiently far away from each other. However, the proposed guarantees only hold for parameters *tuned using some prior knowledge of the means*. Indeed, while in both cases the threshold h can be set as a function of the horizon T and the number of breakpoints Υ_T (also needed by previous approaches to obtain the best possible bounds), the parameter ε for CUSUM and w for M-UCB require the knowledge of Δ^{change} the smallest magnitude of a change-point. In this paper, we propose *the first algorithm that does not require this knowledge*, and still attains a $\mathcal{O}(\sqrt{\Upsilon_T T\ln(T)})$ regret. Moreover we propose the first comparison of the use of local and global restarts within an adaptive algorithm, by studying two variants of our algorithm. This study is supported by both theoretical and empirical results. Finally, on the practical side, while we can note that the proposed GLR test is more complex to implement than the test used by M-UCB, we propose two heuristics to speed it up while not losing much in terms of regret.

3. The Bernoulli GLR Change Point Detector

Sequential change-point detection has been extensively studied in the statistical community (see, *e.g.*, Basseville et al. (1993) for a survey). In this article, we are interested in detecting changes on the mean of a probability distribution with bounded support. Assuming that we collect independent samples $X_1, X_2, \dots$ all from some distribution supported in $[0, 1]$. We want to discriminate between two possible scenarios: all the samples come from distributions that have a common mean μ_0, or there exists a *change-point* $\tau \in \mathbb{N}^*$ such that $X_1, \dots, X_\tau$ have some mean μ_0 and $X_{\tau+1}, X_{\tau+2}, \dots$ have a different mean $\mu_1 \neq \mu_0$. A sequential change-point detector is a stopping time $\hat{\tau}$ with respect to the filtration $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ such that $(\hat{\tau} < \infty)$ means that we reject the hypothesis $\mathcal{H}_0 : (\exists \mu_0 \in [0,1] : \forall i \in \mathbb{N}, \mathbb{E}[X_i] = \mu_0)$.

Generalized Likelihood Ratio tests date back to the seminal work of Barnard (1959) and were for instance studied for change-point detection by Siegmund and Venkatraman (1995). Exploiting the fact that bounded distribution are $(1/4)$-sub Gaussian (*i.e.*, their moment generating function is dominated by that of a Gaussian distribution with the same mean and a variance $1/4$), the (Gaussian) GLRT, recently studied in depth by Maillard (2019), can be use for our problem. We propose instead to exploit the fact that bounded distributions are also dominated by Bernoulli distributions. We call a *sub-Bernoulli distribution* any distribution ν that satisfies $\ln \mathbb{E}_{X\sim\nu}\left[e^{\lambda X}\right] \leq \phi_\mu(\lambda)$ with $\mu = \mathbb{E}_{X\sim\nu}[X]$ and $\phi_\mu(\lambda) = \ln(1-\mu+\mu e^\lambda)$ is the log moment generating function of a Bernoulli distribution with mean μ. Lemma 1 of Cappé et al. (2013) establishes that any bounded distribution supported in $[0, 1]$ is a sub-Bernoulli distribution.

3.1. Presentation of the test

If the samples (X_t) were all drawn from a Bernoulli distribution, our change-point detection problem would reduce to a parametric sequential test of $\mathcal{H}_0 : (\exists \mu_0 : \forall i \in \mathbb{N}, X_i \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_0))$ against the alternative $\mathcal{H}_1 : (\exists \mu_0 \neq \mu_1, \tau \in \mathbb{N}^* : \; X_1, \dots, X_\tau \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_0)$ and $X_{\tau+1}, X_{\tau+2}, \dots \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_1))$. The Generalized Likelihood Ratio statistic for this test is defined by

$$\mathrm{GLR}(n) := \frac{\sup\limits_{\mu_0,\mu_1,\tau<t} \ell(X_1, \dots, X_n; \mu_0, \mu_1, \tau)}{\sup\limits_{\mu_0} \ell(X_1, \dots, X_t; \mu_0)},$$

where $\ell(X_1, \dots, X_t; \mu_0)$ and $\ell(X_1, \dots, X_t; \mu_0, \mu_1, \tau)$ denote the likelihoods of the first n observations under a model in $\mathcal{H}_0$ and $\mathcal{H}_1$. High values of this statistic tend to indicate rejection of $\mathcal{H}_0$. Using the form of the likelihood for Bernoulli distribution, this statistic can be written with the binary relative entropy kl,

$$\mathrm{kl}(x, y) := x \ln\left(\frac{x}{y}\right) + (1-x)\ln\left(\frac{1-x}{1-y}\right). \tag{2}$$

Indeed, one can show that $\mathrm{GLR}(t) = \sup_{s\in[1,n]} \left[s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:t}\right)\right]$, where for $k \leq k'$, $\hat{\mu}_{k:k'}$ denotes the average of the observations collected between the instants k and k'. This motivates the definition of the Bernoulli GLR change point detector.

Definition 1 *The Bernoulli GLR change point detector with threshold function $\beta(t, \delta)$ is*

$$\hat{\tau}_\delta := \inf\left\{n \in \mathbb{N}^* : \sup_{s\in[1,t]} \left[s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right)\right] \geq \beta(n, \delta)\right\}. \tag{3}$$

Asymptotic properties of the GLR for change-point detection have been studied by Lai and Xing (2010) for Bernoulli distributions and more generally for one-parameter exponential families, for which the GLR test is defined as in (3) but with $\mathrm{kl}(x, y)$ replaced by the Kullback-Leibler divergence between two elements in that exponential family that have mean x and y. For example, the Gaussian GLR studied by Maillard (2019) corresponds to (3) with $\mathrm{kl}(x, y) = 2(x - y)^2$ when the variance is set to $\sigma^2 = 1/4$, and non-asymptotic properties of this test are given for any $(1/4)$-subGaussian samples.

In the next section, we provide new non-asymptotic results about the Bernoulli GLR test under the assumption that the samples (X_t) come from a sub-Bernoulli distribution, which holds for any distribution supported in $[0, 1]$. Note that Pinsker's inequality gives that $\mathrm{kl}(x, y) \geq 2(x - y)^2$, hence the Bernoulli GLR may stop earlier that the Gaussian GLR based on the quadratic divergence $2(x - y)^2$.

3.2. Properties of the Bernoulli GLR

In Lemma 2 below, we propose a choice of the threshold function $\beta(n, \delta)$ under which the probability that there exists a *false alarm* under *i.i.d.* data is small. To define β, we need to introduce the function $\mathcal{T}$,

$$\mathcal{T}(x) \;:=\; 2\tilde{h}\left(\frac{h^{-1}(1+x) + \ln(2\zeta(2))}{2}\right) \tag{4}$$

where for $u \geq 1$ we define $h(u) = u - \ln(u)$ and its inverse $h^{-1}(u)$. And for any $x \geq 0$, $\tilde{h}(x) = e^{1/h^{-1}(x)} h^{-1}(x)$ if $x \geq h^{-1}(1/\ln(3/2))$ and $\tilde{h}(x) = (3/2)(x - \ln(\ln(3/2)))$ otherwise. The function $\mathcal{T}$ is easy to compute numerically. Its use for the construction of concentration inequalities that are uniform in time is detailed in Kaufmann and Koolen (2018), where tight upper bound on the function $\mathcal{T}$ are also given: $\mathcal{T}(x) \simeq x + 4\ln\left(1 + x + \sqrt{2x}\right)$ for $x \geq 5$ and $\mathcal{T}(x) \sim x$ when x is large. The proof of Lemma 2, that actually holds for any sub-Bernoulli distribution, is given in Appendix C.1.

Lemma 2 *Assume that there exists $\mu_0 \in [0, 1]$ such that $\mathbb{E}[X_t] = \mu_0$ and that $X_i \in [0, 1]$ for all i. Then the Bernoulli GLR test satisfies $\mathbb{P}_{\mu_0}(\hat{\tau}_\delta < \infty) \leq \delta$ with the threshold function*

$$\beta(n, \delta) = 2\mathcal{T}\left(\frac{\ln(3n\sqrt{n}/\delta)}{2}\right) + 6\ln(1 + \ln(n)). \tag{5}$$

Another key feature of a change-point detector is its *detection delay* under a model in which a change from μ_0 to μ_1 occurs at time τ. We already observed that from Pinsker's inequality, the Bernoulli GLR stops earlier than a Gaussian GLR. Hence, one can leverage some techniques from Maillard (2019) to upper bound the detection delay of the Bernoulli GLR. Letting $\Delta = |\mu_0 - \mu_1|$, one can essentially establish that for τ larger than $(1/\Delta^2)\ln(1/\delta)$ (*i.e.*, enough samples before the change), the delay can be of the same magnitude (*i.e.*, enough samples after the change). In our bandit analysis to follow, the detection delay will be crucially used to control the probability of the good event (in Lemma 12 and 13).

3.3. Practical considerations

Lemma 2 provides the first control of false alarm for the Bernoulli GLR employed for bounded data. However, the threshold (5) is not fully explicit as the function $\mathcal{T}(x)$ can only be computed numerically. Note that for sub-Gaussian distributions, results from Maillard (2019) show that the smaller and more explicit threshold $\beta(n, \delta) = \left(1 + \frac{1}{n}\right)\ln\left(\frac{3n\sqrt{n}}{\delta}\right)$, can be used to prove an upper bound of δ for the false alarm probability of the GLR, with quadratic divergence $\mathrm{kl}(x, y) = 2(x - y)^2$. For the Bernoulli GLR, numerical simulations suggest that the threshold (5) is a bit conservative, and in practice we recommend to keep only the leading term and use $\beta(n, \delta) = \ln(3n\sqrt{n}/\delta)$.

Also note that, as any test based on scan-statistics, the GLR can be costly to implement as at every time step, it considers all previous time steps as a possible positions for a change-point. Thus, in practice the following adaptation may be interesting, based on down-sampling the possible time steps:

$$\tilde{\tau}_\delta = \inf \left\{ n \in \mathcal{N} : \sup_{s \in \mathcal{S}_n} \left[s \times d\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times d\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right) \right] \geq \beta(n, \delta) \right\}, \tag{6}$$

for subsets $\mathcal{N}$ and $\mathcal{S}_n$. Following the proof of Lemma 2, we can easily see that this variant enjoys the exact same false-alarm control. However, the detection delay may be slightly increased. In Appendix F.1 we show that using these practical speedups has little impact on the regret.

4. The GLR-klUCB Algorithm

Our proposed algorithm, GLR-klUCB, combines a bandit algorithm with a change-point detector running on each arm. It also needs a third ingredient, some forced exploration parameterized by $\alpha \in (0, 1)$ to ensure each arm is sampled enough and changes can also be detected on arms currently under-sampled by the bandit algorithm. GLR-klUCB combines the klUCB algorithm (Cappé et al., 2013), known to be optimal for Bernoulli bandits, with the Bernoulli GLR change-point detector introduced in Section 3. This algorithm, formally stated as Algorithm 1, can be used in any bandit model with bounded rewards, and is expected to be very efficient for Bernoulli distributions, which are relevant for practical applications.

Algorithm 1 GLR-klUCB, with **Local** or **Global** restarts

Require: *Problem parameters*: $T \in \mathbb{N}^*$, $K \in \mathbb{N}^*$.
Require: *Algorithm parameters*: exploration probability $\alpha \in (0, 1)$, confidence level $\delta > 0$.
Require: *Option*: **Local** or **Global** restart.
1: **Initialization:** $\forall i \in \{1, \dots, K\}$, $\tau_i \leftarrow 0$ and $n_i \leftarrow 0$
2: **for all** $t = 1, 2, \dots, T$ **do**
3:   **if** $\alpha > 0$ and $t \mod \lfloor \frac{K}{\alpha} \rfloor \in \{1, \dots, K\}$ **then**
4:     $A_t \leftarrow t \mod \lfloor \frac{K}{\alpha} \rfloor$. *(forced exploration)*
5:   **else**
6:     $A_t \leftarrow \arg\max_{i \in \{1,\dots,K\}} \mathrm{UCB}_i(t)$ as defined in (7)
7:   **end if**
8:   Play arm A_t and receive the reward $X_{A_t,t}$: $n_{A_t} \leftarrow n_{A_t} + 1$; $Z_{A_t, n_{A_t}} \leftarrow X_{A_t,t}$.
9:   **if** $\mathrm{GLR}_\delta(Z_{A_t,1}, \dots, Z_{A_t,n_{A_t}})$ = True **then**
10:     **if** Global restart **then**
11:       $\forall i \in \{1, \dots, K\}, \tau_i \leftarrow t$ and $n_i \leftarrow 0$. (*global restart*)
12:     **else**
13:       $\tau_{A_t} \leftarrow t$ and $n_{A_t} \leftarrow 0$. (*local restart*)
14:     **end if**
15:   **end if**
16: **end for**

The GLR-klUCB algorithm can be viewed as a klUCB algorithm allowing for some *restarts* on the different arms. A restart happens when the Bernoulli GLR change-point detector detects a change on the arm that has been played (line 9). To be fully specific, $\mathrm{GLR}_\delta(Z_1, \dots, Z_n) = \mathrm{True}$ if and only if

$$\sup_{1<s<n} \left[s \times \mathrm{kl}\left(\frac{1}{s}\sum_{i=1}^{s} Z_i, \frac{1}{n}\sum_{i=1}^{n} Z_i \right) + (n-s) \times \mathrm{kl}\left(\frac{1}{n-s}\sum_{i=s+1}^{t} Z_i, \frac{1}{n}\sum_{i=1}^{n} Z_i \right) \right] \geq \beta(n, \delta),$$

with $\beta(n, \delta)$ defined in (5), or $\beta(n, \delta) = \ln(3n^{3/2}/\delta)$, as recommended in practice, see Section 3.3. We define the (klUCB-like) index used by our algorithm, by denoting $\tau_i(t)$ the last restart that happened for arm i before time t, $n_i(t) = \sum_{s=\tau_i(t)+1}^{t} \mathbb{1}(A_s = i)$ the number of selections of arm i, and $\hat{\mu}_i(t) = (1/n_i(t)) \sum_{s=\tau_i(t)+1}^{t} X_{i,s}\mathbb{1}(A_s = i)$ their empirical mean (if $n_i(t) \neq 0$). With the exploration function $f(t) = \ln(t) + 3\ln(\ln(t))$ (if $t > 1$ else $f(t) = 0$), the index is defined as

$$\mathrm{UCB}_i(t) := \max\big\{q \in [0,1] : n_i(t) \times \mathrm{kl}\left(\hat{\mu}_i(t), q\right) \leq f(t - \tau_i(t))\big\}. \tag{7}$$

In this work, we simultaneously investigate two possible behavior: *global restart* (reset the history of all arms once a change was detected on one of them, line 11), and *local restart* (reset only the history of the arm on which a change was detected, line 13), which are the two different options in Algorithm 1. Under local restart, in the general case the times $\tau_i(t)$ are not equal for all arms, hence the index policy associated to (7) is *not* a standard UCB algorithm, as each index uses a *different exploration rate*. One can highlight that in the CUSUM-UCB algorithm, which is the only existing algorithm based on local restart, the UCB index are defined differently[1]: $f(t - \tau_i(t))$ is replaced by $f(n_t)$ with $n_t = \sum_{i=1}^{K} n_i(t)$.

The forced exploration scheme used in GLR-klUCB (lines 3-5) generalizes the deterministic exploration scheme proposed for M-UCB by (Cao et al., 2019), whereas CUSUM-UCB performs randomized exploration. A consequence of this forced exploration is given in Proposition 3 (proved in Appendix B).

Proposition 3 *For every pair of instants $s \leq t \in \mathbb{N}^*$ between two restarts on arm i (*i.e.*, for a $k \in \{1, \dots, NC_i\}$, one has $\tau_i^{(k)} = \tau_i(t) < s \leq t < \tau_i^{(k+1)}$) it holds that $n_i(t) - n_i(s) \geq \lfloor \frac{\alpha}{K}(t - s) \rfloor$.*

4.1. Results for GLR-klUCB using Global Changes

Recall that $\tau^{(k)}$ denotes the position of the k-th break-point and let $\mu_i^{(k)}$ be the mean of arm i on the segment between the k and $(k+1)$-th breakpoint: $\forall t \in \{\tau^{(k-1)} + 1, \dots, \tau^{(k)}\}, \mu_i(t) = \mu_i^{(k)}$. We also introduce $k^* = \arg\max_i \mu_i^{(k)}$ and the largest gap at break-point k as $\Delta^{(k)} := \max_{i=1,\dots,K} |\mu_i^{(k)} - \mu_i^{(k-1)}| > 0$.

Assumption 4 below is easy to interpret and standard in non-stationary bandits. It requires that the distance between two consecutive breakpoints is large enough: how large depends on the magnitude of the largest change that happen at those two breakpoints. Under this assumption, we provide in Theorem 5 a finite time problem-dependent regret upper bound. It features the parameters α and δ, the KL-divergence terms $\mathrm{kl}(\mu_i^{(k)}, \mu_{k^*}^{(k)})$ expressing the hardness of the (stationary) MAB problem between two breakpoints, and the $\Delta^{(k)}$ terms expressing the hardness of the change-point detection problem.

Assumption 4 *Define $d^{(k)} = d^{(k)}(\alpha, \delta) = \lceil \frac{4K}{\alpha\left(\Delta^{(k)}\right)^2}\beta(T, \delta) + \frac{K}{\alpha} \rceil$. Then we assume that for all $k \in \{1, \dots, \Upsilon_T\}$, $\tau^{(k)} - \tau^{(k-1)} \geq 2\max(d^{(k)}, d^{(k-1)})$.*

Theorem 5 *For α and δ for which Assumption 4 is satisfied, the regret of* GLR-klUCB *with parameters α and δ based on **Global** Restart satisfies*

$$R_T \leq 2\sum_{k=1}^{\Upsilon_T} \frac{4K}{\alpha\left(\Delta^{(k)}\right)^2}\beta(T, \delta) + \alpha T + \delta(K+1)\Upsilon_T + \sum_{k=1}^{\Upsilon_T} \sum_{i : \mu_i^{(k)} \neq \mu_{k^*}^{(k)}} \frac{\left(\mu_{k^*}^{(k)} - \mu_i^{(k)}\right)}{\mathrm{kl}\left(\mu_i^{(k)}, \mu_{k^*}^{(k)}\right)} \ln(T) + \mathcal{O}\Big(\sqrt{\ln(T)}\Big).$$

1. This alternative choice is currently not fully supported by theory, as we found mistakes in the analysis of CUSUM-UCB: Hoeffding's inequality is used with a *random* number of observations and a *random* threshold to obtain Eq. (31)-(32).

Corollary 6 *For "easy" problems satisfying the corresponding Assumption 4, let Δ^{opt} denote the smallest value of a sub-optimality gap on one of the stationary segments, and Δ^{change} be the smallest magnitude of any change point on any arm.*

1. *Choosing* $\alpha = \sqrt{\frac{\ln(T)}{T}}$, $\delta = \frac{1}{\sqrt{T}}$ *gives* $R_T = \mathcal{O}\Big(\frac{K}{(\Delta^{change})^2}\Upsilon_T\sqrt{T\ln(T)} + \frac{(K-1)}{\Delta^{opt}}\Upsilon_T\ln(T)\Big)$,
2. *Choosing* $\alpha = \sqrt{\frac{\Upsilon_T\ln(T)}{T}}$, $\delta = \frac{1}{\sqrt{\Upsilon_T T}}$ *gives* $R_T = \mathcal{O}\Big(\frac{K}{(\Delta^{change})^2}\sqrt{\Upsilon_T T\ln(T)} + \frac{(K-1)}{\Delta^{opt}}\Upsilon_T\ln(T)\Big)$.

4.2. Results for GLR-klUCB using Local Changes

A few new notation are needed to state a regret bound for GLR-klUCB using local changes. We let $\tau_i^{(\ell)}$ denote the position of the ℓ-th change point *for arm* i: $\tau_i^{(\ell)} = \inf\{t > \tau_i^{(\ell-1)} : \mu_i(t) \neq \mu_i(t+1)\}$, with the convention $\tau_i^{(0)} = 0$, and let $\overline{\mu}_i^{(\ell)}$ be the ℓ-th value for the mean of arm i, such that $\forall t \in [\tau_i^{(\ell-1)}+1, \tau_i^{(\ell)}]$, $\mu_i(t) = \overline{\mu}_i^{(\ell)}$. We also introduce the gap $\Delta_i^{(\ell)} = \overline{\mu}_i^{\ell} - \overline{\mu}_i^{\ell-1} > 0$.

Assumption 7 requires that any two consecutive change-points *on a given arm* are sufficiently spaced (relatively to the magnitude of those two change-points). Under that assumption, Theorem 8 provides a regret upper bound that scales with similar quantities as that of Theorem 5, except that the number of breakpoints Υ_T is replaced with the *total* number of change points $\mathrm{C}_T = \sum_{i=1}^{K} \mathrm{NC}_i \leq K\Upsilon_T$.

Assumption 7 *Define* $d_i^{(\ell)} = d_i^{(\ell)}(\alpha,\delta) = \lceil \frac{4K}{\alpha\left(\Delta_i^{(\ell)}\right)^2}\beta(T,\delta) + \frac{K}{\alpha}\rceil$. *We assume that for all arm i and all* $\ell \in \{1,\dots,NC_i\}$, $\tau_i^{(\ell)} - \tau_i^{(\ell-1)} \geq 2\max(d_i^{(\ell)}, d_i^{(\ell-1)})$.

Theorem 8 *For α and δ for which Assumption 7 is satisfied, the regret of* GLR-klUCB *with parameters α and δ based on Local Restart satisfies*

$$R_T \leq 2\sum_{i=1}^{K}\sum_{\ell=1}^{NC_i}\frac{4K}{\alpha\left(\Delta_i^{(\ell)}\right)^2}\beta(T,\delta) + \alpha T + 2\delta\mathrm{C}_T + \sum_{i=1}^{K}\sum_{\ell=1}^{NC_i}\frac{\ln(T)}{\mathrm{kl}(\overline{\mu}_i^{(\ell)}, \mu_{i,\ell}^*)} + \mathcal{O}\Big(\sqrt{\ln(T)}\Big),$$

where $\mu_{i,\ell}^* = \inf\left\{\mu_{i_t^*}(t) : \mu_{i_t^*}(t) \neq \overline{\mu}_i^{(\ell)}, t \in [\tau_i^{(\ell)}+1, \tau_i^{(\ell+1)}]\right\}$.

Corollary 9 *For "easy" problems satisfying the corresponding Assumption 7, with Δ^{opt} and Δ^{change} defined as in Corollary 6, the following holds.*

1. *Choosing* $\alpha = \sqrt{\frac{\ln(T)}{T}}$, $\delta = \frac{1}{\sqrt{T}}$ *gives* $R_T = \mathcal{O}\Big(\frac{K}{(\Delta^{change})^2}\mathrm{C}_T\sqrt{T\ln(T)} + \frac{1}{(\Delta^{opt})^2}\mathrm{C}_T\ln(T)\Big)$,
2. *Choosing* $\alpha = \sqrt{\frac{K\Upsilon_T\ln(T)}{T}}$, $\delta = \frac{1}{\sqrt{K\Upsilon_T T}}$ *gives* $R_T = \mathcal{O}\Big(\frac{K}{(\Delta^{change})^2}\sqrt{\mathrm{C}_T T\ln(T)} + \frac{1}{(\Delta^{opt})^2}\mathrm{C}_T\ln(T)\Big)$.

4.3. Interpretation

Theorems 5 and 8 both show that there exists a tuning of α and δ as a function and T *and the number of changes* such that the regret is of order $\mathcal{O}_h(K\sqrt{\Upsilon_T T\ln(T)})$ and $\mathcal{O}_h(K\sqrt{\mathrm{C}_T T\ln(T)})$ respectively, where the $\mathcal{O}_h$ notations ignore the gap terms. For very particular instances such that $\Upsilon_T = \mathrm{C}_T$, *i.e.*, at each break-point only one arm changes (*e.g.*, problem 1 in Section 6), the theory advocates the use of local changes. Indeed, while the regret guarantees obtained are similar, those obtained for local changes hold for a wider variety of problems as Assumption 7 is less stringent than 4. Besides those specific instances, our results are essentially worse for local than global changes. However, we only obtain regret upper bounds – thus providing a theoretical safety net for both variants of our algorithm, and the practical story is different, as discussed in Section 6. We find that GLR-klUCB performs better with local restarts.

One can note that with the tuning of α and δ prescribed by Corollaries 6 and 9, our regret bounds hold for problem instances for which two consecutive breakpoints (or change-points on an arm) are separated by more than $\sqrt{T\ln(T)}/(\Delta^{\text{change}})^2$ time steps. Hence those guarantees are valid on "easy" problem instances only, with few changes of a large magnitude (*e.g.*, not for problems 3 or 5). However, this does not prevent our algorithms from performing well on more realistic instances, and numerical experiments support this claim. Note that M-UCB (Cao et al., 2019) is also analyzed for the same type of unrealistic assumptions, while its practical performance is illustrated beyond those.

5. A Unified Regret Analysis

In this section, we sketch a unified proof for Theorem 5 and 8, whose detailed proofs can be found in Appendix D and E respectively. We emphasize that our approach is significantly different from those proposed by Cao et al. (2019) for M-UCB and by Liu et al. (2018) for CUSUM-UCB.

Recall that the regret is defined as $R_T = \mathbb{E}\left[\sum_{t=1}^T (\mu_{i_t^*}(t) - \mu_{I_t}(t))\right]$. Introducing $\mathcal{D}(T,\alpha)$ the (deterministic) set of time steps at which the forced exploration is performed before time T (see lines 3-4 in Algorithm 1), one can write, using notably that $\mu_{i_t^*} - \mu_{I_t} \leq 1$, due to the bounded rewards,

$$\begin{aligned}\sum_{t=1}^T (\mu_{i_t^*}(t) - \mu_{I_t}(t)) &\leq \sum_{t=1}^T \mathbb{1}(t \in \mathcal{D}(T,\alpha)) + \sum_{t=1}^T (\mu_{i_t^*}(t) - \mu_{I_t}(t))\mathbb{1}\left(t \notin \mathcal{D}(T,\alpha), \mathrm{UCB}_{I_t}(t) \geq \mathrm{UCB}_{i_t^*}(t)\right) \\ &\leq \alpha T + \sum_{t=1}^T \mathbb{1}\left(\mathrm{UCB}_{i_t^*}(t) \leq \mu_{i_t^*}(t)\right) + \sum_{i=1}^K \sum_{t=1}^T (\mu_{i_t^*}(t) - \mu_i(t))\mathbb{1}\left(I_t = i, \mathrm{UCB}_i(t) \geq \mu_{i_t^*}(t)\right).\end{aligned}$$

Introducing some *good event* $\mathcal{E}_T$ to be specified in each case, one can write

$$R_T \leq T\mathbb{P}(\mathcal{E}_T^c) + \alpha T + \underbrace{\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^T \mathbb{1}\left(\mathrm{UCB}_{i_t^*}(t) \leq \mu_{i_t^*}(t)\right)\right]}_{(A)} + \underbrace{\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^T (\mu_{i_t^*}(t) - \mu_{I_t}(t))\mathbb{1}\left(\mathrm{UCB}_{I_t}(t) \geq \mu_{i_t^*}(t)\right)\right]}_{(B)}.$$

Each analysis requires to define an *appropriate good event*, stating that *some* change-points are detected within a reasonable delay. Each regret bound then follows from upper bounds on term (A), term (B), and on the failure probability $\mathbb{P}(\mathcal{E}_T^c)$. To control (A) and (B), we split the sum over consecutive segments $[\tau^{(k)}+1, \tau^{(k+1)}]$ for global changes and $[\tau_i^{(k)}+1, \tau_i^{(k+1)}]$ for each arm i for local changes, and use elements from the analysis of klUCB of Cappé et al. (2013).

The tricky part of each proof, which crucially exploits Assumption 4 or 7, is actually to obtain an upper bound on $\mathbb{P}(\mathcal{E}_T^c)$. For example for local changes (Theorem 8), the good event is defined as

$$\mathcal{E}_T(\alpha,\delta) = \left(\forall i \in \{1,\dots,K\}, \forall \ell \in \{1,\dots,\mathrm{NC}_i\}, \hat{\tau}_i^{(\ell)} \in \left[\tau_i^{(\ell)}+1, \tau_i^{(\ell)} + d_i^{(\ell)}\right]\right),$$

where $\hat{\tau}_i^{(\ell)}$ is defined as the ℓ-th change detected by the algorithm on arm i and $d_i^{(\ell)} = d_i^{(\ell)}(\alpha,\delta)$ is defined in Assumption 7. Introducing the event $\mathcal{C}_i^{(\ell)} = \left\{\forall j \leq \ell, \hat{\tau}_i^{(\ell)} \in \left[\tau_i^{(j)}+1, \tau_i^{(j)} + d_i^{(j)}\right]\right\}$ that all the changes up to the ℓ-th have been detected, a union bound yields the following decomposition:

$$\mathbb{P}(\mathcal{E}_T(\alpha,\delta)^c) \leq \sum_{i=1}^{K}\sum_{\ell=1}^{\mathrm{NC}_i} \underbrace{\mathbb{P}\left(\hat{\tau}_i^{(\ell)} \leq \tau_i^{(\ell)} \mid \mathcal{C}_i^{(\ell-1)}\right)}_{(a)} + \sum_{i=1}^{K}\sum_{\ell=1}^{\mathrm{NC}_i} \underbrace{\mathbb{P}\left(\hat{\tau}_i^{(\ell)} \geq \tau_i^{(\ell)} + d_i^{(\ell)} \mid \mathcal{C}_i^{(\ell-1)}\right)}_{(b)}.$$

Term (a) is related to the control of probability of false alarm, which is given by Lemma 2 for a change-point detector run in isolation. Observe that under the bandit algorithm, the change point detector associated to arm i is based on (possibly much) less than $t - \tau_i(t)$ samples from arm i, which makes false alarm even less likely to occur. Hence, it is easy to show that $(a) \leq \delta$.

Term (b) is related to the control of the detection delay, which is more tricky to obtain under the GLR-klUCB adaptive sampling scheme, when compared to a result like Theorem 6 in Maillard (2019) for the change-point detector run in isolation. More precisely, we need to leverage the forced exploration (Proposition 3) to be sure we have enough samples for detection. This explains why delays defined in Assumption 7 are scaled by $1/\alpha$. Using some elementary calculus and a concentration inequality given in Lemma 11, we can finally prove that $(b) \leq \delta$. Finally, the "bad event" is unlikely: $\mathbb{P}(\mathcal{E}_T^c) \leq 2\mathrm{C}_T\delta$.

6. Experimental Results

In this section we report results of numerical simulations performed on synthetic data to compare the performance of GLR-klUCB against other state-of-the-art approaches, on some piece-wise stationary bandit problems. For simplicity, we restrict to rewards generated from Bernoulli distributions, even if GLR-klUCB can be applied to any bounded distributions.

Algorithms and parameters tuning. We include in our study two algorithms designed for the classical MAB, klUCB (Garivier and Cappé, 2011) and Thompson sampling (Agrawal and Goyal, 2012; Kaufmann et al., 2012), as well as an "oracle" version of klUCB, that we call Oracle-Restart. This algorithm knows the exact locations of the breakpoints, and restarts klUCB at those locations (without any delay).

Then, we compare our algorithms to several competitor designed for a piece-wise stationary model. For a fair comparison, all algorithms that use UCB as a sub-routine were adapted to use klUCB instead, which yields better performance[2]. Moreover, all the algorithms are tuned as described in the corresponding paper, using in particular the knowledge of the number of breakpoints Υ_T and the horizon T. We first include three *passively adaptive algorithms*: Discounted klUCB (D-klUCB, Kocsis and Szepesvári (2006)), with discount factor $\gamma = 1 - \sqrt{\Upsilon_T/T}/4$; Sliding-Window klUCB (SW-klUCB, Garivier and Moulines (2011)) using window-size $\tau = 2\sqrt{T\ln(T)/\Upsilon_T}$ and Discounted Thompson sampling (DTS, Raj and Kalyani (2017)) with discount factor $\gamma = 0.95$. For this last algorithm, the discount factor $\gamma = 0.75$ suggested by the authors was performing significantly worse on our problem instances.

Our main goal is to compare against *actively adaptive algorithms*. We include CUSUM-klUCB (Liu et al., 2018), tuned with $M = 150$ and $\varepsilon = 0.1$ for easy problems $(1, 2, 4)$ and $\varepsilon = 0.001$ for hard problems $(3, 5)$, and with $h = \ln(T/\Upsilon_T)$, $\alpha = \sqrt{\Upsilon_T \ln(T/\Upsilon_T)/T}$, as suggested in the paper. Finally, we include M-klUCB (Cao et al., 2019), tuned with $w = 150$, based on a prior knowledge of the problems as the formula using $\delta_{\min}$ given in the paper is too large for small horizons (on all our problem instances), a threshold $b = \sqrt{w\ln(2KT)}$ and $\gamma = \sqrt{\Upsilon_T K(2b + 3\sqrt{w})/(2T)}$ as suggested by Remark 4 in the paper.

For GLR-klUCB, we explore the two different options with **Local** and **Global** restarts, using respectively $\delta = 1/\sqrt{\Upsilon T}$, $\alpha = \alpha_0\sqrt{\Upsilon_T \ln(T)/T}$ and $\delta = 1/\sqrt{K\Upsilon_T T}$, $\alpha = \alpha_0\sqrt{K\Upsilon_T\ln(T)/T}$ from

2. Liu et al. (2018); Cao et al. (2019) both mention that extending their analysis to the use of klUCB should not be too difficult.

Corollaries 6 and 9. The constant is set to $\alpha_0 = 0.05$ (we show in Appendix G a certain robustness, with similar regret as soon as $\alpha \leq 0.1$). To speed up our simulations, two optimizations are used, with $\Delta n = \Delta s = 10$, and CUSUM also uses the first trick with $\Delta n = 10$ (see Appendix F.1 for more details).

Results. We report results obtained on three different piece-wise stationary bandit problems, illustrated in Figure 1 in Appendix A, and described in more details below. Additional results on two more problems can be found in Appendix H, where additional regret curves can be found for all problems. For each experiment, the regret was estimated using 1000 independent runs.

Problem 1. There are $K = 3$ arms changing $\Upsilon_T = 4$ times until $T = 5000$. The arm means $(\mu_i(t))_{1\leq i\leq K, 1\leq t\leq T}$ are shown in Fig. 1a. Note that changes happen on only one arm (*i.e.*, $C_T = \Upsilon_T = 4$), and the optimal arm changes once at $t = \tau_2^{(1)} = 2000$, with a large gap $\Delta = 0.6$.

Problem 2. (see Fig. 1b). This problem is close to Problem 1, with a minimum optimality gap of 0.1. However, all arms change at every breakpoint (*i.e.*, $C_T = K\Upsilon_T = 12$), with identical gap $\Delta = 0.1$. The first optimal arm decreases at every change (2 with ∇), and one arm stays the worst (1 with $\diamond$).

Problem 3. (see Fig. 1c) This problem is harder, with $K = 6$, $\Upsilon_T = 8$ and $T = 20000$. Most arms change at almost every time steps, and means are bounded in $[0.01, 0.07]$. The gaps Δ are much smaller than for the first problems, with amplitudes ranging from 0.02 to 0.001. Note that the assumptions of our regret upper bounds are violated, as well as the assumptions for the analysis of M-UCB and CUSUM-UCB. This problem is inspired from Figure 3 of Cao et al. (2019), where the synthetic data was obtained from manipulations on a real-world database of clicks from *Yahoo!*.

Table 1 shows the final regret R_T obtained for each algorithm. Results highlighted in **bold** show the best non-oracle algorithm for each experiments, with our proposal being the best non-oracle strategy for problems 1 and 2. Thompson sampling and klUCB are efficient, and better than Discounted-klUCB which is very inefficient. DTS and SW-klUCB can sometimes be more efficient than their stationary counterparts, but perform worse than the Oracle and most actively adaptive algorithms. M-klUCB and CUSUM-klUCB outperform the previous algorithms, but GLR-klUCB is often better. On these problems, our proposal with **Local** restarts is always more efficient than with **Global** restarts. Note that on problem 2, all means change at every breakpoint, hence one could expect global restart to be more efficient, yet our experiments show the superiority of local restarts on every instances.

Algorithms \Problems	Pb 1	Pb 2	Pb 3
Oracle-Restart klUCB	$\mathbf{37 \pm 37}$	$\mathbf{45 \pm 34}$	$\mathbf{257 \pm 86}$
klUCB	270 ± 76	162 ± 59	529 ± 148
Discounted-klUCB	1456 ± 214	1442 ± 440	1376 ± 37
SW-klUCB	177 ± 34	182 ± 34	1794 ± 71
Thompson sampling	493 ± 175	388 ± 147	1019 ± 245
DTS	209 ± 38	249 ± 39	2492 ± 52
M-klUCB	290 ± 29	534 ± 93	645 ± 141
CUSUM-klUCB	148 ± 32	152 ± 42	$\mathbf{490 \pm 133}$
GLR-klUCB(Local)	$\mathbf{74 \pm 31}$	$\mathbf{113 \pm 34}$	513 ± 97
GLR-klUCB(Global)	97 ± 32	134 ± 33	621 ± 103

Table 1: Mean regret $\pm$ 1 std-dev, for different algorithms on problems 1, 2 (with $T = 5000$) and 3 ($T = 20000$).

One can note that the best non-oracle strategies are actively adaptive, thus our experiments confirm that an efficient bandit algorithm (*e.g.*, klUCB) combined with an efficient change point detector (*e.g.*, GLR) provides efficient strategies for the piece-wise stationary model.

7. Conclusion

We proposed a new algorithm for the piece-wise stationary bandit problem, $\mathrm{GLR\text{-}klUCB}$, which combines the klUCB algorithm with the Bernoulli GLR change-point detector. This actively adaptive method attains state-of-the-art regret upper-bounds when tuned with a prior knowledge of the number of changes Υ_T, but *without any other prior knowledge on the problem*, unlike $\mathrm{CUSUM\text{-}UCB}$ and $\mathrm{M\text{-}UCB}$ that require to know a lower bound on the smallest magnitude of a change. We also gave numerical evidence of the efficiency of our proposal.

We believe that our new proof technique could be used to analyze $\mathrm{GLR\text{-}klUCB}$ under less stringent assumptions than the one made in this paper (and in previous work), that would require only a few "meaningful" changes to be detected. This interesting research direction is left for future work, but the hope is that the regret would be expressed in term of this number of meaningful changes instead of Υ_T. We shall also investigate whether actively adaptive approaches can attain a $\mathcal{O}(\sqrt{\Upsilon_T T})$ regret upper-bound without the knowledge of Υ_T. Finally, we would like to study in the future possible extension of our approach to the slowly varying model (Wei and Srivastava, 2018).

Acknowledgments

Thanks to Odalric-Ambrym Maillard at Inria Lille for useful discussions, and thanks to Christophe Moy at University Rennes 1. **Open Source:** the simulation code used for the experiments is using Python 3. It is open-sourced at `GitHub.com/SMPyBandits/SMPyBandits` and fully documented at `SMPyBandits.GitHub.io`, for more details see the companion article (Besson, 2018). The page `SMPyBandits.GitHub.io/NonStationaryBandits.html` explains how to reproduce the experiments used for this article.

References

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the Multi-Armed Bandit problem. In *Conference On Learning Theory*. PMLR, 2012.

R. Alami and O.-A. Maillard R. Féraud. Memory Bandits: Towards the Switching Bandit Problem Best Resolution. In *NIPS 2017 - 31st Conference on Neural Information Processing Systems*, 2017.

R. Allesiardo and R. Féraud. Exp3 with Drift Detection for the Switching Bandit Problem. In *IEEE Internation Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–7. IEEE, 2015.

R. Allesiardo, R. Féraud, and O.-A. Maillard. The Non-Stationary Stochastic Multi-Armed Bandit Problem. *International Journal of Data Science and Analytics*, 3(4):267–283, 2017.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multi-armed Bandit Problem. *Machine Learning*, 47(2):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The Non-Stochastic Multi-Armed Bandit Problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

G.A. Barnard. Control charts and stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 239–271, 1959.

M. Basseville, I. Nikiforov, et al. *Detection of Abrupt Changes: Theory And Application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

O. Besbes, Y. Gur, and A. Zeevi. Stochastic Multi-Armed Bandit Problem with Non-Stationary Rewards. In *Advances in Neural Information Processing Systems*, pages 199–207, 2014.

L. Besson. SMPyBandits: an Open-Source Research Framework for Single and Multi-Players Multi-Arms Bandits (MAB) Algorithms in Python, 2018. Code at `https://GitHub.com/SMPyBandits/SMPyBandits/`, documentation at `https://SMPyBandits.GitHub.io/`.

R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, and J. Palicot. Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings. In *12th EAI Conference on Cognitive Radio Oriented Wireless Network and Communication*, CROWNCOM Proceedings, 2017.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

Y. Cao, W. Zheng, B. Kveton, and Y. Xie. Nearly Optimal Adaptive Procedure for Piecewise-Stationary Bandit: a Change-Point Detection Approach. In *AISTATS*, Okinawa, Japan, 2019.

O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler Upper Confidence Bounds For Optimal Sequential Allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

A. Garivier and O. Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *Conference on Learning Theory (COLT)*, pages 359–376. PMLR, 2011.

A. Garivier and E. Moulines. On Upper-Confidence Bound Policies For Switching Bandit Problems. In *Algorithmic Learning Theory (ALT)*, pages 174–188. PMLR, 2011.

A. Garivier, H. Hadiji, P. Ménard, and G. Stoltz. KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. working paper or preprint, 2018. URL `https://hal.archives-ouvertes.fr/hal-01785705`.

C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag. Multi-Armed Bandit, Dynamic Environments and Meta-Bandits. In *NIPS 2006 Workshop, Online Trading Between Exploration And Exploitation*, 2006.

E. Kaufmann and W.M. Koolen. Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. arXiv preprint arXiv:1811.11419, 2018. URL `https://arXiv.org/abs/1811.11419`.

E. Kaufmann, N. Korda, and R. Munos. Thompson Sampling: an Asymptotically Optimal Finite-Time Analysis. In *Algorithmic Learning Theory (ALT)*, pages 199–213. PMLR, 2012.

R. Kerkouche, R. Alami, R. Féraud, N. Varsier, and P. Maillé. Node-based optimization of LoRa transmissions with Multi-Armed Bandit algorithms. In *ICT 2018 - 25th International Conference on Telecommunications*, Saint Malo, France, 2018.

L. Kocsis and C. Szepesvári. Discounted UCB. In *2nd PASCAL Challenges Workshop*, 2006.

T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

T.Z. Lai and H. Xing. Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential Analysis*, 29(2):162–175, 2010.

T. Lattimore and C. Szepesvári. Bandit Algorithms. 2019. URL http://downloads.tor-lattimore.com/book.pdf. Draft of Friday 18th January, 2019, Revision: 1699.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *International Conference on World Wide Web*. ACM, 2010.

F. Liu, J. Lee, and N. Shroff. A Change-Detection based Framework for Piecewise-stationary Multi-Armed Bandit Problem. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, 2018.

S. Maghsudi and E. Hossain. Multi-Armed Bandits with Application to 5G Small Cells. *IEEE Wireless Communications*, 23(3):64–73, 2016.

O.-A. Maillard. Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds. In *Algorithmic Learning Theory (ALT)*, 2019.

J. Mellor and J. Shapiro. Thompson Sampling in Switching Environments with Bayesian Online Change Detection. In *Artificial Intelligence and Statistics*, pages 442–450, 2013.

V. Raj and S. Kalyani. Taming Non-Stationary Bandits: a Bayesian Approach. arXiv preprint arXiv:1707.09727, 2017. URL https://arxiv.org/abs/1707.09727.

H. Robbins. Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

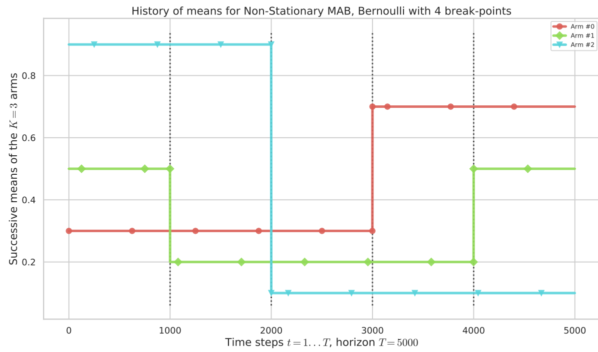
D. Siegmund and E.S. Venkatraman. Using the Generalized Likelihood Ratio Statistic for Sequential Detection of a Change Point. *The Annals of Statistics*, pages 255–271, 1995.

W. R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25, 1933.

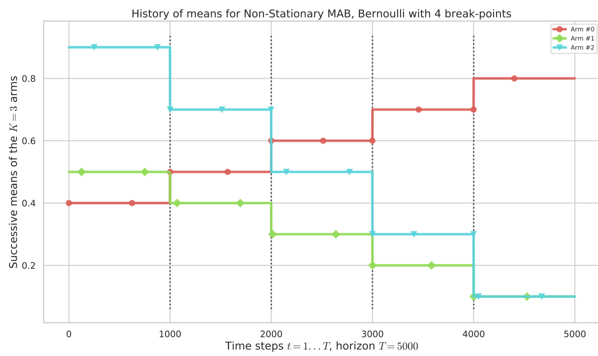
L. Wei and V. Srivastava. On Abruptly-Changing and Slowly-Varying Multiarmed Bandit Problems. arXiv preprint arXiv:1802.08380, 2018. URL https://arxiv.org/abs/1802.08380.

J. Y. Yu and S. Mannor. Piecewise-Stationary Bandit Problems with Side Observations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1177–1184. ACM, 2009.

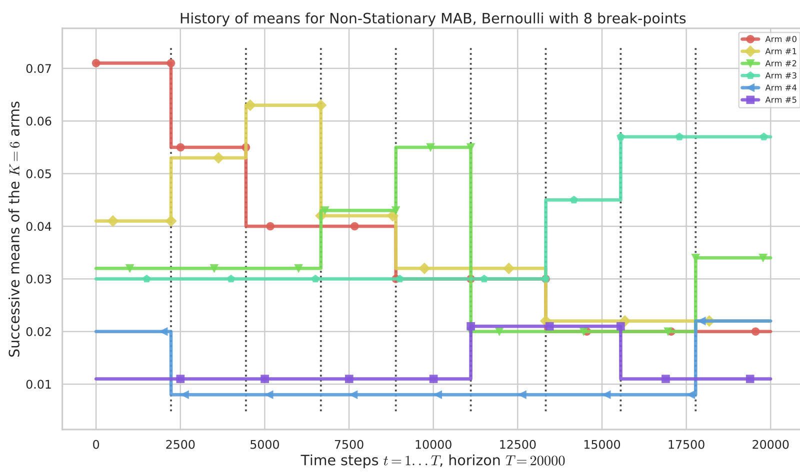
Appendix A. Illustrations of our Benchmark



(a) **Problem** 1: $K = 3$ arms with $T = 5000$, and $\Upsilon = 4$ changes occur on only one arm at a time (*i.e.*, $C = 4$).



(b) **Problem** 2: $K = 3$ arms with $T = 5000$, and $\Upsilon = 4$ changes occur on all arms (*i.e.*, $C = 12$).



(c) **Problem** 3: $K = 6$ arms with $T = 20000$, $\Upsilon = 8$ changes occur on most arms at a time ($C = 19$).

Figure 1: History of means of arms for three problems, $\mu_i(t)$ for $i \in \{1, \dots, K\}$ and $t \in \{1, \dots, T\}$.

Appendix B. Proof of Proposition 3

We consider one arm $i \in \{1, \dots, K\}$, and when the GLR-klUCB algorithm is running, we consider two time steps $s \leq t \in \mathbb{N}^*$, chosen between two restart times for that arm i. Lines 3-4 state that $A_u = u \mod \lceil \frac{K}{\alpha} \rceil$ if $u \mod \lceil \frac{K}{\alpha} \rceil \in \{1, \dots, K\}$ (see Algorithm 1 for details). Thus we have

$$
\begin{aligned}
n_i(t) - n_i(s) &= \sum_{u=s+1}^{t} \mathbb{1}(A_u = i) \\
&\geq \sum_{u=s+1}^{t} \mathbb{1}\left(A_u = i, A_u = u \mod \left\lceil \frac{K}{\alpha} \right\rceil\right) \\
&\geq \sum_{u=s+1}^{t} \mathbb{1}\left(u \mod \left\lceil \frac{K}{\alpha} \right\rceil = i\right) \\
&= \big(t - (s+1) + 1\big) / \left\lceil \frac{K}{\alpha} \right\rceil \geq \left\lfloor \frac{\alpha}{K}(t-s) \right\rfloor.
\end{aligned}
$$

Hence we have the result of Proposition 3.

Appendix C. Concentration Inequalities

C.1. Proof of Lemma 2

Lemma 2 is presented for bounded distributions and is actually valid for any sub-Bernoulli distribution. It could also be presented for more general distributions satisfying

$$
\mathbb{E}[e^{\lambda X}] \leq e^{\phi_\mu(\lambda)} \quad \text{with} \quad \mu = \mathbb{E}[X], \tag{8}
$$

where $\phi\mu(\lambda)$ is the log moment generating of some one-dimensional exponential family. The Bernoulli divergence $\mathrm{kl}(x, y)$ would be replaced by the corresponding divergence in that exponential family (which is the Kullback-Leibler divergence between two distributions of means x and y).

Let's go back to the Bernoulli case with divergence given in (2). A first key observation is

$$
s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right) = \inf_{\mu \in [0,1]} \left[s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \lambda\right) + (n-s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:n}, \lambda\right)\right].
$$

Hence the probability of a false alarm occurring is upper bounded as

$$
\begin{aligned}
\mathbb{P}_{\mu_0}\left(T_\delta < \infty\right) &\leq \mathbb{P}_{\mu_0}\left(\exists (s,n) \in \mathbb{N}^2, s < n : s\,\mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s)\,\mathrm{kl}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right) > \beta(n,\delta)\right) \\
&\leq \mathbb{P}_{\mu_0}\left(\exists (s,n) \in \mathbb{N}^2, s < n : s\,\mathrm{kl}\left(\hat{\mu}_{1:s}, \mu_0\right) + (n-s)\,\mathrm{kl}\left(\hat{\mu}_{s+1:n}, \mu_0\right) > \beta(n,\delta)\right) \\
&\leq \sum_{s=1}^{\infty} \mathbb{P}_{\mu_0}\left(\exists n > s : s\,\mathrm{kl}\left(\hat{\mu}_{1:s}, \mu_0\right) + (n-s)\,\mathrm{kl}\left(\hat{\mu}_{s+1:n}, \mu_0\right) > \beta(n,\delta)\right) \\
&= \sum_{s=1}^{\infty} \mathbb{P}_{\mu_0}\left(\exists r \in \mathbb{N} : s\,\mathrm{kl}\left(\hat{\mu}_{s}, \mu_0\right) + r\,\mathrm{kl}\left(\hat{\mu}'_{r}, \mu_0\right) > \beta(s+r,\delta)\right),
\end{aligned}
$$

where $\hat{\mu}_s$ and $\hat{\mu}'_r$ are the empirical means of respectively s and r *i.i.d.* observations with mean μ_0 and distribution ν, that are independent from the previous ones. As ν is sub-Bernoulli, the conclusion follows from Lemma 10 below and from the definition of $\beta(n, \delta)$:

$$
\begin{aligned}
&\mathbb{P}_{\mu_0}\left(T_\delta < \infty\right) \\
&\leq \sum_{s=1}^{\infty} \mathbb{P}_{\mu_0}\left(\exists r \in \mathbb{N}^* : s \,\mathrm{kl}\left(\hat{\mu}_s, \mu_0\right) + r \,\mathrm{kl}\left(\hat{\mu}'_r, \mu_0\right) > 6 \ln(1 + \ln(s+r)) + 2\mathcal{T}\left(\frac{\ln(3(s+r)^{3/2}/\delta)}{2}\right)\right) \\
&\leq \sum_{s=1}^{\infty} \mathbb{P}_{\mu_0}\left(\exists r \in \mathbb{N}^* : s \,\mathrm{kl}\left(\hat{\mu}_s, \mu_0\right) + r \,\mathrm{kl}\left(\hat{\mu}'_r, \mu_0\right) > 3 \ln(1 + \ln(s)) + 3\ln(1+\ln(r)) + 2\mathcal{T}\left(\frac{\ln(3s^{3/2}/\delta)}{2}\right)\right)
\end{aligned}
$$

And so we have $\mathbb{P}_{\mu_0}\left(T_\delta < \infty\right) \leq \sum_{s=1}^{\infty} \frac{\delta}{3s^{3/2}} \leq \delta$.

Lemma 10 *$(X_i)_{i\in\mathbb{N}}$ and $(Y_i)_{i\in\mathbb{N}}$ two independent* i.i.d. *processes with resp. means μ and μ' such that*

$$\mathbb{E}[e^{\lambda X_1}] \leq e^{\phi_\mu(\lambda)} \quad \textit{and} \quad \mathbb{E}[e^{\lambda Y_1}] \leq e^{\phi_{\mu'}(\lambda)},$$

where $\phi_\mu(\lambda) = \mathbb{E}_{X\sim\nu^\mu}[e^{\lambda X}]$ is the moment generating function of the distribution ν^μ, which is the unique distribution in an exponential family that has mean μ. Let $\mathrm{kl}(\mu,\mu') = \mathrm{KL}(\nu^\mu, \nu^{\mu'})$ be the divergence function associated to that exponential family. Introducing the notation $\hat{\mu}_s = \frac{1}{s}\sum_{i=1}^{s} X_i$ and $\hat{\mu}'_r = \frac{1}{r}\sum_{i=1}^{r} Y_i$, it holds that for every $s, r \in \mathbb{N}^*$,

$$\mathbb{P}\left(\exists r \in \mathbb{N}^* : s \,\mathrm{kl}\left(\hat{\mu}_s, \mu\right) + r \,\mathrm{kl}\left(\hat{\mu}'_r, \mu'\right) > 3 \ln(1 + \ln(s)) + 3\ln(1+\ln(r)) + 2\mathcal{T}\left(\frac{x}{2}\right)\right) \leq e^{-x},$$

where $\mathcal{T}$ is the function defined in (4).

Proof of Lemma 10. Using the same construction as in the proof of Theorem 14 in Kaufmann and Koolen (2018), one can prove that for every $\lambda \in I$ (for an interval I), there exists a non-negative super-martingale $M^\lambda(s)$ with respect to the filtration $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ that satisfies $\mathbb{E}[M^\lambda(s)] \leq 1$ and

$$\forall s \in \mathbb{N}^*, \quad M^\lambda(s) \geq e^{\lambda[s\mathrm{kl}(\hat{\mu}_s,\mu) - 3\ln(1+\ln(s))] - g(\lambda)}$$

for some function $g : I \to \mathbb{R}$. This super-martingale is of the form

$$M^\lambda(s) = \int e^{\eta\sum_{i=1}^{s} X_i - \phi_\mu(\lambda)s} d\pi(\eta)$$

for a well-chosen probability distribution π, and the function g can be chosen to be any

$$
\begin{aligned}
g_\xi : [0; 1/(1+\xi)] \quad &\longrightarrow \quad \mathbb{R} \\
\lambda \quad &\mapsto \quad \lambda(1+\xi)\ln\left(\frac{\pi^2}{3(\ln(1+\xi))^2}\right) - \ln(1-\lambda(1+\xi))
\end{aligned}
$$

for a parameter $\xi \in [0, 1/2]$.

Similarly, there exists an independent super-martingale $W^\lambda(r)$ w.r.t. the filtration $\mathcal{F}'_r = \sigma(Y_1, \dots, Y_r)$ such that

$$\forall r \in \mathbb{N}^*, \quad W^\lambda(r) \geq e^{\lambda[r\mathrm{kl}(\hat{\mu}'_r,\mu) - 3\ln(1+\ln(r))] - g(\lambda)},$$

for the same function $g(\lambda)$. In the terminology of Kaufmann and Koolen (2018), the processes $\boldsymbol{X}(s) = s\,\mathrm{kl}(\hat{\mu}_s, \mu) - 3\ln(1+\ln(s))$ and $\boldsymbol{Y}(s) = r\,\mathrm{kl}(\hat{\mu}_r, \mu) - 3\ln(1+\ln(r))$ are called g-DCC for Doob-Cramér-Chernoff, as Doob's inequality can be applied in combination with the Cramér-Chernoff method to obtain deviation inequalities that are uniform in time.

Here we have to modify the technique used in their Lemma 4 in order to take into account the two stochastic processes, and the presence of super-martingales instead of martingales (for which Doob inequality still works). One can write

$$\begin{aligned}
&\mathbb{P}\left(\exists r \in \mathbb{N}^* : s\,\mathrm{kl}\left(\hat{\mu}_s, \mu\right) + r\,\mathrm{kl}\left(\hat{\mu}'_r, \mu'\right) > 3\ln(1+\ln(s)) + 3\ln(1+\ln(r)) + u\right)\\
&\leq \mathbb{P}\left(\exists r \in \mathbb{N}^* : M^\lambda(s)W^\lambda(r) > e^{\lambda u - 2g(\lambda)}\right)\\
&= \lim_{n\to\infty} \mathbb{P}\left(\exists r \in \{1,\dots,n\} : M^\lambda(s)W^\lambda(r) > e^{\lambda u - 2g(\lambda)}\right)\\
&= \lim_{n\to\infty} \mathbb{P}\left(\sup_{r\in\{1,\dots,n\}} M^\lambda(s)W^\lambda(r) > e^{\lambda u - 2g(\lambda)}\right).
\end{aligned}$$

Using that $\tilde{M}(r) = M^\lambda(s)W^\lambda(r)$ is a super-martingale with respect to the filtration

$$\tilde{\mathcal{F}}_r = \sigma(X_1, \dots, X_s, Y_1, \dots, Y_r),$$

one can apply Doob's maximal inequality to obtain

$$\begin{aligned}
\mathbb{P}\left(\sup_{r\in\{1,\dots,n\}} M^\lambda(s)W^\lambda(r) > e^{\lambda u - 2g(\lambda)}\right) &\leq e^{-(\lambda u - 2g(\lambda))}\mathbb{E}[\tilde{M}(1))]\\
&= e^{-(\lambda u - 2g(\lambda))}\mathbb{E}[M^\lambda(s)W^\lambda(1)]\\
&\leq e^{-(\lambda u - 2g(\lambda))},
\end{aligned}$$

using that $M^\lambda(s)$ and $W^\lambda(1)$ are independent and have an expectation smaller than 1.

Putting things together yields

$$\mathbb{P}\left(\exists r \in \mathbb{N}^* : s\,\mathrm{kl}\left(\hat{\mu}_s, \mu\right) + r\,\mathrm{kl}\left(\hat{\mu}'_r, \mu'\right) > 3\ln(1+\ln(s)) + 3\ln(1+\ln(r)) + u\right) \leq e^{-\left(\lambda u - 2g_\xi(\lambda)\right)},$$

for any function g_ξ defined above. The conclusion follows by optimizing for both λ and ξ, using Lemma 18 in Kaufmann and Koolen (2018).

C.2. A Concentration Result Involving Two Arms

The following result is useful to control the probability of the good event in our two regret analyzes. Its proof follows from a straightforward application of the Cramér-Chernoff method (Boucheron et al., 2013).

Lemma 11 *Let $\hat{\mu}_{i,s}$ be the empirical mean of s i.i.d. observations with mean μ_i, for $i \in \{a, b\}$, that are σ^2-sub-Gaussian. Define $\Delta = \mu_a - \mu_b$. Then for any $s, r > 0$, we have*

$$\mathbb{P}\left(\frac{sr}{s+r}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,r} - \Delta\right)^2 \geq u\right) \quad \leq \quad 2\exp\left(-\frac{u}{2\sigma^2}\right).$$

Proof of Lemma 11. We first note that

$$\begin{aligned}
&\mathbb{P}\left(\frac{sr}{s+r}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,r} - \Delta\right)^2 \geq u\right)\\
&\leq \mathbb{P}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,r} \geq \Delta + \sqrt{\frac{s+r}{sr}u}\right) + \mathbb{P}\left(\hat{\mu}_{b,r} - \hat{\mu}_{a,s} \geq -\Delta + \sqrt{\frac{s+r}{sr}u}\right),
\end{aligned} \tag{9}$$

and those two quantities can be upper-bounded similarly using the Cramér-Chernoff method.

Let (X_i) and (Y_i) be two *i.i.d.* sequences that are σ^2 sub-Gaussian with mean μ_1 and μ_2 respectively. Let n_1 and n_2 be two integers and $\hat{\mu}_{1,n_1}$ and $\hat{\mu}_{2,n_2}$ denote the two empirical means based on n_1 observations from X_i, and n_2 observations from Y_i respectively. Then for every $\lambda > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu}_{1,n_1} - \hat{\mu}_{2,n_2} \geq \mu_1 - \mu_2 + x\right) &\leq \mathbb{P}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}(X_i - \mu_1) - \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i - \mu_2) \geq x\right) \\
&\leq \mathbb{P}\left(e^{\lambda\left(\frac{1}{n_1}\sum_{i=1}^{n_1}(X_i-\mu_1) - \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i-\mu_2)\right)} \geq e^{\lambda x}\right) \\
\text{(using Markov's inequality)} &\leq e^{-\lambda x}\mathbb{E}\left[e^{\lambda\frac{1}{n_1}\sum_{i=1}^{n_1}(X_i-\mu_1)}\right]\mathbb{E}\left[e^{-\lambda\frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i-\mu_2)}\right] \\
&= \exp\left(-\lambda x + n_1\phi_{X_1-\mu_1}\left(\frac{\lambda}{n_1}\right) + n_2\phi_{Y_1-\mu_2}\left(-\frac{\lambda}{n_2}\right)\right) \\
&\leq \exp\left(-\lambda x + \frac{\lambda^2\sigma^2}{2n_2} + \frac{\lambda^2\sigma^2}{2n_1}\right),
\end{aligned}$$

where the last inequality uses the sub-Gaussian property. Choosing the value $\lambda = \frac{x}{2[\sigma^2/(2n_1)+\sigma^2/(2n_2)]}$ which minimizes the right-hand side of the inequality yields

$$\mathbb{P}\left(\hat{\mu}_{1,n_1} - \hat{\mu}_{2,n_2} \geq \mu_1 - \mu_2 + x\right) \leq \exp\left(-\frac{n_1 n_2}{n_1+n_2}\frac{x^2}{2\sigma^2}\right).$$

Using this inequality twice in the right hand side of (9) concludes the proof.

Appendix D. Analysis of GLR-klUCB with Global Changes

We gave in Section 5 the following decomposition of the regret R_T,

$$\begin{aligned}
R_T \quad \leq \quad & T\mathbb{P}\left(\mathcal{E}_T^c\right) + \alpha T + \underbrace{\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^{T}\mathbb{1}\left(\mathrm{UCB}_{i_t^*}(t) \leq \mu_{i_t^*}(t)\right)\right]}_{(A)} \\
& + \underbrace{\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^{T}(\mu_{i_t^*}(t) - \mu_{I_t}(t))\mathbb{1}\left(\mathrm{UCB}_{I_t}(t) \geq \mu_{i_t^*}(t)\right)\right]}_{(B)}. \qquad (10)
\end{aligned}$$

D.1. Proof of Theorem 5.

We first introduce some notation for the proof: let $\hat{\tau}^{(k)}$ be the k-th change detected by the algorithm, leading to the k-th (full) restart and let $\hat{\tau}(t)$ be the last time before t that the algorithm restarted. We denote $n_i(t) = \sum_{s=\tau(t)+1}^{t}\mathbb{1}(A_s = i)$ the number of selections of arm i since the last (global) restart, and $\hat{\mu}_i(t) = \frac{1}{n_i(t)}\sum_{s=\tau(t)+1}^{t} X_{i,s}\mathbb{1}(A_s = i)$ their empirical average (if $n_i(t) \neq 0$).

As explained before, our analysis relies on the general regret decomposition (10), with the following appropriate good event. Let d^k be defined as in Assumption 4, we define

$$\mathcal{E}_T(\delta) = \left(\forall k \in \{1, \dots, \Upsilon_T\}, \hat{\tau}^{(k)} \in \left[\tau^{(k)} + 1, \tau^{(k)} + d^{(k)}\right]\right). \tag{11}$$

Under the good event, all the change points are detected within a delay at most $d^(k)$. Note that from Assumption 4, as the period between two changes are long enough, if $\mathcal{E}_T(\delta)$ holds, then for all change k, one has $\tau^{(k)} \leq \hat{\tau}^{(k)} \leq \tau^{(k+1)}$. Using Assumption 7, one can prove the following.

Lemma 12 *With $\mathcal{E}_T(\delta)$ defined as in (11), the "bad event" is unlikely: $\mathbb{P}(\mathcal{E}_T^c(\delta)) \leq \delta(K+1)\Upsilon_T$.*

We now turn our attention to upper bounding the two terms (A) and (B) in (10).

Upper bound on Term (A)

$$\begin{aligned}
(A) &\leq \mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^{T} \mathbb{1}\left(n_{i_t^*}(t)\,\mathrm{kl}\left(\hat{\mu}_{i_t^*}(t), \mu_{i_t^*}(t)\right) \geq f(t-\tau(t))\right)\right] \\
&\leq \sum_{k=1}^{\Upsilon_T} \mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=\tau^{(k)}+1}^{\tau^{(k+1)}} \mathbb{1}\left(n_{k^*}(t)\,\mathrm{kl}\left(\hat{\mu}_{k^*}(t), \mu_{k^*}^{(k)}\right) \geq f(t-\hat{\tau}(t))\right)\right] \\
&\leq \sum_{k=1}^{\Upsilon_T} \mathbb{E}\left[d^{(k)} + \mathbb{1}(\mathcal{E}_T)\sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \mathbb{1}\left(n_{k^*}(t)\,\mathrm{kl}\left(\hat{\mu}_{k^*}(t), \mu_{k^*}^{(k)}\right) \geq f(t-\hat{\tau}^{(k)})\right)\right] \\
&\leq \sum_{k=1}^{\Upsilon_T} d^{(k)} + \sum_{k=1}^{\Upsilon_T} \mathbb{E}\left[\mathbb{1}(\mathcal{C}^{(k)})\sum_{t=\hat{\tau}^{(k)}}^{\tau^{(k+1)}} \mathbb{1}\left(n_{k^*}(t)\,\mathrm{kl}\left(\hat{\mu}_{k^*}(t), \mu_{k^*}\right) \geq f(t-\hat{\tau}^{(k)})\right)\right],
\end{aligned}$$

where we introduce the event $\mathcal{C}^{(k)}$ that all the changes up to the k-th have been detected:

$$\mathcal{C}^{(k)} = \left\{\forall j \leq k, \hat{\tau}^{(j)} \in \{\tau^{(j)} + 1, \dots, \tau^{(j)} + d^{(j)}\}\right\}. \tag{12}$$

Clearly, $\mathcal{E}_T \subseteq \mathcal{C}^{(k)}$ and $\mathcal{C}^{(k)}$ is $\mathcal{F}_{\hat{\tau}^{(k)}}$-measurable. Observe that conditionally to $\mathcal{F}_{\hat{\tau}^{(k)}}$, when $\mathcal{C}^{(k)}$ holds, $\hat{\mu}_{k^*}(t)$ is the average of samples that have all mean $\mu_{k^*}^{(k)}$. Thus, introducing $\hat{\mu}_s$ as a sequence of *i.i.d.* random variables with mean $\mu_{k^*}^{(k)}$, one can write

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}(\mathcal{C}^{(k)})\sum_{t=\hat{\tau}^{(k)}}^{\tau^{(k+1)}} \mathbb{1}\left(n_{k^*}(t)\,\mathrm{kl}\left(\hat{\mu}_{k^*}(t), \mu_{k^*}^{(k)}\right) \geq f(t-\hat{\tau}^{(k)})\right)\middle|\, \mathcal{F}_{\hat{\tau}^{(k)}}\right] \\
&= \mathbb{1}(\mathcal{C}^{(k)})\sum_{t=\hat{\tau}^{(k)}}^{\tau^{(k+1)}} \mathbb{E}\left[\mathbb{1}\left(n_{k^*}(t)\,\mathrm{kl}\left(\hat{\mu}_{k^*}(t), \mu_{k^*}^{(k)}\right) \geq f(t-\hat{\tau}^{(k)})\right) \mid \mathcal{F}_{\hat{\tau}^{(k)}}\right] \\
&\leq \mathbb{1}(\mathcal{C}^{(k)}) \sum_{t'=1}^{\tau^{(k+1)}-\hat{\tau}^{(k)}} \mathbb{P}\left(\exists s \leq t' : s\,\mathrm{kl}(\hat{\mu}_s, \mu_{k^*}^{(k)}) \geq f(t')\right) \\
&\leq \sum_{t=1}^{T} \frac{1}{t\ln(t)} \leq \ln(\ln(T)),
\end{aligned}$$

where the last but one inequality relies on the concentration inequality given in Lemma 2 of Cappé et al. (2013) and the fact that $f(t) = \ln(t) + 3\ln(\ln(t))$. Finally, using the law of total expectation yields

$$(A) \leq \sum_{k=1}^{\Upsilon_T} \left[d^{(k)} + \ln(\ln(T)) \right]. \tag{13}$$

Upper bound on Term (B). We let $\tilde{\mu}_{i,s}^{(k)}$ denote the empirical mean of the first s observations of arm i made after time $t = \hat{\tau}^{(k)} + 1$. Rewriting the sum in t as the sum of consecutive intervals $[\tau^{(k)} + 1, \tau^{(k+1)}]$,

$$\begin{aligned}
(B) &\leq \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \sum_{k=1}^{\Upsilon_T} \sum_{t=\tau^{(k)}}^{\tau^{(k+1)}} \left(\mu_{k^*}^{(k)} - \mu_{I_t}^{(k)} \right) \mathbb{1}\left(\mathrm{UCB}_{I_t}(t) \geq \mu_{k^*}^{(k)} \right) \Big] \\
&\leq \sum_{k=1}^{\Upsilon_T} \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \hat{\tau}^{(k)} + \mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \left(\mu_{k^*}^{(k)} - \mu_{I_t}^{(k)} \right) \mathbb{1}\left(\mathrm{UCB}_{I_t}(t) \geq \mu_{k^*}^{(k)} \right) \Big] \\
&\leq \sum_{k=1}^{\Upsilon_T} d_i^{(k)} + \sum_{i=1}^{K} \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \left(\mu_{k^*}^{(k)} - \mu_i^{(k)} \right) \mathbb{1}\left(I_t = i, \mathrm{UCB}_i(t) \geq \mu_{k^*}^{(k)} \right) \Big] \\
&\leq \sum_{k=1}^{\Upsilon_T} d_i^{(k)} + \sum_{i=1}^{K} \sum_{k=1}^{\Upsilon_T} \left(\mu_{k^*}^{(k)} - \mu_i^{(k)} \right) \times \\
&\qquad \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \sum_{s=1}^{t-\hat{\tau}^{(k)}} \mathbb{1}\left(I_t = i, n_i(t) = s \right) \mathbb{1}\left(s \,\mathrm{kl}(\tilde{\mu}_{i,s}^{(k)}, \mu_{k^*}^{(k)}) \leq f(\tau^{(k+1)} - \hat{\tau}^{(k)}) \right) \Big] \\
&\leq \sum_{k=1}^{\Upsilon_T} d_i^{(k)} + \sum_{i=1}^{K} \sum_{k=1}^{\Upsilon_T} \left(\mu_{k^*}^{(k)} - \mu_i^{(k)} \right) \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \sum_{s=1}^{n_i(\tau^{(k+1)})} \mathbb{1}\left(s \,\mathrm{kl}(\tilde{\mu}_{i,s}^{(k)}, \mu_{k^*}^{(k)}) \leq f(\tau^{(k+1)} - \tau^{(k)}) \right) \Big] \\
&\leq \sum_{k=1}^{\Upsilon_T} d_i^{(k)} + \sum_{i=1}^{K} \sum_{k=1}^{\Upsilon_T} \left(\mu_{k^*}^{(k)} - \mu_i^{(k)} \right) \mathbb{E}\Big[\mathbb{1}(\mathcal{C}^{(k)}) \sum_{s=1}^{n_i(\tau^{(k+1)})} \mathbb{1}\left(s \,\mathrm{kl}(\tilde{\mu}_{i,s}^{(k)}, \mu_{k^*}^{(k)}) \leq f(\tau^{(k+1)} - \tau^{(k)}) \right) \Big].
\end{aligned}$$

Conditionally to $\mathcal{F}_{\hat{\tau}^{(k)}}$, when $\mathcal{C}^{(k)}$ holds, for $s \in \{1, \dots, n_i(\tau^{(k+1)})\}$, $\tilde{\mu}_{i,s}^{(k)}$ is the empirical mean from *i.i.d.* observations of mean $\mu_i^{(k)}$. Therefore, introducing $\hat{\mu}_s$ as a sequence of *i.i.d.* random variables with mean $\mu_i^{(k)}$, it follows from the law of total expectation that

$$(B) \leq \sum_{k=1}^{\Upsilon_T} d^{(k)} + \sum_{i=1}^{K} \sum_{k=1}^{\Upsilon_T} \left(\mu_{k^*}^{(k)} - \mu_i^{(k)} \right) \sum_{s=1}^{\tau^{(k+1)} - \tau^{(k)}} \mathbb{P}\left(s \times \mathrm{kl}(\hat{\mu}_s, \mu_{k^*}^{(k)}) \leq f(\tau^{(k+1)} - \tau^{(k)}) \right).$$

If $\mu_{k^*}^{(k)} > \mu_i^{(k)}$ by definition, we can use the analysis of klUCB from Cappé et al. (2013) to further upper bound the right-most part, and we obtain

$$(B) \leq \sum_{k=1}^{\Upsilon_T} d^{(k)} + \sum_{i=1}^{K} \sum_{k=1}^{\Upsilon_T} \mathbb{1}\left(\mu_i^{(k)} \neq \mu_{k^*}^{(k)} \right) \left[\frac{\left(\mu_{k^*}^{(k)} - \mu_i^{(k)} \right)}{\mathrm{kl}(\mu_i^{(k)}, \mu_{k^*}^{(k)})} \ln(T) + \mathcal{O}\Big(\sqrt{\ln(T)} \Big) \right]. \tag{14}$$

Combining the regret decomposition (10) with Lemma 12 and the two upper bounds in (13) and (14),

$$R_T \leq 2\sum_{k=1}^{\Upsilon_T} \frac{4K}{\alpha\left(\Delta^{(k)}\right)^2}\beta(T,\delta)+\alpha T+\delta(K+1)\Upsilon_T+\sum_{k=1}^{\Upsilon_T}\sum_{i:\mu_i^{(k)}\neq\mu_{k^*}^{(k)}} \frac{\left(\mu_{k^*}^{(k)}-\mu_i^{(k)}\right)}{\mathrm{kl}\left(\mu_i^{(k)},\mu_{k^*}^{(k)}\right)}\ln(T)+\mathcal{O}\left(\sqrt{\ln(T)}\right),$$

which concludes the proof.

D.2. Controlling the probability of the good event: Proof of Lemma 12

Recall that $\mathcal{C}^{(k)}$ defined in (12) is the event that all the breakpoints up to the k-th have been correctly detected. Using a union bound, one can write

$$\mathbb{P}(\mathcal{E}_T^c) \leq \sum_{k=1}^{\Upsilon_T}\mathbb{P}\left(\hat{\tau}^{(k)} \notin \{\tau^{(k)}+1,\dots,\tau^{(k)}+d^{(k)}\}\middle|\ \mathcal{C}^{(k-1)}\right)$$

$$\leq \sum_{k=1}^{\Upsilon_T}\underbrace{\mathbb{P}\left(\hat{\tau}^{(k)}\leq\tau^{(k)}\mid\mathcal{C}^{(k-1)}\right)}_{(a)}+\sum_{k=1}^{\Upsilon_T}\underbrace{\mathbb{P}\left(\hat{\tau}^{(k)}\geq\tau^{(k)}+d^{(k)}\mid\mathcal{C}^{(k-1)}\right)}_{(b)}.$$

The final result follows by proving that $(a)\leq K\delta$ and $(b)\leq\delta$, as detailed below.

Upper bound on (a): controlling the false alarm. $\hat{\tau}^{(k)}\leq\tau^{(k)}$ implies that there exists an arm whose associated change point detector has experienced a false-alarm:

$$(a)\leq\mathbb{P}\left(\exists i,\exists s<t\leq n_i(\tau_i^{(k)}): s\,\mathrm{kl}\left(\tilde{\mu}_{i,1:s}^{(k-1)},\tilde{\mu}_{i,1:t}^{(k-1)}\right)+(t-s)\,\mathrm{kl}\left(\tilde{\mu}_{i,s+1:t}^{(k-1)},\tilde{\mu}_{i,1:t}^{(k-1)}\right)>\beta(t,\delta)\mid\mathcal{C}^{(k-1)}\right)$$
$$\leq\sum_{i=1}^{K}\mathbb{P}\left(\exists s<t: s\,\mathrm{kl}(\hat{\mu}_{1:s},\mu_i^{(k-1)})+(t-s)\,\mathrm{kl}(\hat{\mu}_{s+1:t},\mu_i^{(k-1)})>\beta(t,\delta)\right),$$

with $\hat{\mu}_{s:s'}=\sum_{r=s}^{s'}Z_{i,r}$ where $Z_{i,r}$ is an *i.i.d.* sequence with mean $\mu_i^{(k-1)}$. Indeed, conditionally to $\mathcal{C}^{(k-1)}$, the $n_i(\tau^{(k)})$ successive observations of arm i arm starting from $\hat{\tau}^{(k)}$ are *i.i.d.* with mean $\mu_i^{(k-1)}$. Using Lemma 10, term (a) is upper bounded by $K\delta$.

Upper bound on term (b): controlling the delay. From the definition of $\Delta^{(k)}$, there exists an arm i such that $\Delta^{(k)}=|\mu_i^{(k)}-\mu_i^{(k-1)}|$. We shall prove that it is unlikely that the change-point detector associated to i doesn't trigger within the delay $d^{(k)}$.

First, it follows from Proposition 3 that there exists $\bar{t}\in\{\tau^{(k)},\dots,\tau^{(k)}+d^{(k)}\}$ such that $n_i(\bar{t})-n_i(\tau^{(k)})=\overline{r}$ where $\overline{r}=\lfloor\frac{\alpha}{K}d^{(k)}\rfloor$ (as the mapping $t\mapsto n_i(t)-n_i(\tau^{(k)})$ is non-decreasing, is 0 at $t=\tau^{(k)}$ and its value at $\tau^{(k)}+d^{(k)}$ is larger than $\overline{r}$). Using that

$$\left(\hat{\tau}^{(k)}\geq\tau^{(k)}+d^{(k)}\right)\subseteq\left(\hat{\tau}^{(k)}\geq\bar{t}\right)$$

the event $\left(\hat{\tau}^{(k)}\geq\tau^{(k)}+d^{(k)}\right)$ further implies that

$$n_i(\tau^{(k)})\,\mathrm{kl}\left(\tilde{\mu}_{i,n_i(\tau^{(k)})}^{k-1},\tilde{\mu}_{i,n_i(\bar{t})}^{k-1}\right)+\overline{r}\,\mathrm{kl}\left(\tilde{\mu}_{i,n_i(\tau^{(k)}):n_i(\bar{t})}^{k-1},\tilde{\mu}_{i,n_i(\bar{t})}^{k-1}\right)\leq\beta(n_i(\tau^{(k)})+\overline{r},\delta),$$

where $\tilde{\mu}_{i,s}^{k-1}$ denotes the empirical mean of the s first observation of arm i since the $(k-1)$-th restart $\hat{\tau}^{(k-1)}$ and $\tilde{\mu}_{i,s:s'}^{k-1}$ the empirical mean that includes observation number s to number s'. Conditionally to $\mathcal{C}^{(k-1)}$, $\tilde{\mu}_{i,n_i(\tau^{(k)})}^{k-1}$ is the empirical mean of $n_i(\tau^{(k)})$ *i.i.d.* replications of mean μ^{k-1}, whereas $\tilde{\mu}_{i,n_i(\tau^{(k)}):n_i(\bar{t})}^{k-1}$ is the empirical mean of $\overline{r}$ *i.i.d.* replications of mean μ_i^k.

Moreover, due to Proposition 3, $n_i(\tau^{(k)})$ lies in the interval $\left[\left\lfloor \frac{\alpha}{K}\left(\tau^{(k)} - \hat{\tau}^{(k-1)}\right)\right\rfloor, \left(\tau^{(k)} - \hat{\tau}^{(k-1)}\right)\right]$. Conditionally to $\mathcal{C}^{(k-1)}$, one obtains furthermore using that $d^{(k-1)} \leq (\tau^{(k)} - \tau^{(k-1)})/2$ – which follows from Assumption 4 – that

$$n_i(\tau^{(k)}) \quad \in \quad \left\{\left\lfloor \frac{\alpha}{2K}(\tau^{(k)} - \tau^{(k-1)})\right\rfloor, \dots, \tau^{(k)} - \tau^{(k-1)}\right\} := \mathcal{I}_k.$$

Introducing $\hat{\mu}_{a,s}$ (resp. $\hat{\mu}_{b,s}$) the empirical mean of s *i.i.d.* observations with mean $\hat{\mu}_i^{(k-1)}$ (resp. $\hat{\mu}_i^{(k)}$), such that $\hat{\mu}_{a,s}$ and $\hat{\mu}_{b,r}$ are independent, it follows that

$$(b) \leq \mathbb{P}\left(\exists s \in \mathcal{I}_k : s\,\mathrm{kl}\left(\hat{\mu}_{a,s}, \frac{s\hat{\mu}_{a,s} + \overline{r}\hat{\mu}_{b,\overline{r}}}{s+\overline{r}}\right) + \overline{r}\,\mathrm{kl}\left(\hat{\mu}_{b,\overline{r}}, \frac{s\hat{\mu}_{a,s} + \overline{r}\hat{\mu}_{b,\overline{r}}}{s+\overline{r}}\right) \leq \beta(s+\overline{r}, \delta)\right),$$

where we have also used that $\tilde{\mu}_{i,n_i(\bar{t})}^{k-1} = \left(n_i(\tau^{(k)})\tilde{\mu}_{i,n_i(\tau^{(k)})}^{k-1} + \overline{r}\tilde{\mu}_{i,n_i(\tau^{(k)}):n_i(\bar{t})}^{k-1}\right)/(n_i(\tau^{(k)}) + \overline{r})$.

Using Pinsker's inequality and introducing the gap $\Delta_i^{(k)} = \mu_i^{(k-1)} - \mu_i^{(k)}$ (which is such that $\Delta^{(k)} = |\Delta_i^k|$), one can write

$$\begin{aligned}
(b) &\leq \mathbb{P}\left(\exists s \in \mathcal{I}_k : \frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}}\right)^2 \leq \beta(s+\overline{r}, \delta)\right) \\
&\leq \mathbb{P}\left(\exists s \in \mathbb{N} : \frac{2sr}{s+r}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,s} - \Delta_i^{(k)}\right)^2 \geq \beta(s+r, \delta)\right) \\
&+ \mathbb{P}\left(\exists s \in \mathcal{I}_k : \frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}} - \Delta_i^{(k)}\right)^2 \leq \beta(s+\overline{r}, \delta), \frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}}\right)^2 \leq \beta(s+\overline{r}, \delta)\right)
\end{aligned}$$

Using Lemma 11 (given above in Appendix C.2) and a union bound, the first term in the right hand side is upper bounded by δ (as $\beta(r+s, \delta) \geq \beta(r, \delta) \geq \ln(3s\sqrt{s}/\delta)$). For the second term, we use the observation

$$\frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}} - \Delta_i^{(k)}\right)^2 \leq \beta(s+\overline{r}, \delta) \;\;\Rightarrow\;\; |\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}}| \geq |\Delta_i^{(k)}| - \sqrt{\frac{s+\overline{r}}{2\overline{r}s}\beta(s+\overline{r}, \delta)}$$

and, using that $\Delta^{(k)} = |\Delta_i^k|$, one obtains

$$(b) \leq \delta + \mathbb{P}\left(\exists s \in \mathcal{I}_k : \Delta^{(k)} \leq 2\sqrt{\frac{s+\overline{r}}{2s\overline{r}}\beta(s+\overline{r}, \delta)}\right). \tag{15}$$

Define $s_{\min} = \left\lfloor \frac{\alpha}{K}(\tau^{(k)} - \tau^{(k-1)})/2\right\rfloor$. Using that the mappings $s \mapsto (s+\overline{r})/s\overline{r}$ and $s \mapsto \beta(s+\overline{r}, \delta)$ are respectively decreasing and increasing in s, one has, for all $s \in \mathcal{I}_k$,

$$2\frac{s+\overline{r}}{s\overline{r}}\beta\left(s+\overline{r}, \delta\right) \quad \leq \quad 2\frac{s_{\min}+\overline{r}}{s_{\min}\overline{r}}\beta\left(T, \delta\right) \leq \frac{4\beta(T, \delta)}{\left\lfloor \frac{\alpha}{K}d^{(k)}\right\rfloor},$$

where the last inequality follows from the fact that $\overline{r} \leq s_{\min}$ as $d^{(k)} \leq (\tau^{(k)} - \tau^{(k-1)})/2$ by Assumption 4. Now the definition of $d^{(k)}$ readily implies that

$$\left\lfloor \frac{\alpha}{K} d^{(k)} \right\rfloor > \frac{4\beta(T,\delta)}{\left(\Delta^{(k)}\right)^2},$$

which yields

$$\forall s \in \mathcal{I}_k, \;\; 2\frac{s+\overline{r}}{s\overline{r}}\beta\left(s+\overline{r},\delta\right) \leq \left(\Delta^{(k)}\right)^2.$$

Hence, the probability in the right-hand side of (15) is zero, which yields $(b) \leq \delta$.

Appendix E. Analysis of GLR-klUCB with Local Changes

E.1. Proof of Theorem 8

Our analysis relies on the general regret decomposition (10), with the following appropriate good event.

$$\mathcal{E}_T(\alpha,\delta) = \left(\forall i \in \{1,\dots,K\}, \forall \ell \in \{1,\dots,\mathrm{NC}_i\}, \hat{\tau}_i^{(\ell)} \in \left[\tau_i^{(\ell)}+1, \tau_i^{(\ell)}+d_i^{(\ell)}\right]\right), \tag{16}$$

where $\hat{\tau}_i^{(\ell)}$ is defined as the ℓ-th change detected by the algorithm on arm i and $d_i^{(\ell)} = d_i^{(\ell)}(\alpha,\delta)$ is defined as in Assumption 7. Using this assumption, one can prove the following.

Lemma 13 *The "bad event" is highly unlikely, as it satisfies* $\mathbb{P}((\mathcal{E}_T(\alpha,\delta))^c) \leq 2\delta \mathrm{C}_T$.

We now turn our attention to upper bounding terms (A) and (B) in (10).

Upper bound on term (A).

$$\begin{aligned}
(A) &\leq \mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^{T}\mathbb{1}\left(n_{i_t^*}(t)\mathrm{kl}\left(\hat{\mu}_{i_t^*}(t), \mu_{i_t^*}(t)\right) \geq f(t-\tau_{i_t^*}(t))\right)\right] \\
&\leq \sum_{i=1}^{K}\sum_{k=1}^{\mathrm{NC}_i}\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^{T}\mathbb{1}(i_t^*=i)\mathbb{1}(\tau_i(t)=\hat{\tau}_i^{(\ell)})\mathbb{1}\left(n_i(t)\mathrm{kl}\left(\hat{\mu}_i(t),\mu_i\right) \geq f(t-\hat{\tau}_i^{(\ell)})\right)\right] \\
&\leq \sum_{i=1}^{K}\sum_{\ell=0}^{\mathrm{NC}_i}\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=\hat{\tau}_i^{(\ell)}}^{\tau_i^{(\ell+1)}}\mathbb{1}\left(n_i(t)\mathrm{kl}\left(\hat{\mu}_i(t),\mu_i\right) \geq f(t-\hat{\tau}_i^{(\ell)})\right) + d_i^{(\ell)}(T)\right] \\
&\leq \sum_{i=1}^{K}\sum_{\ell=0}^{\mathrm{NC}_i} d_i^{(\ell+1)}(T) + \sum_{i=1}^{K}\sum_{\ell=0}^{\mathrm{NC}_i}\mathbb{E}\left[\mathbb{1}(\mathcal{C}_i^{(\ell)})\sum_{t=\hat{\tau}_i^{(\ell)}}^{\tau_i^{(\ell+1)}}\mathbb{1}\left(n_i(t)\mathrm{kl}\left(\hat{\mu}_i(t),\mu_i\right) \geq f(t-\hat{\tau}_i^{(\ell)})\right)\right],
\end{aligned}$$

where we introduce the event $\mathcal{C}_i^{(\ell)}$ that all the changes up to the ℓ-th have been detected:

$$\mathcal{C}_i^{(\ell)} = \left\{\forall j \leq \ell, \hat{\tau}_i^{(j)} \in \left[\tau_i^{(j)}+1, \tau_i^{(j)}+d_i^{(\ell)}\right]\right\}. \tag{17}$$

Clearly, $\mathcal{E}_T \subseteq \mathcal{C}_i^{(\ell)}$ and $\mathcal{C}_i^{(\ell)}$ is $\mathcal{F}_{\hat{\tau}_i^{(\ell)}}$-measurable. Observe that conditionally to $\mathcal{F}_{\hat{\tau}_i^{(\ell)}}$, when $\mathbb{1}(\mathcal{C}_i^{(\ell)})$ holds, $\hat{\mu}_i(t)$ is the average of samples that have all mean $\mu_i^{(\ell)}$. Thus, introducing $\hat{\mu}_s$ as a sequence of *i.i.d.* random variables with mean $\mu_i^{(\ell)}$, one can write

$$\mathbb{E}\left[\mathbb{1}(\mathcal{C}_i^{(\ell)}) \sum_{t=\hat{\tau}_i^{(\ell)}}^{\tau_i^{(\ell+1)}} \mathbb{1}\left(n_i(t)\mathrm{kl}\left(\hat{\mu}_i(t), \mu_i\right) \geq f(t-\hat{\tau}_i^{(\ell)})\right) \middle| \mathcal{F}_{\hat{\tau}_i^{(\ell)}}\right]$$

$$= \mathbb{1}(\mathcal{C}_i^{(\ell)}) \sum_{t=\hat{\tau}_i^{(\ell)}}^{\tau_i^{(\ell+1)}} \mathbb{E}\left[\mathbb{1}\left(n_i(t)\mathrm{kl}\left(\hat{\mu}_i(t), \mu_i\right) \geq f(t-\hat{\tau}_i^{(\ell)})\right) \mid \mathcal{F}_{\hat{\tau}_i^{(\ell)}}\right]$$

$$\leq \mathbb{1}(\mathcal{C}_i^{(\ell)}) \sum_{t=1}^{\tau_i^{(\ell+1)}-\hat{\tau}_i^{(\ell)}} \mathbb{P}\left(\exists s \leq t' : s \times \mathrm{kl}(\hat{\mu}_s, \mu_i^{(\ell)}) \geq f(t')\right)$$

$$\leq \sum_{t=1}^{T} \frac{1}{t\ln(t)} \leq \ln(\ln(T)),$$

where the last but one inequality relies on the concentration inequality given in Lemma 2 of Cappé et al. (2013), and the fact that $f(t) = \ln(t) + 3\ln(\ln(t))$. Finally, using the law of total expectation yields

$$(A) \leq \sum_{i=1}^{K}\sum_{\ell=0}^{\mathrm{NC}_i}\left[d_i^{(\ell+1)}(T) + \ln(\ln(T))\right]. \tag{18}$$

Upper bound on term (B). Recall that $\mu_{i,\ell}^*$ is defined from the statement of Theorem 8 as the smallest value that still outperforms arm i on the interval between the ℓ and $(\ell+1)$-st change of i. We let $\tilde{\mu}_{i,s}^{(\ell)}$ denote the empirical mean of the first s observations of arm i made after time $t = \hat{\tau}_i^{(\ell)} + 1$. To upper bound Term B, we introduce a sum over all arms and rewrite the sum in t as a sum of consecutive intervals $[\tau_i^{(\ell)} + 1, \tau_i^{(\ell+1)}]$. The decomposition using furthermore that on $\mathcal{E}_T$, $\tau_i^{\ell} \leq \hat{\tau}_i^{\ell} \leq \tau_i^{\ell+1}$ for all changes, by Assumption 7.

$$(B) \leq \sum_{i=1}^{K}\mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T)\sum_{\ell=1}^{\mathrm{NC}_i}\sum_{t=\tau_i^{(\ell)}}^{\tau_i^{(\ell+1)}}\left(\mu_{i_t^*}(t) - \overline{\mu}_i^{(\ell)}\right)\mathbb{1}\left(I_t = i, \mathrm{UCB}_i(t) \geq \mu_{i,\ell}^*\right)\Big]$$

$$\leq \sum_{i=1}^{K}\sum_{\ell=1}^{\mathrm{NC}_i}\mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T)\hat{\tau}_i^{(\ell)} + \mathbb{1}(\mathcal{E}_T)\sum_{t=\hat{\tau}_i^{(\ell)}+1}^{\tau_i^{(\ell+1)}}\mathbb{1}\left(I_t = i, \mathrm{UCB}_i(t) \geq \mu_{i,\ell}^*\right)\Big]$$

$$\leq \sum_{i=1}^{K}\sum_{\ell=1}^{\mathrm{NC}_i} d_i^{(\ell)} + \sum_{i=1}^{K}\sum_{\ell=1}^{\mathrm{NC}_i}\mathbb{E}\Big[\mathbb{1}(\mathcal{C}_i^{(\ell)})\sum_{s=1}^{n_i(\tau_i^{(\ell+1)})}\mathbb{1}\left(s \times \mathrm{kl}(\tilde{\mu}_{i,s}^{(\ell)}, \mu_{i,\ell}^*) \leq f(\tau_i^{(\ell+1)} - \tau_i^{(\ell)})\right)\Big].$$

The last inequality relies on introducing a sum over $\mathbb{1}(n_i(t) = s)$ and swapping the sums. Conditionally to $\mathcal{F}_{\hat{\tau}_i^{(\ell)}}$, when $\mathcal{C}_i^{(\ell)}$ holds, for $s \in \{1, \ldots, n_i(\tau_i^{(\ell+1)})\}$, $\tilde{\mu}_{i,s}^{(\ell)}$ is the empirical mean from *i.i.d.* observations of mean $\overline{\mu}_i^{(\ell)}$. Therefore, introducing $\hat{\mu}_s$ as a sequence of *i.i.d.* random variables with mean $\overline{\mu}_i^{(\ell)}$, it follows from the law of total expectation that

$$(B) \leq \sum_{i=1}^{K} \sum_{\ell=1}^{\mathrm{NC}_i} d_i^{(\ell)} + \sum_{i=1}^{K} \sum_{\ell=1}^{\mathrm{NC}_i} \sum_{s=1}^{\tau_i^{(\ell+1)} - \tau_i^{(\ell)}} \mathbb{P}\left(s \times \mathrm{kl}(\hat{\mu}_s, \mu_{i,\ell}^*) \leq f(\tau_i^{(\ell+1)} - \tau_i^{(\ell)})\right).$$

As $\mu_{\ell,i}^* > \mu_i^{(\ell)}$ by definition, we can use the same analysis as in the proof of Fact 2 in Appendix A.2 of Cappé et al. (2013) to show that

$$(B) \leq \sum_{i=1}^{K} \sum_{\ell=1}^{\mathrm{NC}_i} d_i^{(\ell)} + \sum_{i=1}^{K} \sum_{\ell=1}^{\mathrm{NC}_i} \left[\frac{\ln(T) + 3\ln\ln(T)}{\mathrm{kl}(\overline{\mu}_i^{(\ell)}, \mu_{i,\ell}^*)} + O\left(\sqrt{\ln(T)}\right)\right]. \tag{19}$$

The result follows by combining the decomposition (10) with Lemma 13 and the bounds (18) and (19).

E.2. Proof of Lemma 13

With the event $\mathcal{C}_i^{(\ell)}$ defined in (17), a simple union bound yields

$$\mathbb{P}(\mathcal{E}_T^c) \leq \sum_{i=1}^{K} \sum_{\ell=1}^{\mathrm{NC}_i} \underbrace{\mathbb{P}\left(\hat{\tau}_i^{(\ell)} \leq \tau_i^{(\ell)} \mid \mathcal{C}_i^{(\ell-1)}\right)}_{(a)} + \sum_{i=1}^{K} \sum_{\ell=1}^{\mathrm{NC}_i} \underbrace{\mathbb{P}\left(\hat{\tau}_i^{(\ell)} \geq \tau_i^{(\ell)} + d_i^{(\ell)} \mid \mathcal{C}_i^{(\ell-1)}\right)}_{(b)}.$$

The final result follows by proving that the terms (a) and (b) are both upper bounded by δ.

Upper bound on (a): controlling the false alarms. Under the bandit algorithm, the change point detector associated to arm i is based on (possibly much) less than $t - \tau_i(t)$ samples from arm i, which makes false alarm even less likely to occur. More precisely, we upper bound term (a) by

$$\begin{aligned}(a) &\leq \mathbb{P}\left(\exists s < t \leq n_i(\tau_i^{(\ell)}) : s \times \mathrm{kl}\left(\tilde{\mu}_{i,1:s}^{(\ell-1)}, \tilde{\mu}_{i,1:t}^{(\ell-1)}\right) + (t-s) \times \mathrm{kl}\left(\tilde{\mu}_{i,s+1:t}^{(\ell-1)}, \tilde{\mu}_{i,1:t}^{(\ell-1)}\right) > \beta(t,\delta) \mid \mathcal{C}_i^{(\ell-1)}\right) \\ &\leq \mathbb{P}\left(\exists s < t : s \times \mathrm{kl}(\hat{\mu}_{1:s}, \mu_i^{(\ell-1)}) + (t-s) \times \mathrm{kl}(\hat{\mu}_{s+1:t}, \mu_i^{(\ell-1)}) > \beta(t,\delta)\right),\end{aligned}$$

with $\hat{\mu}_{s:s'} = \sum_{r=s}^{s'} Z_{i,r}$ where $Z_{i,r}$ is an *i.i.d.* sequence with mean $\mu_i^{(\ell-1)}$. Indeed, conditionally to $\mathcal{C}_i^{(\ell-1)}$, the $n_i(\tau_i^{(\ell)})$ successive observations of arm i arm starting from $\hat{\tau}_i^{(\ell)}$ are *i.i.d.* with mean $\mu_i^{(\ell-1)}$. Using Lemma 10, term (a) is upper bounded by δ.

Upper bound on term (b): controlling the delay. Controlling the detection delay on arm i under an adaptive sampling scheme can be tricky. Here we need to leverage the forced exploration (Proposition 3) to be sure we have enough samples to ensure detection: the effect is that delays will be scaled by the exploration parameter α.

First, it follows from Proposition 3 that there exists $\overline{t} \in \left\{\tau_i^{(\ell)}, \dots, \tau_i^{(\ell)} + d_i^{(\ell)}\right\}$ such that $n_i(\overline{t}) - n_i(\tau_i^{(\ell)}) = \overline{r}$ where $\overline{r} = \lfloor \frac{\alpha}{K} d_i^{(\ell)} \rfloor$ (as the mapping $t \mapsto n_i(t) - n_i(\tau_i^{(\ell)})$ is non-decreasing, is 0 at $t = \tau_i^{(\ell)}$ and its value at $\tau_i^{(\ell)} + d_i^{(\ell)}$ is larger than $\overline{r}$). Using that $(\hat{\tau}_i^{(\ell)} \geq \tau_i^{(\ell)} + d_i^{(\ell)}) \subseteq (\hat{\tau}_i^{(\ell)} \geq \overline{t})$, the event $(\hat{\tau}_i^{(\ell)} \geq \tau_i^{(\ell)} + d_i^{(\ell)})$ further implies that

$$n_i(\tau_i^{(\ell)})\, \mathrm{kl}\left(\tilde{\mu}_{i,n_i(\tau_i^{(\ell)})}^{\ell-1}, \tilde{\mu}_{i,n_i(\overline{t})}^{\ell-1}\right) + \overline{r}\, \mathrm{kl}\left(\tilde{\mu}_{i,n_i(\tau_i^{(\ell)}):n_i(\overline{t})}^{\ell-1}, \tilde{\mu}_{i,n_i(\overline{t})}^{\ell-1}\right) \leq \beta(n_i(\tau_i^{(\ell)}) + \overline{r}, \delta),$$

where $\tilde{\mu}^{\ell-1}_{i,s}$ denotes the empirical mean of the s first observation of arm i since the $(\ell-1)$-th restart $\hat{\tau}_i^{(\ell-1)}$ and $\tilde{\mu}^{\ell-1}_{i,s:s'}$ the empirical mean that includes observation number s to number s'. Conditionally to $\mathcal{C}_i^{(\ell-1)}$, $\tilde{\mu}^{\ell-1}_{i,n_i(\tau_i^{(\ell)})}$ is the empirical mean of $n_i(\tau_i^{(\ell)})$ *i.i.d.* replications of mean $\mu_i^{\ell-1}$, whereas $\tilde{\mu}^{\ell-1}_{i,n_i(\tau_i^{(\ell)}):n_i(\bar{t})}$ is the empirical mean of $\overline{r}$ *i.i.d.* replications of mean μ_i^{ℓ}.

Moreover, due to Proposition 3, $n_i(\tau_i^{(\ell)})$ lies in $\left\{\left\lfloor \frac{\alpha}{K}\left(\tau_i^{(\ell)} - \hat{\tau}_i^{(\ell-1)}\right)\right\rfloor, \dots, \tau_i^{(\ell)} - \hat{\tau}_i^{(\ell-1)}\right\}$. Conditionally to $\mathcal{C}_i^{(\ell-1)}$, one obtains furthermore using that $d_i^{(\ell-1)} \leq (\tau_i^{(\ell)} - \tau_i^{(\ell-1)})/2$ – which follows from Assumption 7 – that

$$\begin{aligned} n_i(\tau_i^{(\ell)}) &\in \left\{\left\lfloor \frac{\alpha}{K}\left(\tau_i^{(\ell)} - \tau_i^{(\ell-1)} - d_i^{(\ell-1)}\right)\right\rfloor, \dots, \left(\tau_i^{(\ell)} - \tau_i^{(\ell-1)}\right)\right\} \\ n_i(\tau_i^{(\ell)}) &\in \left\{\left\lfloor \frac{\alpha}{2K}\left(\tau_i^{(\ell)} - \tau_i^{(\ell-1)}\right)\right\rfloor, \dots, \left(\tau_i^{(\ell)} - \tau_i^{(\ell-1)}\right)\right\} = \mathcal{I}_k. \end{aligned}$$

Introducing $\hat{\mu}_{a,s}$ (resp. $\hat{\mu}_{b,s}$) the empirical mean of s *i.i.d.* observations with mean $\hat{\mu}_i^{(\ell-1)}$ (resp. $\hat{\mu}_i^{(\ell)}$), such that $\hat{\mu}_{a,s}$ and $\hat{\mu}_{b,r}$ are independent, it follows that

$$(b) \leq \mathbb{P}\left(\exists s \in \mathcal{I}_k : s\,\mathrm{kl}\left(\hat{\mu}_{a,s}, \frac{s\hat{\mu}_{a,s} + \overline{r}\hat{\mu}_{b,\overline{r}}}{s+\overline{r}}\right) + \overline{r}\,\mathrm{kl}\left(\hat{\mu}_{b,\overline{r}}, \frac{s\hat{\mu}_{a,s} + \overline{r}\hat{\mu}_{b,\overline{r}}}{s+\overline{r}}\right) \leq \beta(s+\overline{r},\delta)\right),$$

where we have also used that $\tilde{\mu}^{\ell-1}_{i,n_i(\bar{t})} = \frac{n_i(\tau_i^{(\ell)})\tilde{\mu}^{\ell-1}_{i,n_i(\tau_i^{(\ell)})} + \overline{r}\tilde{\mu}^{\ell-1}_{i,n_i(\tau_i^{(\ell)}):n_i(\bar{t})}}{n_i(\tau_i^{(\ell)})+\overline{r}}$.

Using Pinsker's inequality and introducing the gap $\Delta_i^{(\ell)} = \mu_i^{(\ell-1)} - \mu_i^{(\ell)}$, one can write

$$\begin{aligned} (b) &\leq \mathbb{P}\left(\exists s \in \mathcal{I}_k : \frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}}\right)^2 \leq \beta(s+\overline{r},\delta)\right) \\ &\leq \mathbb{P}\left(\exists s \in \mathbb{N} : \frac{2sr}{s+r}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,s} - \Delta_i^{(\ell)}\right)^2 \geq \beta(s+r,\delta)\right) \\ &+ \mathbb{P}\left(\exists s \in \mathcal{I}_k : \frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}} - \Delta_i^{(\ell)}\right)^2 \leq \beta(s+\overline{r},\delta), \frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}}\right)^2 \leq \beta(s+\overline{r},\delta)\right) \end{aligned}$$

Using Lemma 11 stated in Appendix C.2 and a union bound, the first term in the right hand side is upper bounded by δ (as $\beta(r+s,\delta) \geq \beta(r,\delta) \geq \ln(3s\sqrt{s}/\delta)$). For the second term, we use the observation

$$\frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}} - \Delta_i^{(\ell)}\right)^2 \leq \beta(s+\overline{r},\delta) \;\Rightarrow\; |\hat{\mu}_{a,s} - \hat{\mu}_{b,\overline{r}}| \geq |\Delta_i^{(\ell)}| - \sqrt{\frac{s+\overline{r}}{2\overline{r}s}\beta(s+\overline{r},\delta)}$$

and finally get

$$(b) \leq \delta + \mathbb{P}\left(\exists s \in \mathcal{I}_k : |\Delta_i^{(\ell)}| \leq 2\sqrt{\frac{s+\overline{r}}{2s\overline{r}}\beta(s+\overline{r},\delta)}\right). \tag{20}$$

Define $s_{\min} = \left\lfloor \frac{\alpha}{K}(\tau_i^{(\ell)} - \tau_i^{(\ell-1)})/2\right\rfloor$. Using that the mappings $s \mapsto (s+\overline{r})/s\overline{r}$ and $s \mapsto \beta(s+\overline{r},\delta)$ are respectively decreasing and increasing in s, one has, for all $s \in \mathcal{I}_k$,

$$2\frac{s+\overline{r}}{s\overline{r}}\beta\left(s+\overline{r},\delta\right) \;\leq\; 2\frac{s_{\min}+\overline{r}}{s_{\min}\overline{r}}\beta\left(T,\delta\right) \leq \frac{4\beta(T,\delta)}{\left\lfloor \frac{\alpha}{K}d_i^{(\ell)}\right\rfloor},$$

where the last inequality follows from the fact that $\overline{r} \leq s_{\min}$ as $d_i^{(\ell)} \leq (\tau_i^{(\ell)} - \tau_i^{(\ell-1)})/2$ by Assumption 7. Now the definition of $d_i^{(\ell)}$ readily implies that

$$\left\lfloor \frac{\alpha}{K} d_i^{(\ell)} \right\rfloor > 4\beta(T,\delta) / \left(\Delta_i^{(\ell)}\right)^2 ,$$

which yields

$$\forall s \in \mathcal{I}_k, \;\; 2\frac{s+\overline{r}}{s\overline{r}} \beta\left(s+\overline{r},\delta\right) \leq \left(\Delta_i^{(\ell)}\right)^2 .$$

Hence, the probability in the right-hand side of (20) is zero, which yields $(b) \leq \delta$.

Appendix F. Time and Memory Costs of GLR-klUCB

As demonstrated, our proposal is empirically efficient in terms of regret, but it is important to also evaluate its cost in terms of both *time* and *memory*. Remember that Υ_T denotes the number of change-points. If we denote $d_{\max}$ the longest duration of a stationary sequence, the worst case is $d_{\max} = T$ for a stationary problem, and the easiest case is $d_{\max} \simeq T/\Upsilon_T$ typically for Υ_T evenly spaced change-points. We begin by reviewing the costs of the algorithms designed for stationary problems and then of other approaches.

For a stationary bandit problem, almost all *classical algorithms* (*i.e.*, designed for stationary problems) need and use a storage proportional to the number of arms, *i.e.*, $\mathcal{O}(K)$, as most of them only need to store a number of pulls and the empirical mean of rewards for each arm. They also have a time complexity $\mathcal{O}(K)$ at every time step $t \in \{1,\dots,T\}$, hence an optimal total time complexity of $\mathcal{O}(KT)$. In particular, this case includes UCB, Thompson sampling and klUCB.

Most algorithms designed for abruptly changing environments are more costly, both in terms of storage and computation time, as they need more storage and time to test for changes. The *oracle algorithm* presented in Section 6, combined with any efficient index policy, needs a storage at most $\mathcal{O}(K\Upsilon_T)$ as it stores the change-points, and have an optimal time complexity of $\mathcal{O}(KT)$ too.

On the first hand, passively adaptive algorithms should intuitively be more efficient, but as they use a non-constant storage, they are actually as costly as the oracle. For instance SW-UCB uses a storage of $\mathcal{O}(K\tau)$, increasing as T increases, and similarly for other passively adaptive algorithms. We highlight that to our knowledge, the Discounted Thompson sampling algorithm (DTS) is the only algorithm tailored for abruptly changing problems that is both efficient in terms of regret (see the simulations results, even though it has no theoretical guarantee), and optimal in terms of both computational and storage costs. Indeed, it simply needs a storage proportional to the number of arms, $\mathcal{O}(K)$, and a time complexity of $\mathcal{O}(KT)$ for a horizon T (see the pseudo-code in Raj and Kalyani (2017)). Note that the discounting scheme in Discounted-UCB (D-UCB) from Kocsis and Szepesvári (2006) requires to store the whole history, and not only empirical rewards of each arm, as after observing a reward, all previous rewards must be multiplied by γ^n if that arm was not seen for $n > 0$ times. So the storage cannot be simply proportional to K, but needs to grows as t grows. Therefore, D-UCB costs $\mathcal{O}(KT)$ in memory and $\mathcal{O}\left(KT^2\right)$ in time.

On the other hand, limited memory actively adaptive algorithms, like M-UCB, are even more costly. For instance, M-UCB would have the same cost of $\mathcal{O}(KTd_{\max})$, except that Cao et al. introduce a window-size w and run their CPD algorithm only using the last w observations of each arm. If w is constant w.r.t. the horizon T, their algorithm has a storage cost bounded by $\mathcal{O}(Kw)$ and a running time of $\mathcal{O}(KTw)$, being comparable to the cost $\mathcal{O}(KT)$ of the oracle approach. However in practice, as well as in the theoretical results, the window size should depend on T and a prior knowledge of the minimal change size $\widehat{\delta}$ (see Remark 1 in Cao et al. (2019)), and $w = \mathcal{O}(\ln(T)/\widehat{\delta}^2)$. Hence it makes more sense to consider that M-UCB has a time cost bounded by $\mathcal{O}(KT\ln(T))$ and a memory costs bounded by $\mathcal{O}(K\ln(T))$, which is better than our proposal but more costly than the oracle or DTS or stationary algorithms.

Time and memory cost of GLR-klUCB**.** On the other hand, actively adaptive algorithms are more efficient (when tuned correctly) but at the price of being more costly, both in terms of time and memory. The two algorithms using the CUSUM or GLR tests (as well as PHT), when used with an efficient and low-cost index policy (that is, choosing the arm to play only costs $\mathcal{O}(K)$ at any time t), are found to be efficient in terms of regret. However, they need to store all past rewards and pulls history, to be able to reset them when the CPD algorithm tells to do, so they have a memory cost of $\mathcal{O}(Kd_{\max})$, that is $\mathcal{O}(KT)$ in the worst case (compared to $\mathcal{O}(K)$ for algorithms designed for the stationary setting). They are also costly in terms of computation time, as at current time t, when trying to detect a change with n_i observations of arm i (*i.e.*, $(Z_{i,n})_{1\leq n\leq n_i}$ in Algorithm 1), the CPD algorithm (CUSUM or GLR) costs a time $\mathcal{O}(n_i)$. Indeed, it needs to compute sliding averages for every s in an interval of size n_i (*i.e.*, $\mu_{\text{left}} = \mu_{1:s}$ and $\mu_{\text{right}} = \mu_{s+1:n_i}$) and a test for each s which costs a constant time $\mathcal{O}(1)$ (*e.g.*, computing two kl and checking for a threshold for our GLR test). So for every s, the running time is $\mathcal{O}(1)$, if the sliding averages are computed iteratively based on a simple scheme: first, one compute the total average $\mu_{t_0:t}$ and set $\mu_{\text{left}} = 0$ and $\mu_{\text{right}} = \mu_{t_0:t}$. Then for every successive values of s, both the left and right sliding window means can be updated with a single memory access and two computations (*i.e.*, $\mathcal{O}(1)$):

$$z \leftarrow Z_{i,s+1},\ \mu_{\text{left}} \leftarrow \frac{s\mu_{\text{left}} + z}{s+1},\ \mu_{\text{right}} \leftarrow \frac{(n_i + 1 - s)\mu_{\text{right}} - z}{n_i - s},\ s \leftarrow s + 1.$$

To sum up, at every time step the CPD algorithm needs a time $\mathcal{O}(n_i) = \mathcal{O}(d_{\max})$, and at the end, the time complexity of CUSUM-klUCB as well as GLR-klUCB is $\mathcal{O}(KTd_{\max})$, which can be up-to $\mathcal{O}\big(KT^2\big)$, much more costly than $\mathcal{O}(KT)$ for klUCB for instance.

Our proposal GLR-klUCB requires a storage of the order of $\mathcal{O}(Kd_{\max})$ and a running time of the order of $\mathcal{O}(KTd_{\max})$, and the two bounds are validated experimentally, see Table 2.

Empirical measurements of computation times and memory costs. A theoretical analysis shows that there is a large gap between the costs of stationary or passively adaptive algorithms, and the costs of actively adaptive algorithms, for both computation time and memory consumption. We include here an extensive comparison of memory costs of the different algorithms. For instance on the same experiment as the one used for Table 4, that is problem 1 with $T = 5000$, and then with $T = 10000$ and $T = 20000$, and 100 independent runs, in our Python implementation (using the SMPyBandits library, Besson (2018)), we can measure the (mean) real memory cost[3] of the different algorithms. The Tables 2 and 3 included below give the mean ($\pm$ 1 standard-deviation) of real computation time and memory consumption, used by the algorithms. The computation time is normalized by the horizon, to reflect the (mean) time used for each time steps $t \in \{1, \dots, T\}$. We also found that our two optimizations described below in F.1 do not reduce the memory, thus we used $\Delta n = \Delta s = 20$ to speed-up the simulations. The conclusions to draw for these two Tables 2 and 3 are twofold.

First, we verify the results stated above for the time complexity of different algorithms. On the first hand, stationary and passively adaptive algorithms all have a time complexity scaling as $\mathcal{O}(T)$, as their normalized computation time is almost constant w.r.t. T. We check that TS and DTS are the fastest algorithms, 2.5 to 3 times faster than klUCB-based algorithms, due to the fact that sampling from a Beta posterior is typically faster than doing a (small) numerical optimization step to compute the sup in the klUCB indexes. We also check that the passively adaptive algorithms add a non-trivial but constant

3. We used one core of a Intel $i5$ Core CPU, GNU/Linux machine running Ubuntu 18.04 and Python v3.6, with 8 Gb of RAM.

Algorithms \Problems	$T = 5000$	$T = 10000$	$T = 20000$
Thompson sampling	$\mathbf{55\,\mu s \pm 11\,\mu s}$	$\mathbf{51\,\mu s \pm 6\,\mu s}$	$\mathbf{49\,\mu s \pm 4\,\mu s}$
DTS	$62\,\mu s \pm 11\,\mu s$	$60\,\mu s \pm 8\,\mu s$	$59\,\mu s \pm 7\,\mu s$
klUCB	$122\,\mu s \pm 13\,\mu s$	$125\,\mu s \pm 13\,\mu s$	$128\,\mu s \pm 11\,\mu s$
Discounted-klUCB	$94\,\mu s \pm 10\,\mu s$	$97\,\mu s \pm 12\,\mu s$	$103\,\mu s \pm 12\,\mu s$
SW-klUCB	$162\,\mu s \pm 21\,\mu s$	$169\,\mu s \pm 18\,\mu s$	$167\,\mu s \pm 12\,\mu s$
Oracle-Restart klUCB	$159\,\mu s \pm 31\,\mu s$	$157\,\mu s \pm 21\,\mu s$	$149\,\mu s \pm 15\,\mu s$
M-klUCB	$202\,\mu s \pm 25\,\mu s$	$220\,\mu s \pm 26\,\mu s$	$230\,\mu s \pm 19\,\mu s$
CUSUM-klUCB	$264\,\mu s \pm 53\,\mu s$	$227\,\mu s \pm 31\,\mu s$	$270\,\mu s \pm 33\,\mu s$
GLR-klUCB	$314\,\mu s \pm 50\,\mu s$	$399\,\mu s \pm 33\,\mu s$	$920\,\mu s \pm 180\,\mu s$

Table 2: **Normalized** computation time, for each time step $t \in \{1, \dots, T\}$, for different horizons.

Algorithms \Problems	$T = 5000$	$T = 10000$	$T = 20000$
Thompson sampling	813 B $\pm$ 63 B	819 B $\pm$ 27 B	818 B $\pm$ 35 B
DTS	946 B $\pm$ 38 B	946 B $\pm$ 38 B	946 B $\pm$ 39 B
klUCB	937 B $\pm$ 164 B	931 B $\pm$ 172 B	933 B $\pm$ 164 B
Discounted-klUCB	1 KiB $\pm$ 79 B	1 KiB $\pm$ 94 B	1 KiB $\pm$ 78 B
SW-klUCB	6 KiB $\pm$ 976 B	8 KiB $\pm$ 1 KiB	12 KiB $\pm$ 2 KiB
Oracle-Restart klUCB	11 KiB $\pm$ 3 KiB	19 KiB $\pm$ 7 KiB	31 KiB $\pm$ 17 KiB
M-klUCB	4 KiB $\pm$ 1 KiB	6 KiB $\pm$ 2 KiB	9 KiB $\pm$ 4 KiB
CUSUM-klUCB	8 KiB $\pm$ 4 KiB	11 KiB $\pm$ 6 KiB	15 KiB $\pm$ 10 KiB
GLR-klUCB	19 KiB $\pm$ 7 KiB	32 KiB $\pm$ 15 KiB	75 KiB $\pm$ 26 KiB

Table 3: **Non normalized** memory costs, for the same problem (Pb 1) with different horizons.

overhead on the computation times of their based algorithm, *e.g.*, SW-klUCB compared to klUCB, or DTS compared to TS. On the other hand, we also check that actively adaptive algorithms are most costly. M-klUCB normalized computation time is not increasing much when the horizon is doubled, as the window-size M was set to a constant w.r.t. the horizon T for this experiment. The complexity of GLR-klUCB follows the $\mathcal{O}(KT^2)$ bound we presented above.

Second, we also verify the results for the memory costs of the different algorithms. Similarly, stationary and passively adaptive algorithms based on a discount factor (D-UCB, DTS) have a memory cost constant w.r.t. the horizon T, as stated above, while algorithms based on a sliding-window have a memory cost increasing w.r.t. the horizon T. The Oracle-Restart and the actively adaptive algorithms see their memory costs increase similarly. These measurements validate the upper-bound we gave on their memory costs, $\mathcal{O}(Kd_{\max})$, as $d_{\max} \simeq T/\Upsilon$ for this problem with evenly-spaced change-points.

F.1. Two ideas of numerical optimization

In order to empirically improve this weakness of our proposal, we suggest two simple optimizations tweaks on GLR-klUCB (see Algorithm 1) to drastically speed-up its computation time.

1. The first optimization, parametrized by a constant $\Delta n \in \mathbb{N}^*$, is the following idea. We can test for statistical changes not at all time steps $t \in \{1, \dots, T\}$ but only every Δn time steps (*i.e.*, for t satisfying $t \mod \Delta n = 0$). In practice, instead of sub-sampling for the *time* t, we propose to

sub-sample for the number of samples of arm i before calling GLR to check for a change on arm i, that is, $n_i(t)$ in Algorithm 1. Note that the first heuristic using Δn can be applied to M-UCB as well as CUSUM-UCB and PHT-UCB (and variants using klUCB), with similar speed-up and typically leading to similar consequences on the algorithm's performance.

2. The second optimization is in the same spirit, and uses a parameter $\Delta s \in \mathbb{N}^*$. When running the GLR test with data $Z_1, \dots, Z_t$, instead of considering every splitting time steps $s \in \{1, \dots, t\}$, in the same spirit, we can skip some and test not at all time steps s but only every Δs time steps.

The new GLR test is using the stopping time $\tilde{T}_\delta$ defined in (6), with $\mathcal{T} = \{t \in [1, T], t \mod \Delta n = 0\}$ and $\mathcal{S}_t = \{s \in [1, t], s \mod \Delta s = 0\}$. The goal is to speed up the computation time of every call to the GLR test (*e.g.*, choosing $\Delta s = 10$, every call should be about 10 times faster), and to speed up the overhead cost of running the tests on top of the index policy (klUCB), by testing for changes less often (*e.g.*, choosing $\Delta n = 10$ should speed up the all computation by a factor 10).

Empirical validation of these optimization tricks. We consider the problem 1 presented above (Figure 1a), with $T = 5000$ and 100 repetitions, and we give the means ($\pm$ 1 standard-deviation) of both regret and computation time of GLR-klUCB with **Local** restarts, for different parameters Δn and Δs, in Table 4 below. The other parameters of GLR-klUCB are chosen as $\delta = 1/\sqrt{K \Upsilon_T T}$ and $\alpha = 0.1\sqrt{K \ln(T)/T}$ (from Corollary 9). The algorithm analyzed in Section 5 corresponds to $\Delta n = \Delta s = 1$.

On the same problem, the Oracle-Restart klUCB obtained a mean regret of 37 ± 37 for a running time of 711 ms $\pm$ 95 ms, while klUCB obtained a regret of 270 ± 76 for a time of 587 ms $\pm$ 47 ms. In comparison with the two other efficient approaches, M-klUCB obtained a regret of 290 ± 29 for a time of 943 ms $\pm$ 102 ms, and CUSUM-klUCB obtained a regret of 148 ± 32 for a time of 46 s $\pm$ 5 s. This shows that our proposal is very efficient compared to stationary algorithms, and comparable to the state-of-the-art actively adaptive algorithm. Moreover, this shows that two heuristics efficiently speed-up the computation times of GLR-klUCB. Choosing small values, like $\Delta n = 20, \Delta s = 20$, can speed-up GLR-klUCB, making it fast enough to be comparable to recent efficient approaches like M-UCB and even comparable to the oracle policy. It is very satisfying to see that the use of these optimizations do not reduce much the regret of GLR-klUCB, as it still outperforms most state-of-the-art algorithms, and significantly reduces the computation time as wanted. With such numerical optimization, GLR-klUCB is not significantly slower than klUCB while being much more efficient for piece-wise stationary problems.

Appendix G. Sensibility analysis of the exploration probability α

As demonstrated in our experiments, the choice of $\alpha = \alpha_0\sqrt{\Upsilon_T \ln(T)/T}$ for the exploration probability is a good choice for GLR-klUCB to be efficient. The dependency w.r.t. the horizon T comes from Corollary 9 when Υ_T is unknown, and we observe in the Table 5 below that the value of α_0 does not influence much the performance of our proposal, as long as $\alpha_0 \leq 1$. Different values of α_0 are explored, for problems 1, 2 and 4, and we average over 100 independent runs. Other parameters are set to **Local** restarts, $\delta_T = 1/\sqrt{\Upsilon_T T}$, and $\Delta n = \Delta s = 10$ to speed-up the experiments. In the three experiments, GLR-klUCB performs close to the Oracle-Restart klUCB, and outperforms all or almost all the other approaches, for all choices of α_0. We observe in Table 5 that the parameter α_0 does not have a significative impact on the performance, and that surprisingly, choosing $\alpha_0 = 0$ does not reduce the empirical performance of GLR-klUCB, which means that on some problems there is no need of a forced exploration. However, the analysis of GLR-klUCB is based on the forced exploration, and we found

$\Delta n \setminus \Delta s$	1	5	10	20
1	44 ± 29	44 ± 28	50 ± 31	53 ± 28
5	48 ± 29	41 ± 30	44 ± 28	47 ± 31
10	51 ± 32	43 ± 26	47 ± 28	46 ± 29
20	46 ± 31	46 ± 34	46 ± 31	49 ± 31

$\Delta n \setminus \Delta s$	1	5	10	20
1	**50** s $\pm$ **4.5** s	11.1 s $\pm$ 1.16 s	5.8 s $\pm$ 0.5 s	3.3 s $\pm$ 0.3 s
5	17.9 s $\pm$ 1.6 s	5.08 s $\pm$ 3.3 s	2.5 s $\pm$ 0.3 s	1.7 s $\pm$ 0.2 s
10	14.9 s $\pm$ 1.9 s	3.47 s $\pm$ 0.4 s	2.1 s $\pm$ 0.2 s	1.4 s $\pm$ 0.2 s
20	12.1 s $\pm$ 1.1 s	3.02 s $\pm$ 0.3 s	1.9 s $\pm$ 0.2 s	**1.4** s $\pm$ **0.1** s

Table 4: Effects of the two optimizations parameters Δn and Δs, on the mean regret R_T (top) and mean computation time (bottom) for GLR-klUCB on a simple problem. Using the optimizations with $\Delta n = \Delta s = 20$ does not reduce the regret much but speeds up the computations by about a **factor 50**.

that for a larger number of arms, or for problems when the optimal arm change constantly, the forced exploration is required.

Choice of α_0 **Problems**	Problem 1	Problem 2	Problem 4
$\alpha_0 = 1$	51 ± 29	79 ± 35	82 ± 45
$\alpha_0 = 0.5$	38 ± 29	70 ± 35	76 ± 39
$\alpha_0 = 0.1$	33 ± 29	69 ± 31	68 ± 32
$\boldsymbol{\alpha_0 = 0.05}$	$\mathbf{36 \pm 29}$	$\mathbf{65 \pm 33}$	$\mathbf{67 \pm 36}$
$\alpha_0 = 0.01$	38 ± 33	66 ± 33	71 ± 37
$\alpha_0 = 0.005$	40 ± 27	69 ± 30	73 ± 56
$\alpha_0 = 0.001$	38 ± 30	69 ± 34	67 ± 34
$\alpha_0 = 0$	36 ± 32	66 ± 36	67 ± 33

Table 5: Mean regret $\pm$ 1 standard-deviation, for different choices of $\alpha_0 \in [0, 1]$ on three problems of horizon $T = 5000$, for GLR-klUCB with $\alpha_T = \alpha_0 \sqrt{\Upsilon_T \ln(T)/T}$.

Appendix H. Additional Numerical Results

This section includes additional figures and numerical results, not presented in the main text. We describe and illustrate in Figures 2 and 3 two more problems.

Problem 4. Like problem 1, it uses $K = 3$ arms, $\Upsilon = 4$ change-points and $T = 5000$, but the stationary sequences between successive change-points no longer have the same length, as illustrated in Figure 2. Classical (stationary) algorithms such as klUCB can be "tricked" by large enough stationary sequences, as they learn with a large confidence the optimal arm, and then fail to adapt to a new optimal arm after a change-point. We observe below in Table 6 that they can suffer higher regret when the change-points are more spaced out, as this problem starts with a longer stationary sequence of length $T/2$.

Problem 5. Like problem 3, this hard problem is inspired from synthetic data obtained from a real-world database of clicks from *Yahoo!* (see Figure 3 from Liu et al. (2018)). It is harder, with $\Upsilon = 81$ change-points on $K = 5$ arms for a longer horizon of $T = 100000$. Some arms change at almost every

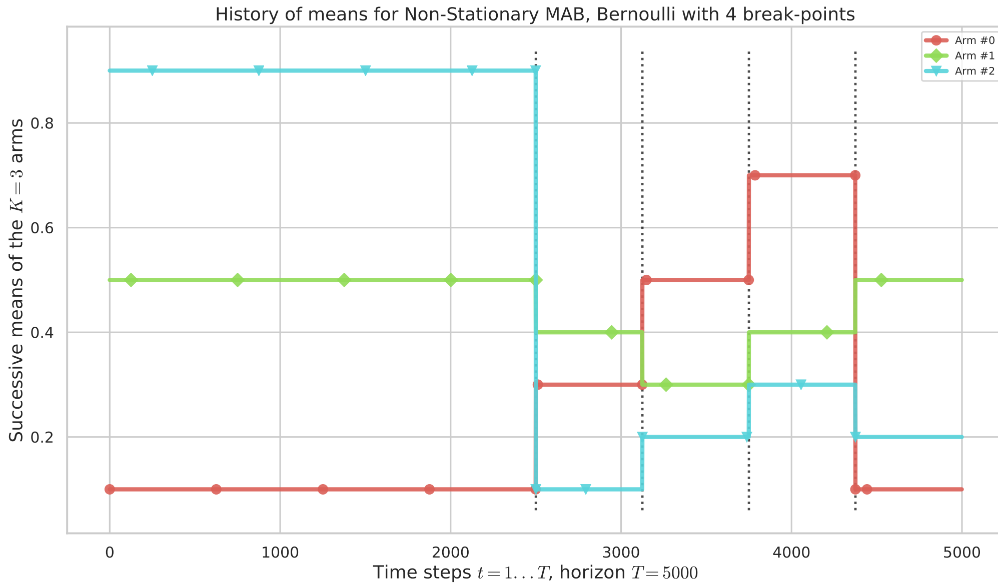


Figure 2: **Problem 4**: $K = 3$ arms with $T = 5000$, $\Upsilon = 4$ changes occur on all arms at a time (*i.e.*, $C = 12$).

time steps, for a total number of breakpoints $C = 179$, but the optimal arm is almost always the same one (arm 0, with ●). It is a good benchmark to see if the actively adaptive policies do not detect *too many changes*, as the Oracle-Restart policy suffers higher regret in comparison to klUCB. Means are also bounded in $[0.01, 0.07]$, with small gaps of amplitude ranging from 0.02 to 0.001, as shown in Figure 3.

Interpretations. We include below the figures showing the simulation results for the 5 problems presented in Section 6 and Section H. The results in terms of mean regret R_T are given in Tables 1 and 6, but it is also interesting to observe two plots for each experiments. First, we show the mean regret as a function of time (*i.e.*, R_t for $1 \leq t \leq T$), for the 9 algorithms considered (Discounted-klUCB is removed as it is always the most inefficient and suffers high regret). Efficient stationary algorithms, like TS and klUCB, typically suffer a linear regret after a change on the optimal arm changes if they had "too" many samples before the change-points (*e.g.*, on Figure 4 and even more on Figures 8 and 9). This illustrates the conjecture that classical algorithms can suffer linear regret even on simple piece-wise stationary problems.

On very simple problems, like problem 1, all the algorithms being designed for piece-wise stationary environments perform similarly, but as soon as the gaps are smaller or there is more changes, we clearly observe that our approach GLR-klUCB can outperform the two other actively adaptive algorithms CUSUM-klUCB and M-klUCB (*e.g.*, on Figure 6), and performs much more than passively adaptive algorithms DTS and SW-klUCB (*e.g.*, on Figure 7). Our approach, with the two options of **Local** or **Global** restarts, performs very closely to the oracle for problem 4.

Finally, in the case of hard problems, like problems 3 and 5, that have a lot of changes but where the optimal arm barely changes, we verify in Figure 9 that klUCB and TS can outperform the oracle policy. Indeed the oracle policy is suboptimal as it restarts as soon as one arm change but is unaware of the meaningful changes, and stationary policies which quickly identify the best arm will play it most of the times, achieving a smaller regret. We note that, sadly, all actively adaptive policies fail to outperform

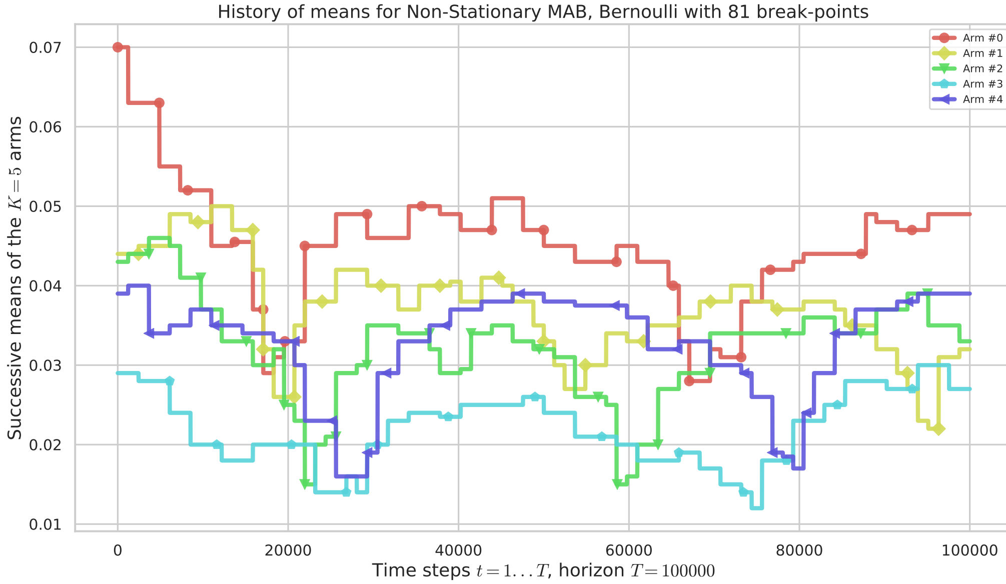


Figure 3: **Problem 5**: $K = 5$, $T = 100000$, $\Upsilon = 81$ changes occurring on some arms at every time ($C = 179$).

stationary policies on such hard problems, because they do not observe enough rewards from each arm between two restarts (*i.e.*, the Assumptions 4 and 7 for our Theorems 5 and 8 are not satisfied). We can also verify that the two options, **Local** and **Global** restart, for GLR-klUCB, give close results, and that the **Local** option is always better.

We also show the empirical distribution of the regret R_T, on Figure 5. It shows that all algorithms have a rather small variance on their regret, except Thompson Sampling which has a large tail due to its large mean regret on this (easy) non-stationary problems.

Algorithms \Problems	Pb 4 ($T = 5000$)	Pb 4 ($T = 10000$)	Pb 5
Oracle-Restart klUCB	$\mathbf{68 \pm 40}$	$\mathbf{86 \pm 50}$	126 ± 54
klUCB	615 ± 74	1218 ± 123	106 ± 36
SW-klUCB	202 ± 33	322 ± 47	228 ± 27
Discounted-klUCB	911 ± 210	1741 ± 200	2085 ± 910
Thompson sampling	756 ± 65	1476 ± 137	$\mathbf{88 \pm 39}$
DTS	250 ± 39	481 ± 58	238 ± 24
M-klUCB	337 ± 46	544 ± 47	116 ± 36
CUSUM-klUCB	267 ± 69	343 ± 94	117 ± 34
GLR-klUCB(Local)	$\mathbf{99 \pm 32}$	$\mathbf{128 \pm 42}$	149 ± 34
GLR-klUCB(Global)	128 ± 32	185 ± 47	152 ± 32

Table 6: Mean regret $\pm$ 1 std-dev. Problem 4 use $K = 3$ arms, and a first long stationary sequence. Problem 5 use $K = 5$, $T = 20000$ and is much harder with $\Upsilon = 82$ breakpoints and $C = 179$ changes.

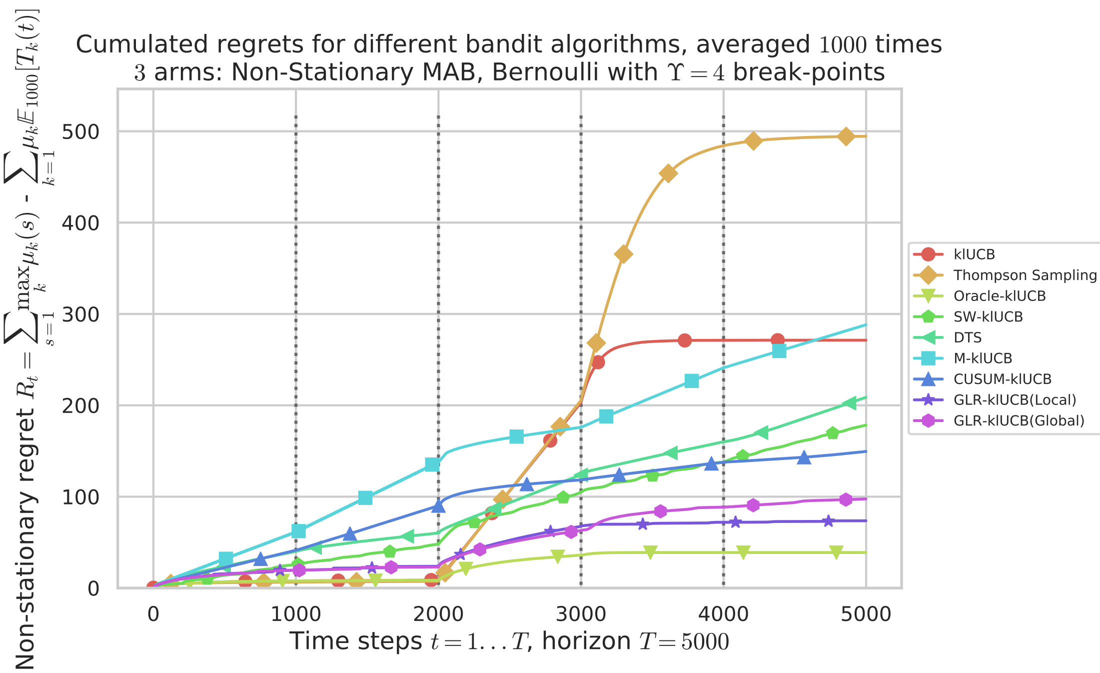


Figure 4: Mean regret as a function of time, R_t ($1 \leq t \leq T = 5000$) for problem 1.

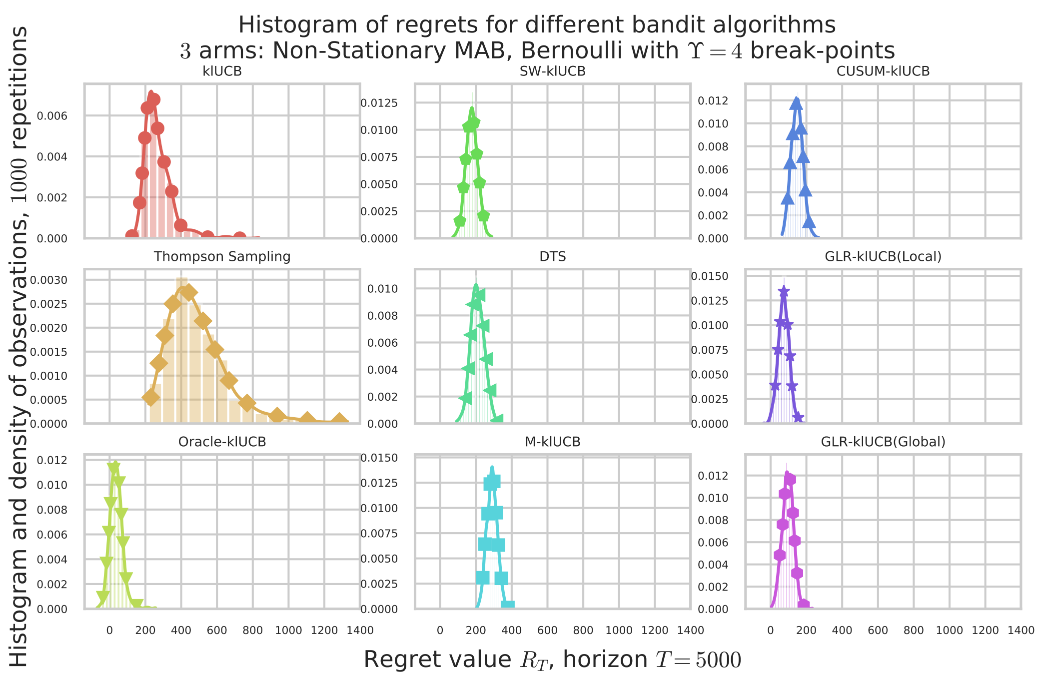


Figure 5: Histograms of the distributions of regret R_T ($T = 5000$) for problem 1.

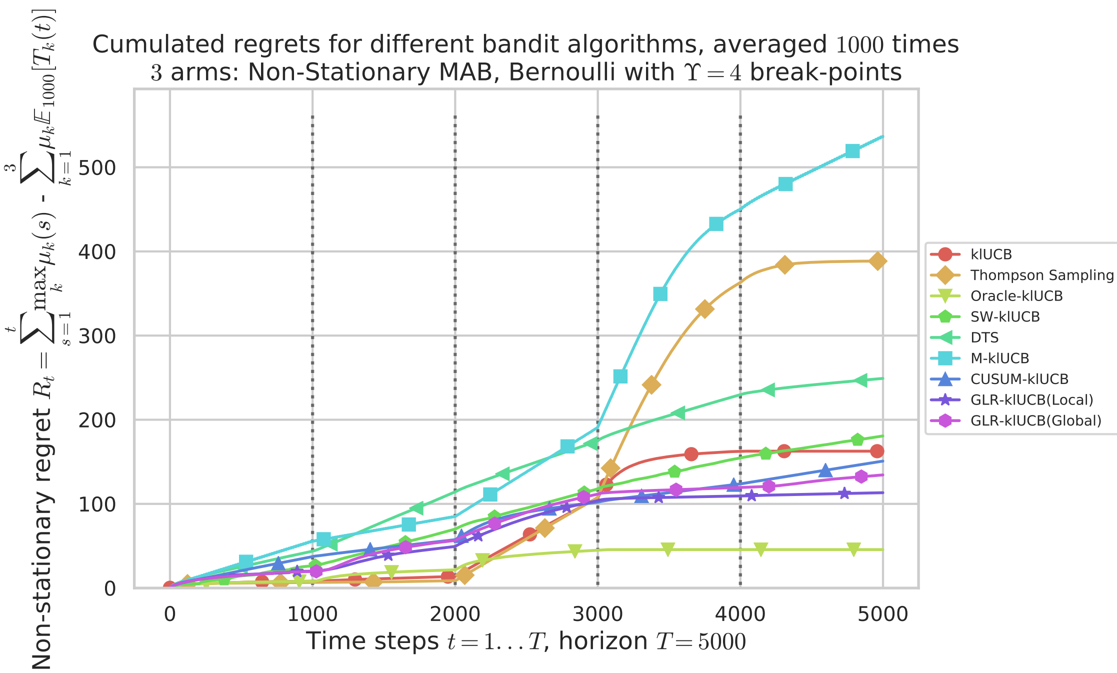


Figure 6: Mean regret as a function of time, R_t ($1 \leq t \leq T = 5000$) for problem 2.

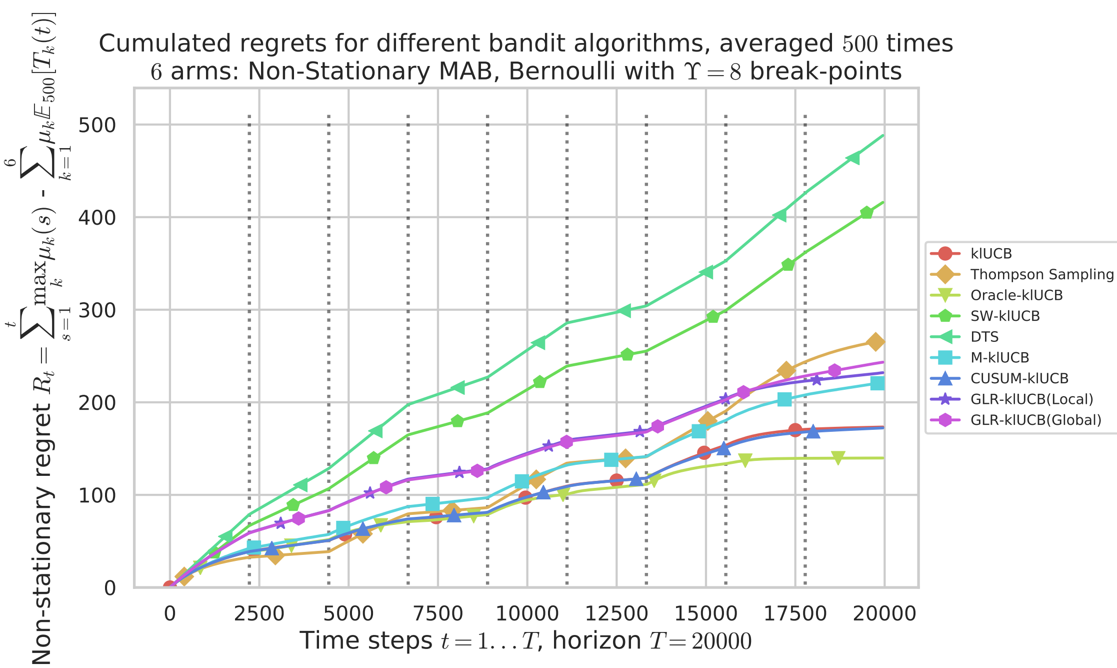


Figure 7: Mean regret as a function of time, R_t ($1 \leq t \leq T = 20000$) for problem 3.

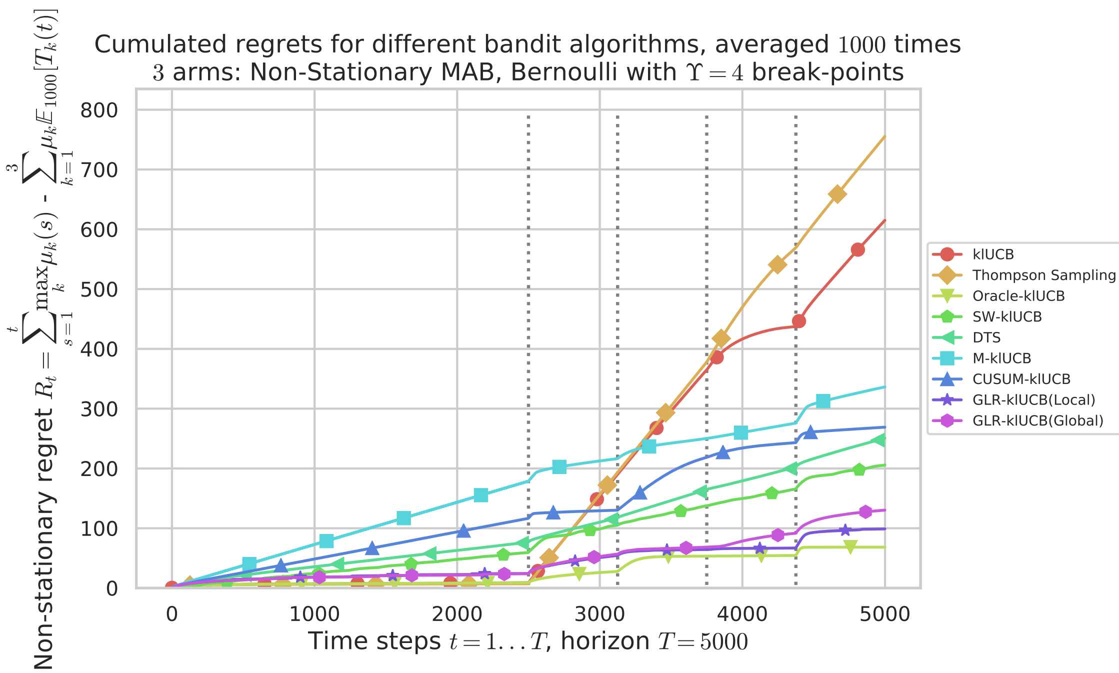


Figure 8: Mean regret as a function of time, R_t ($1 \leq t \leq T = 5000$) for problem 4.

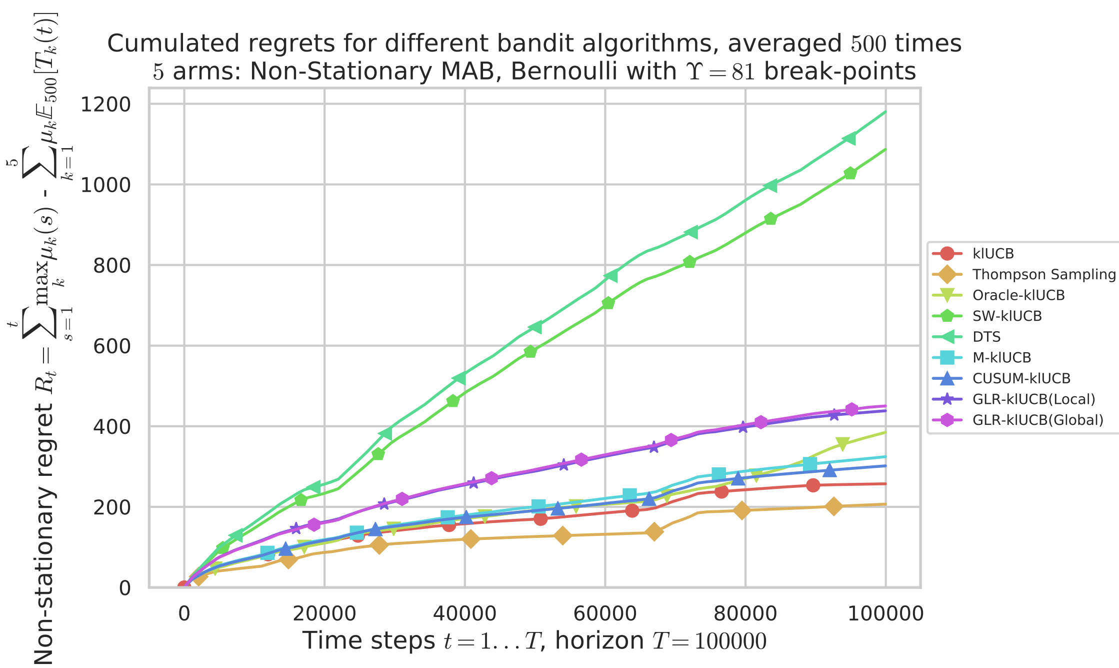


Figure 9: Mean regret as a function of time, R_t ($1 \leq t \leq T = 100000$) for problem 5.