Efficient Change-Point Detection for Tackling Piecewise-Stationary Bandits

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, Julien Seznec

Journal of Machine Learning Research 23 (2022) 1-41 Submitted 12/20; Revised 1/22; Published 3/22

Efficient Change-Point Detection for Tackling Piecewise-Stationary Bandits

Lilian Besson
ENS Rennes, IRISA, Inria Rennes, France
LILIAN.BESSON@ENS-RENNES.FR

Emilie Kaufmann
Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
EMILIE.KAUFMANN@UNIV-LILLE.FR

Odalric-Ambrym Maillard
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
ODALRIC.MAILLARD@INRIA.FR

Julien Seznec
Inria and Lelivrescolaire.fr *Éditions*
JULIEN.SEZNEC@INRIA.FR

Editor: Alexandre Proutiere

Abstract

We introduce GLR-klUCB, a novel algorithm for the piecewise *i.i.d.* non-stationary bandit problem with bounded rewards. This algorithm 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. Unlike previous non-stationary bandit algorithms using a change-point detector, GLR-klUCB does not need to be calibrated based on prior knowledge on the arms' means. We prove that this algorithm can attain a $\mathcal{O}(\sqrt{TA\Upsilon_T \ln(T)})$ regret in T rounds on some "easy" instances in which there is sufficient delay between two change-points, where A is the number of arms and Υ_T the number of change-points, *without prior knowledge of* Υ_T. In contrast with recently proposed algorithms that are agnostic to Υ_T, we perform a numerical study showing that GLR-klUCB is also very efficient in practice, beyond easy instances.

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 $A_t \in \{1, \dots, A\}$ or "arm" among a set of A arms (Robbins, 1952; Lattimore and Szepesvári, 2019). In the most standard version of the stochastic bandit model, each arm a is associated with an *i.i.d.* sequence of rewards $(X_{a,t})$ that follow some distribution of mean μ_a. Upon selecting arm A_t, the agent receives the reward $X_{A_t,t}$. Her goal is to design a sequential arm selection strategy that maximizes the expected sum of these rewards, or, equivalently, that minimizes regret, defined as the difference between the total sum of rewards of an oracle strategy always selecting the arm with largest mean and that of her strategy.

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 other applications, such as 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)). In all these applications, the assumption that the arms distributions do not evolve over time is often violated: patients adapt to medical treatments, new devices can enter or leave the radio network, hence impacting the availability of radio channels, and the popularity of items is subject to trends. This aroused interest in how to take non-stationary aspects into account within a multi-armed bandit model.

As a possible way to cope with non-stationarity, the piecewise stationary MAB was introduced by Kocsis and Szepesvári (2006). In this model, the (random) reward of arm a at round t has some mean $\mu_a(t)$ and the regret is measured with respect to the current best arm $a_t^\star = \arg\max_a \mu_a(t)$. It is furthermore assumed that there are relatively few breakpoints between which the $\mu_a(t)$ remain constant for all arms a. Alternative models in which the mean reward function $\mu_a(t)$ can slowly vary in every time step have also been considered in the literature (see, e.g. Wei and Srivastava (2018)) and may also be useful for the above mentioned applications. Still, a piecewise stationary model can be a relevant approximation when there are a few important changes whose magnitude is much larger than the typical variation between consecutive time steps. For example in cognitive radio, an important drop in performance could be caused by a new set of radio devices being deployed on a subset of the channels.

Despite many approaches already proposed for minimizing regret under the piecewise stationary MAB model (see Section 2), research on this topic has been very active in the last years, notably in two different directions. The first is the design of a good combination of a bandit algorithm and a change-point detector (CPD) supported by regret guarantees and enjoying good empirical performance (Liu et al., 2018; Cao et al., 2019). These algorithms share with many others the downside of having to know the number of breakpoints Υ_T to guarantee state-of-the-art regret. The second direction proposes algorithms that achieve optimal regret without the knowledge of Υ_T (Auer et al., 2019b; Chen et al., 2019), but without an emphasis on actual practical performance (yet).

In this paper, we propose the first algorithm based on a change-point detector that is very efficient in practice and *does not require the knowledge of* Υ_T to provably achieve optimal regret, at least on some "easy" instances, with few breakpoints of large enough magnitude. An interesting feature of our algorithm compared to other CPD-based algorithms is that it *does not require any prior knowledge on the arms means*. Like CUSUM (Liu et al., 2018) and $\mathrm{M\text{-}UCB}$ (Cao et al., 2019), our algorithm relies on combining a standard bandit algorithm with a change-point detector. For the bandit component, we propose the use of the klUCB (Cappé et al., 2013) which is known to outperform UCB (Auer et al., 2002a) used in previous works. For the change-point detector, we suggest using 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 the change-point detectors used in previous works, the Bernoulli GLRT *does not require a lower bound on the minimal amount of change to detect*, which leads to a bandit algorithm which is agnostic to the arms' means. In contrast, both CUSUM and $\mathrm{M\text{-}UCB}$ require the knowledge of the smallest magnitude of a change in the arm's mean.

In this work we jointly investigate two versions of GLR-klUCB, one using *global restarts* (resetting the history of all arms once a change-point is detected on one of them) and one using *local restarts* (resetting the history of an arm each time a change-point is detected on that arm). We prove that GLR-klUCB based on global restart achieves a $\mathcal{O}(\sqrt{TA\Upsilon_T \ln(T)}/(\Delta^{\text{change}})^2)$ regret where Δ^{change} is the smallest magnitude of a breakpoint. If all breakpoints have a large magnitude, this $\mathcal{O}(\sqrt{TA\Upsilon_T \ln(T)})$ regret is matching the lower bound of Seznec et al. (2020) up to a $\sqrt{\ln(T)}$ factor. Following a similar analysis, we prove slightly weaker results for the version based on local restart. Numerical simulations in Section 6 reveal that these two versions are both competitive in practice with state-of-the-art algorithms.

To summarize, our contributions are the following: (1) A non-asymptotic analysis of the Bernoulli-GLR change-point detector. (2) A new bandit algorithm for the piecewise stationary setting based on this test that needs no prior knowledge on the number of change-points and no information on the arms means to attain near-optimal regret. (3) An extensive numerical study illustrating the good performance of two versions of GLR-klUCB compared to other algorithms with state-of-the-art regret.

Outline The paper is structured as follows. We introduce the model and review related works in Section 2. In Section 3, we present some properties of the Bernoulli-GLR change-point detector. We introduce the two variants of GLR-klUCB in Section 4. In Section 5 we present regret upper bounds for GLR-klUCB for Global Restart and sketch our regret analysis. Numerical experiments are presented in Section 6.

2. Setup and Related Work

A piecewise stationary bandit model is characterized by a stream of (random) rewards $(X_{a,t})_{t\in\mathbb{N}^\star}$ associated to each arm $a \in \{1, \dots, A\}$. We assume that the rewards are bounded in a known range, and without loss of generality we assume that $X_{a,t} \in [0, 1]$. We denote by $\mu_a(t) := \mathbb{E}[X_{a,t}]$ the mean reward of arm a at round t. At each round t, a decision maker has to select an arm $A_t \in \{1, \dots, A\}$, based on past observation and receives the corresponding reward $r(t) = X_{A_t,t}$. At time t, we denote by $a_t^\star$ an arm with maximal expected reward, *i.e.*, $\mu_{a_t^\star}(t) = \max_a \mu_a(t)$, called an optimal arm.

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 (dynamic) regret, the difference between the expected reward obtained by an oracle policy playing an optimal arm $a_t^\star$ at time t, and that of the policy π:

$$R_T^\pi = \mathbb{E}\left[\sum_{t=1}^{T} \left(\mu_{a_t^\star}(t) - \mu_{A_t}(t)\right)\right].$$

In the piecewise *i.i.d.* model, we furthermore assume that there is a (relatively small) number of *breakpoints*, denoted by $\Upsilon_T := \sum_{t=1}^{T-1} \mathbb{1}\left(\exists a \in \{1, \dots, A\} : \mu_t(a) \neq \mu_{t+1}(a)\right)$. We define the k-th breakpoint by $\tau^{(k)} = \inf\{t > \tau^{(k-1)} : \exists a : \mu_a(t) \neq \mu_a(t+1)\}$ with $\tau^{(0)} = 1$. Hence for $t \in [\tau^{(k)} + 1, \tau^{(k+1)}]$, the rewards $(X_{a,t})$ associated to all arms are *i.i.d.*, with mean denoted by $\mu_a^{(k)}$. The magnitude of a breakpoint k is defined as $\Delta^{c,(k)} =: \max_{a=1,\dots,A} \left|\mu_a^{(k)} - \mu_a^{(k-1)}\right|$ and we let $\Delta^{\text{change}} =: \min_{k=1,\dots,\Upsilon_T} \Delta^{c,(k)}$.

Breakpoint versus change-point Note than when a breakpoint occurs, we do not assume that all the arms means change, but that there exists an arm which experiences a *change-point*, i.e. whose

mean satisfies $\mu_a(t) \neq \mu_a(t+1)$. Depending on the application, many scenarios can be meaningful: changes occurring for all arms simultaneously (due to some exogenous event), or only a few arms that experience a change-point in each breakpoint. Denoting by C_T the total number of change-points before horizon T, it holds that $C_T \in \{\Upsilon_T, \dots, A\Upsilon_T\}$.

2.1 An Adversarial View on Non-Stationary Bandits

A natural way to cope with non-stationary is to model the decision making problem as an adversarial bandit problem (Auer et al., 2002b), under which the rewards are arbitrarily generated. For adversarial environments, the most studied performance measure is the pseudo-regret, which compares the accumulated reward of a given strategy with that of the best fixed-arm policy. However in some changing environments it is more natural to measure regret against the best sequence of actions. Auer et al. (2002b) propose the Exp3.S algorithm, that achieves a regret of $\mathcal{O}(\sqrt{A\Upsilon_T T \ln(T)})$ against the best sequence of actions with $\Upsilon_T - 1$ switches. This regret rate matches the corresponding lower bound. Exp3.S is simple to implement and run with time and space complexity $\mathcal{O}(A)$ but requires the knowledge of T and Υ_T to reach near-minimax optimal regret rate.

When the piecewise i.i.d. assumption holds (with Υ_T stationary part), the best sequence of actions with $\Upsilon_T - 1$ switches corresponds to the optimal oracle policy. The minimax optimal rate against piecewise i.i.d. rewards sequences is also $\mathcal{O}(\sqrt{A\Upsilon_T T})$. It is similar to the fixed-arm case where the adversarial pseudo-regret rate and the minimax stochastic rate are the same ($\mathcal{O}(\sqrt{AT})$, Audibert and Bubeck (2010)). However, in the fixed-arm setup, the stochastic stationary assumption allows a problem-dependent analysis: some algorithms (e.g. UCB or Thomson Sampling) suffer $\mathcal{O}(\ln T/\Delta_i)$ regret on each arm i with a reward gap of Δ_i compared to the best arm. When Δ_i is large enough, this problem-dependent guarantee is much better than the $\mathcal{O}(\sqrt{T})$ minimax rate. Unfortunately, in the piecewise i.i.d. setup, Garivier and Moulines (2011) show that any algorithm whose regret is $R_T(\boldsymbol{\mu})$ on a stationary bandit instance $\boldsymbol{\mu}$ is such that there exists a piecewise stationary instance $\boldsymbol{\mu}'$ with at most two breakpoints such that $R_T(\boldsymbol{\mu}') \geq cT/R_T(\boldsymbol{\mu})$, for some absolute constant c. In particular, this implies that an algorithm that attains $\mathcal{O}(\sqrt{T})$ regret for any piecewise stationary bandit model has no hope to reach $\mathcal{O}(\ln(T))$ regret on easy instances. The intuition behind this result is that if an algorithm achieves very low regret on a specific problem then it has to pull suboptimal arms very scarcely. By doing so, it is unable to perform well on a similar problem where the identified suboptimal arms' surreptitiously increase to become optimal. Therefore, it is important to pull every arm often enough (e.g. every $\mathcal{O}(\sqrt{AT/\Upsilon_T})$ rounds) even when one is clearly underperforming.

Nevertheless, the piecewise i.i.d. bandit problem remained actively studied since the seminal paper of Auer et al. (2002b). The outcome of this line of work is threefold. First, designing strategies leveraging tools from the stochastic MAB can greatly improve the empirical performance compared to adversarial algorithms like Exp3.S. Second, we would like to build strategies that are near-optimal without the knowledge of Υ_T[1] (unlike Exp3.S). Third, it is possible to further restrain the setup to make the problem-dependent analysis possible by forbidding the aforementioned surreptitious increase of one arm. For instance, Mukherjee and Maillard (2019) consider the "global change" setup in which all the arms change significantly when a breakpoint occurs. Seznec et al. (2020) consider

1. In the adversarial setup, Υ_T appears in the definition of the pseudo-regret, hence it is quite natural that the learner knows this parameter. In the piecewise i.i.d. setup, the regret is against the optimal oracle policy which is defined independently of Υ_T.

the rotting setup where the arms cannot increase. In both cases, the authors proved a logarithmic problem-dependent upper bound on the regret of their algorithms.

In this paper, we bring theoretical and empirical contributions to the two first points. We also discuss a possible adaptation of GLR-klUCB that may recover logarithmic regret in the easier setups of Mukherjee and Maillard (2019); Seznec et al. (2020).

2.2 Algorithms Exploiting the Stochastic Assumption

The piecewise 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. Most approaches exploiting the stochastic assumption combine a bandit algorithm with a mechanism to forget old rewards. We make a distinction between *passively adaptive strategies*, which use a fixed forgetting mechanism, and *actively adaptive strategies*, for which this mechanism is also data-dependent.

Passively Adaptive Strategies A simple mechanism to forget the past consists in either discounting rewards (multiplying past reward by γ^n where n is the time elapsed since that reward was collected, for a discount factor $\gamma \in (0, 1)$), or using a sliding window (only the rewards gathered in the τ last rounds 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 $\mathcal{O}(A\sqrt{\Upsilon_T T}\ln(T))$ regret bound, if $\gamma = 1 - \sqrt{\Upsilon_T/T}/4$. The same authors proposed the Sliding-Window UCB (SW-UCB) and prove a $\mathcal{O}(A\sqrt{\Upsilon_T T \ln(T)})$ 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 on the reported experiments with $\gamma = 0.75$, but no theoretical guarantees are given for this particular tuning. The RExp3 algorithm (Besbes et al., 2014) is another passively adaptive strategy that is based on (non-adaptive) restarts of the Exp3 algorithm (Auer et al., 2002b). RExp3 is analyzed in terms of a different measure of interest, the total variation budget V_T which satisfies $\Delta^{\text{change}}\Upsilon_T \leq V_T \leq \Upsilon_T$. RExp3 is proved to have a $\mathcal{O}((A\ln A)^{1/3}V_T^{1/3}T^{2/3})$ regret, which translates to a sub-optimal rate in our setting.

Actively Adaptive Strategies 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 does not apply to our setting as it takes into account side observations. Another line of research on actively adaptive algorithms uses a Bayesian point of view, where the process of change-point occurrences is modeled and tracked using Bayesian updates. 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 et al. (2017). Since none of these algorithms have theoretical guarantees and they are designed for a different setup, we do not include them in our experiments.

Our closest competitors rather use frequentist CPD algorithms 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 guarantees 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 sub-optimal arm is detected to become optimal and it achieves a (sub-optimal)

$\mathcal{O}(\Upsilon_T A\sqrt{T\ln(T)})$ regret (yet without the knowledge of Υ_T). More recently, CUSUM-UCB (Liu et al., 2018) and Monitored UCB (M-UCB, Cao et al. (2019)) have achieved $\mathcal{O}(\sqrt{\Upsilon_T AT\ln(T)})$ regret, when Υ_T is known.

CUSUM-UCB is based on a variant of a 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. 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 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. 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 them.

On a stationary batch, a UCB index algorithm tends to pull each arm at a logarithmic rate asymptotically. According to the aforementioned Garivier and Moulines (2011)'s lower bound, it is not enough to shield against increases of the suboptimal arms' values. Thus, CPD-based algorithms usually rely on additional *forced exploration*: each arm is pulled regularly either according to a constant probability of uniform exploration (Liu et al., 2018) or according to a deterministic scheme (Cao et al., 2019). To avoid linear regret, the total budget dedicated to this forced exploration is tuned with the knowledge of T and Υ_T (e.g. $\mathcal{O}\left(\sqrt{A\Upsilon_T T}\right)$). Mukherjee and Maillard (2019) suggest canceling the forced exploration when all the arms change at the same rounds. Indeed, in that case, we can aim to detect the changes on any arms' sequences and then restart all the arms' indexes. Similarly, Seznec et al. (2020) do not use forced exploration and study the Rotting Adaptive Window UCB (RAW-UCB) - a UCB index policy with an adaptive window designed for non-increasing sequences of rewards. Both these algorithms can get logarithmic regret on some problem instances and, therefore, cannot be minimax optimal for the general piecewise i.i.d bandit problem, which is our focus in this paper.

2.3 On the Knowledge of the Number of Breakpoints

All algorithms mentioned above for the general piecewise stationary bandit problem require some tuning that should depend on Υ_T to attain state-of-the-art $\mathcal{O}(\sqrt{\Upsilon_T AT\ln(T)})$ regret. Two algorithms achieving this regret without the knowledge of Υ_T were recently proposed: Ada-ILTCB$^+$(Chen et al., 2019) and AdSwitch (Auer et al., 2019b), that also rely on detecting non-stationarities (Auer et al., 2019a). While the former is tailored for the more general adversarial and contextual setting, the latter is specifically proposed for the piecewise i.i.d. model.

AdSwitch is an elimination strategy based on confidence interval (like Improved UCB (Auer and Ortner, 2010)) with global restarts when a change-point is detected on one arm. AdSwitch performs an adaptive forced exploration scheme on the eliminated arms which adds two main components to the aforementioned uniform random exploration. First, AdSwitch uses a counter l (initialized at 1) for the number of detected changes by the CPD subroutine as a proxy for Υ_T to tune the random exploration probability on each eliminated arms at $\mathcal{O}\left(\sqrt{l/KT}\right)$. Second, AdSwitch also selects at random a change size Δ on a geometric grid, with a probability proportional to Δ. The arm is then pulled $\mathcal{O}\left(1/\Delta^2\right)$ consecutive pulls to check if there is a change of size Δ. The consecutive sampling is particularly helpful theoretically to analyze the algorithm when the CPD misses some breakpoints.

However, unlike in our work, the underlying change-point detectors used in AdSwitch have not been optimized for efficiency or tractability[2]. Neither (Chen et al., 2019) nor (Auer et al., 2019b) report simulation to assess the empirical or numerical efficiency of their algorithms. In this paper, we include (a tweaked, tractable version of) AdSwitch in our experiments for short horizons.

An alternative idea to adapt to Υ_T is the "Bandit over Bandit" approach of Cheung et al. (2019), which uses an exponential weights algorithm for expert aggregation on top of several copies of Sliding-Windows UCB with different (fixed) window size. Yet this approach does not yield optimal 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); Jie and Gupta (2000); Wu (2007))). In this article, we are interested in detecting changes on the mean of a probability distribution with bounded support. Assume 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* $\tau \in \mathbb{N}^\star$ 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 have been used for a very long time (see, e.g. Wilks (1938)) and were for instance studied for change-point detection by Siegmund and Venkatraman (1995). Exploiting the fact that bounded distributions are $(1/4)$-sub-Gaussian (*i.e.*, have a moment generating function dominated by that of a Gaussian with the same mean and variance $1/4$), the (Gaussian) GLRT, recently studied in depth by Maillard (2019), can be used 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 $\mathcal{H}_1 : (\exists \mu_0 \neq \mu_1, \tau \in \mathbb{N}^\star : \ X_1, \dots, X_\tau \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_0)$ and $X_{\tau+1}, X_{\tau+2}, \dots \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_1))$. The (log)-Generalized Likelihood Ratio statistic for this test is defined by

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

where $\ell(X_1, \dots, X_n; \mu_0)$ and $\ell(X_1, \dots, X_n; \mu_0, \mu_1, \tau)$ denote the likelihood 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 $\mathcal{H}_0$.

2. Indeed, at each time step t, the test employed by AdSwitch requires $\Theta(At^3)$ operations, resulting in a very expensive $\Theta(AT^4)$ time complexity when compared to $\Theta(AT)$ for simple algorithms like UCB and $\Theta(AT^2)$ for other adaptive approaches based on scan statistics like GLR-klUCB.

Using the notation $S_{s:t} = \sum_{\ell=s}^{t} X_s$ for $s < t \leq n$, one has

$$\begin{aligned} \ell(X_1, \ldots, X_n; \mu_0) &= \mu_0^{S_{1:n}} (1-\mu_0)^{n-S_{1:n}} \\ \ell(X_1, \ldots, X_n; \mu_0, \mu_1, \tau) &= \mu_0^{S_{1:\tau}} (1-\mu_0)^{\tau - S_{1:\tau}} \mu_1^{S_{\tau+1:n}} (1-\mu_1)^{n-\tau-S_{\tau+1:n}}. \end{aligned}$$

and the GLR statistic can be expressed using the binary relative entropy kl, defined as

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

Indeed, one can show that $\mathrm{GLR}(n) = \sup_{s \in [1,n]} Z_{s,n}$ where $Z_{s,n} = s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right)$ and 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 $\beta(n, \delta)$ is*

$$\hat{\tau}_\delta := \inf\left\{ n \in \mathbb{N}^\star : \sup_{s \in [1,n]} \left[s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right) \right] \geq \beta(n, \delta) \right\}. \tag{2}$$

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 (2) 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 (2) 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 $(1/4)$-sub-Gaussian 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 than the Gaussian GLR based on the quadratic divergence $2(x-y)^2$ (assuming that the two statistics were compared to the same threshold).

GLR versus confidence-based CPD An alternative to the GLR also based on scan statistics, used by Mukherjee and Maillard (2019) consists in building individual confidence intervals for the mean in each segment, of the form

$$\left[\hat{\mu}_{1:s} \pm \sqrt{\frac{\tilde{\beta}(s, \delta)}{2s}} \right] \text{ and } \left[\hat{\mu}_{s+1:n} \pm \sqrt{\frac{\tilde{\beta}(n-s, \delta)}{2(n-s)}} \right]$$

and report that there is a change-point if there exists s such that these confidence interval are disjoint, i.e.

$$\hat{\tau}'_\delta = \inf\left\{ n \in \mathbb{N}^\star : \exists s \in [1, n], \left| \hat{\mu}_{1:s} - \hat{\mu}_{s+1:n} \right| > \sqrt{\frac{\tilde{\beta}(s, \delta)}{2s}} + \sqrt{\frac{\tilde{\beta}(n-s, \delta)}{2(n-s)}} \right\}.$$

By measuring distances with the appropriate KL divergence function, the Bernoulli GLR test better exploits the geometry of (sub-)Bernoulli distributions.

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 introduce the function $\mathcal{T}$, originally introduced by Kaufmann and Koolen (2021),

$$\mathcal{T}(x) \;:=\; 2\tilde{h}\left(\frac{h^{-1}(1+x)+\ln(\pi^2/3)}{2}\right) \tag{3}$$

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 (2021), where tight upper bounds 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 is given in Appendix B.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*

$$\beta(n,\delta) = 2\mathcal{T}\left(\frac{\ln(3n\sqrt{n}/\delta)}{2}\right) + 6\ln(1+\ln(n)). \tag{4}$$

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" that all the change-points are detected within a reasonable delay (Lemma 8 and 15).

3.3 Practical considerations

Lemma 2 provides the first control of false alarm for the Bernoulli GLR employed for bounded distributions. However, the threshold (4) 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 (4) is a bit conservative, and in practice we recommend to keep only the leading term and use $\beta(n,\delta) = \ln(n\sqrt{n}/\delta)$.

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

$$\tilde{\tau}_\delta = \inf\left\{n \in \mathcal{N} : \sup_{s \in \mathcal{S}_n} Z_{s,n} \geq \beta(n,\delta)\right\}, \tag{5}$$

for any strict subsets $\mathcal{N} \subseteq \mathbb{N}$ and $\mathcal{S}_n \subset \{1, \dots, n\}$. 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. Yet, in our experiments we will use $\mathcal{S}_n = \{k \leq n : k \bmod 5 = 0\}$ in the Bernoulli GLR and still get improved regret compared to state-of-the-art algorithms.

4. The GLR-klUCB Algorithm

GLR-klUCB (Algorithm 1) 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. It also needs a third ingredient: some extra exploration to ensure each arm is sampled enough and changes can also be detected on arms currently under-sampled by klUCB. This forced exploration is parameterized by a sequence $\alpha = (\alpha_k)_{k \in \mathbb{N}}$ of exploration frequencies $\alpha_k \in (0, 1)$. GLR-klUCB can be used in any bandit model with bounded rewards, and is expected to be very efficient for Bernoulli distributions.

Algorithm 1: GLR-klUCB (**Local** or **Global** restarts)

Input: $(\alpha_k)_{k \in \mathbb{N}^\star}$ (sequence of exploration probabilities),
$\delta \in (0, 1)$ (maximum error probability for the test);
Input: *Option*: **Local** or **Global** restart;
1 **Initialization:** $\forall a \in \{1, \dots, A\}$, $\tau_a \leftarrow 0$ (last restart) and $n_a \leftarrow 0$ (number of selections since last restart)
2 $k \leftarrow 1$ (number of episodes)
3 **for** $t = 1, 2, \dots, T$ **do**
4 **if** $t \mod \left\lceil \frac{A}{\alpha_k} \right\rceil \in \{1, \dots, A\}$ **then**
5 $A_t \leftarrow t \mod \left\lceil \frac{A}{\alpha_k} \right\rceil$
6 **else**
7 $A_t \leftarrow \arg\max_{a \in \{1, \dots, A\}} \mathrm{UCB}_a(t)$ as defined in (6)
8 **end**
9 Play arm A_t and receive the reward $X_{A_t,t}$: $n_{A_t} \leftarrow n_{A_t} + 1$; $Y_{A_t, n_{A_t}} \leftarrow X_{A_t,t}$
10 **if** $\mathrm{GLR}_\delta(Y_{A_t,1}, \dots, Y_{A_t,n_{A_t}}) = \mathrm{True}$ **then**
11 **if** Local *restart* **then**
12 $\tau_{A_t} \leftarrow t$ and $n_{A_t} \leftarrow 0$ and $k \leftarrow k + 1$
13 **else**
14 $\forall a \in \{1, \dots, A\}, \tau_a \leftarrow t$ and $n_a \leftarrow 0$ and $k \leftarrow k + 1$
15 **end**
16 **end**

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, $\mathrm{GLR}_\delta(Y_1, \dots, Y_n) = \mathrm{True}$ if and only if the GLR statistic associated to those n samples,

$$\sup_{1 \leq s \leq n} \left[s \times \mathrm{kl}(\hat{Y}_{1:s}, \hat{Y}_{1:n}) + (n - s) \times \mathrm{kl}(\hat{Y}_{s+1:n}, \hat{Y}_{1:n}) \right],$$

is larger than the threshold $\beta(n, \delta)$ defined in (4), or $\beta(n, \delta) = \ln(n^{3/2}/\delta)$, as recommended in Section 3.3. Each restart (on any arm) triggers a new episode and we denote by k_t the number of episodes started after t samples (i.e. the index of the on-going episode at time t).

Letting $\tau_a(t)$ denote the instant of the last restart that happened for arm a before time t, $n_a(t) = \sum_{s=\tau_a(t)+1}^{t} \mathbb{1}(A_s = a)$ the number of selections of arm a and $\hat{\mu}_a(t) = (1/n_a(t)) \sum_{s=\tau_a(t)+1}^{t} X_{a,s}\mathbb{1}(A_s = a)$ the empirical mean (if $n_a(t) \neq 0$), the index used by the algorithm is defined as

$$\mathrm{UCB}_a(t) := \max\big\{q : n_a(t) \times \mathrm{kl}\,(\hat{\mu}_a(t), q) \leq f(t - \tau_a(t))\big\}. \tag{6}$$

Algorithm 1 presents two variants of GLR-klUCB, one using *local restarts* (line 11), and one using *global restarts* (line 13). Under local restarts, in the general case the times $\tau_a(t)$ are not equal for all arms, hence the index policy associated to (6) 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 restarts, the UCB index are defined differently[3]: $f(t - \tau_a(t))$ is replaced by $f(n_t)$ with $n_t = \sum_{a=1}^{A} n_a(t)$.

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 (Liu et al., 2018). A consequence of this forced exploration is given below in Proposition 3, which is proved in Appendix A.

Proposition 3 *Let s, t be two time instants between two consecutive restarts on arm a (i.e. $\tau_a(t) < s < t$). Then it holds that $n_a(t) - n_a(s) \geq \left\lfloor \frac{\alpha_{k_t}}{A}(t - s) \right\rfloor$, with k_t the number of episodes before round t.*

While M-UCB uses a constant exploration sequence such that $\alpha_k \simeq \sqrt{\Upsilon_T \ln(T)/T}$, we advocate in the next section the use of an *increasing exploration sequence* such that $\alpha_k \simeq \sqrt{k \ln(T)/T}$, which yields near optimal regret without the knowledge of Υ_T.

5. Regret Analysis

In this section, we prove regret bounds for GLR-klUCB using Global Restart. Our results for GLR-klUCB with Local Restart are a bit weaker and are deferred to Appendix D.

5.1 Regret Upper Bounds

Recall that $\tau^{(k)}$ denotes the position of the k-th breakpoint and let $\mu_a^{(k)}$ be the mean of arm a on the segment $\{\tau^{(k-1)} + 1, \ldots, \tau^{(k)}\}$. We also introduce $k^\star = \arg\max_a \mu_a^{(k)}$, the sub-optimality gap $\Delta_a^{(k)} = \mu_{k^\star}^{(k)} - \mu_a^{(k)}$ and the recall that the magnitude of breakpoint k is $\Delta^{c,(k)} = \max_{a=1,\ldots,A} |\mu_a^{(k)} - \mu_a^{(k-1)}| > 0$.

We first introduce the assumption under which we are able to upper bound the regret of GLR-klUCB. It requires that the distance between two consecutive breakpoints is large enough and how large depends on the magnitude of the largest change that happens at those two breakpoints. A similar assumption was made for the analysis of M-UCB (Cao et al. (2019), Assumption 1), whereas

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

the analysis of CUSUM-UCB requires a known lower bound on the magnitude of any change-point (see Liu et al. (2018), Assumption 2).

Assumption 4 *Define the delay $d^{(k)} = d^{(k)}(\alpha, \delta)$ as $d^{(0)}(\alpha, \delta) = 0$ and for $k > 0$,*

$$d^{(k)}(\alpha, \delta) = \left\lceil \frac{4A}{\alpha_k \left(\Delta^{c,(k)}\right)^2} \beta\left(\frac{3}{2}(\tau^{(k)} - \tau^{(k-1)}), \delta\right) + \frac{A}{\alpha_k} \right\rceil,$$

we assume that $\forall k \in \{1, \dots, \Upsilon_T\}, \tau^{(k)} - \tau^{(k-1)} \geq 2(d^{(k)} \vee d^{(k-1)})$.

Under Assumption 4, we provide in Theorem 5 a finite time problem-dependent regret upper bound. It features the parameters α and δ, the gaps $\Delta_a^{(k)}$ and KL-divergence terms $\mathrm{kl}(\mu_a^{(k)}, \mu_{k^\star}^{(k)})$ expressing the hardness of the stationary MAB problem between two breakpoints, and the detection delays $d^{(k)}(\alpha, \delta)$, which feature the gap $\Delta^{c,(k)}$ and express the hardness of the detection of each breakpoint.

Theorem 5 *For $\alpha = (\alpha_1, \alpha_2, \dots)$ an increasing exploration sequence and δ for which Assumption 4 is satisfied, the regret of* GLR-klUCB *with parameters α and δ based on* ***Global*** *Restart satisfies*

$$R_T \leq (A+1)\Upsilon_T \delta T + A\delta T + \alpha_{\Upsilon_T+1} T$$
$$+ \sum_{k=0}^{\Upsilon_T} \sum_{a:\Delta_a^{(k)}>0} \min\left\{ \Delta_a^{(k)}\left(\tau^{(k+1)} - \tau^{(k)}\right); \Delta_a^{(k)}\left[d^{(k)}(\alpha, \delta) + \frac{\ln\left(\tau^{(k+1)} - \tau^{(k)}\right)}{\mathrm{kl}\left(\mu_a^{(k)}, \mu_{k^\star}^{(k)}\right)} \right] + \mathcal{O}\left(\sqrt{\ln\left(\tau^{(k+1)} - \tau^{(k)}\right)}\right) \right\}.$$

We express below the scaling of this regret bound when the exploration sequence α and the parameter δ are carefully tuned using the knowledge of the horizon T, but *without the knowledge of the number of breakpoints* Υ_T. We express this scaling as a function of the smallest value of a sub-optimality gap on one of the stationary segments and the gap of the hardest breakpoint to detect, respectively defined as

$$\Delta^{\text{opt}} := \min_{k=1,\dots,\Upsilon_T} \min_{\{a:\Delta_a^{(k)}>0\}} \Delta_a^{(k)}$$
$$\text{and} \quad \Delta^{\text{change}} := \min_{k=1,\dots,\Upsilon_T} \Delta^{c,(k)} = \min_{k=1,\dots,\Upsilon_T} \max_{a=1,\dots,A} |\mu_a^{(k)} - \mu_a^{(k-1)}|.$$

Corollary 6 *For any $\alpha_0 \in \mathbb{R}^+$ and $\gamma \in (1/2, 1]$, choosing*

$$\alpha_k = \alpha_0 \sqrt{\frac{kA\ln(T)}{T}} \quad \textit{and} \quad \delta = \frac{1}{T^\gamma},$$

on problem instances satisfying the corresponding Assumption 4, the regret of GLR-klUCB *satisfies*

$$R_T = \mathcal{O}\left((1+\gamma) \frac{\sqrt{\Upsilon_T A T \ln(T)}}{\left(\Delta_{\text{change}}\right)^2} + \frac{(A-1)}{\Delta^{\text{opt}}} (\Upsilon_T + 1) \ln(T) \right),$$

and $R_T = \mathcal{O}\left(\frac{\sqrt{\Upsilon_T A T \ln(T)}}{\left(\Delta_{\text{change}}\right)^2} \right)$.

If Δ^{change} is viewed as a constant, our regret upper bound is matching the lower bound of Seznec et al. (2020) up to a $\sqrt{\ln(T)}$ factor. Hence we propose a tuning of GLR-klUCB which attains near-optimal regret without the knowledge of the number of breakpoints, on "easy" problems. In particular, a sufficient condition for Assumption 4 to hold for the parameter tuning suggested by Corollary 6 and used in our experiments ($\alpha_k = \sqrt{kA\ln(T)/T}$, $\delta = 1/\sqrt{T}$ and $\beta(n,\delta) = \ln(n^{3/2}/\delta)$) is that the length of the k-th stationary period, $D_k = \tau^{(k)} - \tau^{(k-1)}$ satisfies

$$D_k \geq \frac{8}{\Delta^2_{\text{change}}}\sqrt{\frac{AT}{k\ln(T)}}\left(\frac{1}{2}\ln(T) + \frac{3}{2}\ln\left(\frac{3}{2}D_k\right) + 1\right),$$

which holds in particular if $D_k \geq 24\sqrt{AT\ln(T)}/\Delta^2_{\text{change}}$. Under this constraint, observe that the number of breakpoints Υ_T can still be reasonably large, of order $\Delta^2_{\text{change}}\sqrt{T/(A\ln(T))}$.

As shown is Section 6, GLR-klUCB also performs well on problem instances for which Assumption 4 is violated, which was similarly observed by Cao et al. (2019) for M-UCB. Compared to other algorithms based on stationary bandit strategies combined with change-point detectors, GLR-klUCB is the only one that does not require a tuning based on Υ_T to attain the best possible regret. Indeed, it uses an increasing exploration sequence instead of a constant sequence, which allows the trick (8) in the proof of Corollary 6. If Υ_T is known, observe that one can also run GLR-klUCB with the constant exploration sequence $\alpha_k = \alpha = \alpha_0\sqrt{kA\ln(T)/T}$ and obtain the same regret as in Corollary 6. We tried the two alternatives in our experiments, and got similar performances. Hence, the use of an exploration sequence that is agnostic to Υ_T does not hinder the practical performance of GLR-klUCB.

Finally, there exist algorithms which attain near-optimal regret without the knowledge of Υ_T and with no $\left(\Delta^{\text{change}}\right)^{-2}$ multiplicative factor in their regret bound (Auer et al., 2019b; Chen et al., 2019). However, these algorithms are very conservative in order to make their analysis possible. For instance, AdSwitch is an elimination policy, which is often a poor choice in practice for regret minimization. In Section 6, we indeed show that GLR-klUCB greatly outperforms AdSwitch. We note that the dependency in $\left(\Delta^{\text{change}}\right)^{-2}$ is also present in the regret bound for other algorithms combining UCB-style algorithms and change-point detectors (Liu et al., 2018; Cao et al., 2019). It may come from a limitation of the current analysis of such algorithms, which require every breakpoint to be detected.

Regret of GLR-klUCB with Local changes As stated precisely in Appendix D, for GLR-klUCB with Local changes, we can prove a $\mathcal{O}(\sqrt{C_T AT\ln(T)}/\Delta^2)$ regret where

$$C_T = \sum_{t=0}^{T-1}\sum_{a=1}^{A}\mathbb{1}(\mu_a(t) \neq \mu_a(t+1))$$

is the total number of *change-points* and Δ is the minimal magnitude of any *change-point*.

This result is in general weaker than that of Corollary 6, as C_T can be (much) larger than Υ_T and Δ can be (much) smaller than Δ_{change}. However, Theorem 12 in Appendix D holds under Assumption 11, a counterpart of Assumption 4 giving a minimal spacing between two change-points depending on their magnitude. A sufficient condition for Assumption 11 to hold is that the minimal delay between two change-points on an arm is of order $\sqrt{AT\ln(T)}/\Delta^2$, which can be less stringent than the consequence of Assumption 4 on problem instances in which very few arms change in each breakpoint.

5.2 Proof of Corollary 6

With the choice $\delta = T^{-\gamma}$ and $\alpha_k = \alpha_0\sqrt{k\ln(T)A/T}$, Theorem 5 yields the following upper bound on the regret of GLR-klUCB:

$$(A+1)\Upsilon_T T^{1-\gamma} + AT^{-\gamma} + \alpha_0\sqrt{(\Upsilon_T+1)AT\ln(T)} + \sum_{k=1}^{\Upsilon_T}\frac{4A}{\alpha_k\left(\Delta^{c,(k)}\right)^2}\beta(\tfrac{3}{2}T, T^{-\gamma})$$
$$+\sum_{k=0}^{\Upsilon_T}\sum_{a:\mu_a^{(k)}\neq\mu_{k^\star}^{(k)}}\frac{\left(\mu_{k^\star}^{(k)}-\mu_a^{(k)}\right)}{\mathrm{kl}\left(\mu_a^{(k)},\mu_{k^\star}^{(k)}\right)}\ln(T) + \mathcal{O}\Big(\sqrt{\ln(T)}\Big)\ . \tag{7}$$

For $\gamma > 1/2$, the leading term in this expression is

$$\alpha_0\sqrt{(\Upsilon_T+1)AT\ln(T)} + \sum_{k=1}^{\Upsilon_T}\frac{4A}{\alpha_k\left(\Delta^{c,(k)}\right)^2}\beta(\tfrac{3}{2}T, T^{-\gamma}) + \sum_{k=0}^{\Upsilon_T}\sum_{a:\mu_a^{(k)}\neq\mu_{k^\star}^{(k)}}\frac{\left(\mu_{k^\star}^{(k)}-\mu_a^{(k)}\right)}{\mathrm{kl}\left(\mu_a^{(k)},\mu_{k^\star}^{(k)}\right)}\ln(T)\ .$$

Using that $\beta(n,\delta) \le C\ln(n/\delta)$ for some absolute constant C together with the fact that

$$\sum_{k=1}^{\Upsilon_T}\frac{1}{\alpha_k} = \frac{1}{\alpha_0}\sqrt{\frac{T}{A\ln(T)}}\sum_{k=1}^{\Upsilon_T}\frac{1}{\sqrt{k}} \le \frac{1}{\alpha_0}\sqrt{\frac{\Upsilon_T T}{A\ln(T)}} \tag{8}$$

yields the following control on the expected regret

$$R_T = \mathcal{O}\left((1+\gamma)\frac{\sqrt{\Upsilon_T AT\ln(T)}}{\left(\min_{k=1}^{\Upsilon_T}\Delta^{c,(k)}\right)^2} + \sum_{k=0}^{\Upsilon_T}\sum_{a:\mu_a^{(k)}\neq\mu_{k^\star}^{(k)}}\frac{\left(\mu_{k^\star}^{(k)}-\mu_a^{(k)}\right)}{\mathrm{kl}\left(\mu_a^{(k)},\mu_{k^\star}^{(k)}\right)}\ln(T)\right)\ .$$

The conclusion follows from Pinsker's inequality: $\mathrm{kl}(\mu_a^{(k)},\mu_{k^\star}^{(k)}) \ge 2\left(\Delta_a^{(k)}\right)^2$ and from lower bounding all sub-optimality gaps by Δ^{opt}.

Rather than using the problem-dependent complexity of the MAB problem on each stationary segment, using Theorem 5 and standard techniques one can also obtain the following "worse-case" upper bound:

$$(A+1)\Upsilon_T T^{1-\gamma} + AT^{-\gamma} + \alpha_0\sqrt{(\Upsilon_T+1)AT\ln(T)} + \sum_{k=1}^{\Upsilon_T}\frac{4A}{\alpha_k\left(\Delta^{c,(k)}\right)^2}\beta(\tfrac{3}{2}T, T^{-\gamma}) + C\sum_{k=0}^{\Upsilon_T}\sqrt{A\left(\tau^{(k+1)}-\tau^{(k)}\right)\ln(T)}\,.$$

for some constant C. One can further upper bound the last term using the Cauchy-Schwarz inequality:

$$\sum_{k=0}^{\Upsilon_T}\sqrt{A\left(\tau^{(k+1)}-\tau^{(k)}\right)\ln(T)} \le \sqrt{\sum_{k=0}^{\Upsilon_T}\left(\tau^{(k+1)}-\tau^{(k)}\right)}\sqrt{\sum_{k=0}^{\Upsilon_T}A\ln(T)} = \sqrt{(\Upsilon_T+1)AT\ln(T)}.$$

Using the same upper bound derived above for the first three terms yields

$$R_T = O\left((1+\gamma)\frac{\sqrt{\Upsilon_T AT\ln(T)}}{\left(\min_{k=1}^{\Upsilon_T}\Delta^{c,(k)}\right)^2} + \sqrt{(\Upsilon_T+1)AT\ln(T)}\right) = \mathcal{O}\left(\frac{\sqrt{\Upsilon_T AT\ln(T)}}{\left(\Delta^{\text{change}}\right)^2}\right).$$

5.3 Proof of Theorem 5

We first introduce some notation for the proof. Recall that $\tau^{(k)}$ denote the k-th breakpoint, we add the convention that $\tau^{(\Upsilon_T+1)} = T$. We denote by $\hat{\tau}^{(k)}$ the k-th breakpoint detected by GLR-klUCB.

Distinguishing the exploration steps and the steps in which GLR-klUCB uses the UCBs to select the next arm to play, one can upper bound the regret as

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(t \bmod \left\lceil \frac{A}{\alpha_{k_t}} \right\rceil \in [1, A]\right) + \sum_{t=1}^{T}(\mu_{a_t^\star}(t) - \mu_{A_t}(t))\mathbb{1}\big(\mathrm{UCB}_{A_t}(t-1) \geq \mathrm{UCB}_{a_t^\star}(t-1)\big)\right] \quad (9)$$

We now introduce some high-probability event in which all the breakpoints are detected within a reasonable delay. With $d^{(k)} = d^{(k)}(\alpha, \delta)$ in Assumption 4, we define

$$\mathcal{E}_T = \mathcal{E}_T(\alpha, \delta) := \left(\forall k \in \{1, \dots, \Upsilon_T\}, \hat{\tau}^{(k)} \in \left[\tau^{(k)} + 1, \tau^{(k)} + d^{(k)}\right], \hat{\tau}^{(\Upsilon_T+1)} > T\right). \quad (10)$$

Note that from Assumption 4, as the period between two changes are long enough, if $\mathcal{E}_T$ holds, then for all change k, one has $\tau^{(k)} \leq \hat{\tau}^{(k)} \leq \tau^{(k+1)}$ for all $k \in \{1, \dots, \Upsilon_T\}$. Also, when $\mathcal{E}_T$ holds, GLR-klUCB experiences exactly Υ_T restarts which permits to upper bound the exploration term in (9), using the convention that $\hat{\tau}^{(\Upsilon_T+1)} = T$:

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{1}\left(t \bmod \left\lceil \frac{A}{\alpha_{k_t}} \right\rceil \in \{1, \dots, A\}\right) &\leq \sum_{k=0}^{\Upsilon_T}\sum_{t=\hat{\tau}^{(k)}+1}^{\hat{\tau}^{(k+1)}}\mathbb{1}\left(t \bmod \left\lceil \frac{A}{\alpha_{k+1}} \right\rceil \in \{1, \dots, A\}\right) \\
&\leq \sum_{k=0}^{\Upsilon_T}\alpha_{k+1}\left(\hat{\tau}^{(k+1)} - \hat{\tau}^{(k)}\right) \leq \alpha_{\Upsilon_T+1}\sum_{k=0}^{\Upsilon_T}\left(\hat{\tau}^{(k+1)} - \hat{\tau}^{(k)}\right) \\
&= \alpha_{\Upsilon_T+1}T.
\end{aligned}$$

On $\mathcal{E}_T$, the second term in (9) can also be decomposed along the $\Upsilon_T + 1$ episodes experienced by the algorithm. Recalling that $k^\star$ denotes the optimal arm for $t \in \left[\tau^{(k)} + 1, \tau^{(k+1)}\right]$, one can write

$$R_T \leq T\mathbb{P}\left(\mathcal{E}_T^c\right) + \alpha_{\Upsilon_T+1}T + \sum_{k=0}^{\Upsilon_T}\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=\tau^{(k)}+1}^{\tau^{(k+1)}}\left(\mu_{k^\star}^{(k)} - \mu_{A_t}^{(k)}\right)\mathbb{1}(\mathrm{UCB}_{A_t}(t-1) \geq \mathrm{UCB}_{k^\star}(t-1))\right]. \quad (11)$$

The conclusion follows from the two lemmas stated below, whose proofs are given in Appendix C. The first one hinges on some elements of the analysis of the klUCB algorithm proposed by Cappé et al. (2013) whereas the second exploits the change-point detection mechanism.

Lemma 7 *With $\Delta_a^{(k)} = \mu_{k^\star}^{(k)} - \mu_a^{(k)}$, the following upper bound holds:*

$$(11) \leq \sum_{k=0}^{\Upsilon_T}\sum_{a:\Delta_a^{(k)}>0}\min\left\{\Delta_a^{(k)}\left(\tau^{(k+1)} - \tau^{(k)}\right); d^{(k)} + \frac{\Delta_a^{(k)}\ln\left(\tau^{(k+1)} - \tau^{(k)}\right)}{\mathrm{kl}\left(\mu_a^{(k)}, \mu_{k^\star}^{(k)}\right)} + \mathcal{O}\left(\sqrt{\ln\left(\tau^{(k+1)} - \tau^{(k)}\right)}\right)\right\}.$$

Lemma 8 *Under Assumption 4, it holds that $\mathbb{P}(\mathcal{E}_T^c(\alpha, \delta)) \leq \delta(A+1)\Upsilon_T + A\delta$.*

The tricky part in the analysis is the proof of Lemma 8, which crucially exploits Assumption 4, that we briefly sketch here (with a detailed proof in Appendix C.2). Introducing the event $\mathcal{C}^{(k)} =$

$\left\{\forall \ell \leq k, \hat{\tau}^{(\ell)} \in \left[\tau^{(\ell)}+1, \tau^{(\ell)}+d^{(\ell)}\right]\right\}$ that all the changes up to the k-th have been detected and using the convention $\tau^{(\Upsilon_T+1)} = T$, a union bound permits to upper bound $\mathbb{P}(\mathcal{E}_T^c)$ by the sum of two terms:

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

The event in (a) implies that the change-point detector associated with some arm a experiences a false alarm. The probability of such an event is upper bounded by Lemma 2 for a change-point detector run in isolation. Under the bandit algorithm, arm a's change-point detector is based on less than $t - \tau_a(t)$ samples, which makes a false alarm even less likely. We finally show that $(a) \leq A\delta$ (with union bound over the A arms).

Term (b) is related to the control of the detection delay, which is more involved 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 the detection delay for the k-th breakpoint defined in Assumption 4 is scaled by α_k. Using some elementary calculus and a concentration inequality given in Lemma 10, we finally prove that $(b) \leq \delta$.

6. Experimental Results

In this section, we report numerical simulations performed on synthetic data to compare the performance of GLR-klUCB against other state-of-the-art approaches. Experiments were performed with a library written in the Julia language which is available online.[4]

6.1 Algorithms and Parameters Tuning

We include two baselines: the klUCB algorithm (not designed for the non-stationary setting) and an algorithm that we call Oracle-klUCB, which knows the exact locations of the breakpoints, and restarts klUCB for all arms at those locations. Then, we include algorithms with state-of-the-art regret for the piecewise stationary MAB presented in Section 2. For a fair comparison, all algorithms that use UCB as a sub-routine were adapted to use klUCB instead, which yields better performance[5]. For all the algorithms, we performed the tuning recommended in the corresponding paper, using in particular the knowledge of the number of breakpoints Υ_T and the horizon T when needed. Only two algorithms do not require the knowledge of Υ_T: AdSwitch and GLR-klUCB.

We experiment with Exp3.S (with theoretically optimal tuning in Corollary 8.3 of Auer et al. (2002b)), and the two passively adaptive algorithms Discounted klUCB (D-klUCB) with discount factor $\gamma = 1 - \sqrt{\Upsilon_T/T}/4$ and Sliding-Window klUCB (SW-klUCB) with window-size $\tau = \lceil 2\sqrt{T\ln(T)/\Upsilon_T}\rceil$. As for actively adaptive algorithms, we experiment with AdSwitch (Auer et al., 2019b) and three algorithms combining a change-point detector with klUCB: CUSUM-klUCB, M-klUCB and GLR-klUCB. These three algorithms share the use of an exploration parameter that we call α (or an exploration sequence α_k for GLR-klUCB). Liu et al. (2018) and Cao et al. (2019) recommend two slightly different tuning for CUSUM-klUCB and M-klUCB respectively, that both scale in $\sqrt{\Upsilon_T \ln(T)/T}$. This is also the order of magnitude of α_{Υ_T} given by Corollary 6

4. https://github.com/EmilieKaufmann/PiecewiseStationaryBandits
5. Liu et al. (2018); Cao et al. (2019) both mention that extending their analysis to the use of klUCB should be easy.

for GLR-klUCB. Hence, in order to compare algorithm that adds a similar amount of exploration, we set $\alpha = \sqrt{\Upsilon_T A \ln(T)/T}$ for all algorithms using a constant exploration probability and $\alpha_k = \sqrt{kA\ln(T)/T}$ for the exploration sequence of GLR-klUCB.

Regarding the parameters of the change-point detectors, we use a threshold $h = \ln(T/\Upsilon_T)$ for CUSUM-klUCB, as recommended by Liu et al. (2018), and experience with different values of (M, ϵ) that have to be tuned using some prior knowledge of the problem. For M-klUCB, we experiment with different values of the windows parameter w (often choosing the tuning $w = 800$ that was found to be robust in the experiments of Cao et al. (2019)) and use the recommended threshold $b = \sqrt{w\ln(2AT^2)/2}$. For the Bernoulli-GLR test, we use the threshold function $\beta(n, \delta) = \ln(n^{3/2}/\delta)$ and set $\delta = 1/\sqrt{T}$, which is the largest value licensed by Corollary 6.

For GLR-klUCB and AdSwitch, which are computationally more demanding due to the tests based on scan-statistics, we use some implementation tweaks. For GLR-klUCB, we use some down-sampling as discussed in Section 3.3, performing the test only every $\Delta t = 10$ time steps and scanning every $\Delta s = 5$ observations for a possible change-point. To be able to implement AdSwitch up to a horizon $T = 5000$, we set $\Delta t = 50$. The computational bottleneck in AdSwitch is the checks on good arms that compare the empirical mean between s and t to that between s_1 and s_2 for all possible $s < t$ and $s_1 \leq s_2 < t$. We only test values of s_1, s_2 and s satisfying $s' \mod \Delta s = 0$, for $\Delta s = 20$. This reduces the time complexity by $(\Delta s)^3$, which is a significant speed-up in practice. Finally, the parameter C_1 that governs the elimination of good arms and should be chosen large enough was set to $C_1 = 1$.

6.2 Results on two simple benchmarks

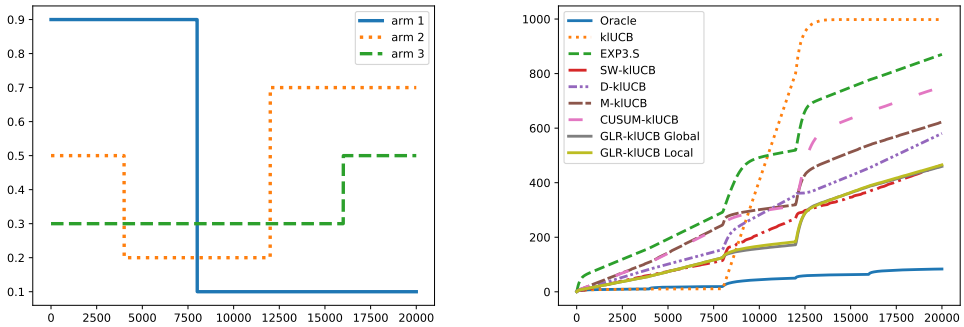
We design two simple piecewise stationary bandit problems with $A = 3$ arms and $\Upsilon_T = 4$ breakpoints. These breakpoints are evenly spaced up to the horizon, for which we investigate 4 values for each problem: $T = 5000$, $T = 10000$, $T = 20000$ and $T = 100000$. In **Problem 1**, a single arm changes in each breakpoint ($\Upsilon_T = C_T = 4$) and $\Delta^{\text{change}} = 0.3$, whereas in **Problem 2**, all arms means change at every breakpoint ($\Upsilon_T = 4, C_T = 16$) and $\Delta^{\text{change}} = 0.2$. For each problem, we display the reward functions of each arm in the top left corner of Figures 1 and 2.

For the different values of the horizon, the reward functions are simply expanded: the size Δ^{change} remains the same and the breakpoints are still evenly spaced. Hence, it is a way to vary the difficulty of the underlying change-point detection problems. Indeed, when T goes larger, the distance between two consecutive breakpoints increases, and Assumption 4 is closer to be satisfied. On these simple problems with equally spaced breakpoints ($\tau^{(k)} - \tau^{(k-1)} = T/5$), with our choice of α_k and δ, Assumption 4 amounts to

$$\sqrt{T} \geq \frac{80\sqrt{A}}{\min_{k=1}^{\Upsilon_T} \sqrt{k}(\Delta^{c,(k)})^2} \sqrt{\ln(T)} + 1$$

which is only satisfied for T much larger than 100000 for both Problem 1 and Problem 2. Therefore, in this experiment, we investigate the performance of GLR-klUCB for difficult problems on which it does not have theoretical guarantees.

In Figure 1 (respectively Figure 2), we display the results for Problem 1 (respectively Problem 2). We display the regret of each algorithm as a function of the rounds for one horizon (top right corner); and we also tabulate the regret at the horizon and the number of restarts for all the algorithms and all the horizons. In Problem 1, we observe that the regret of GLR-klUCB with Global and

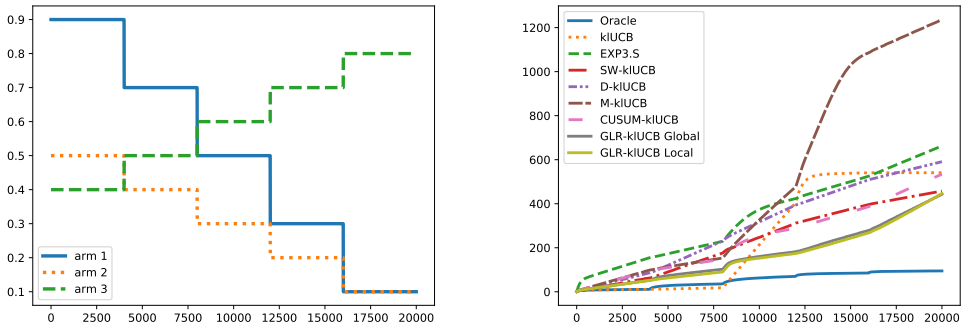


Algorithm	$T = 5000$		$T = 10000$		$T = 20000$		$T = 100000$	
	$(\alpha \simeq 0.142)$		$(\alpha \simeq 0.105)$		$(\alpha \simeq 0.077)$		$(\alpha \simeq 0.019)$	
Oracle-klUCB	59 ± 11	4	70 ± 13	4	84 ± 15	4	114 ± 17	4
klUCB	270 ± 66	0	515 ± 110	0	998 ± 202	0	4769 ± 784	0
EXP3.S	379 ± 31	0	578 ± 44	0	870 ± 64	0	2125 ± 142	0
SW-klUCB	$\mathbf{186 \pm 19}$	0	$\mathbf{298 \pm 29}$	0	$\mathbf{465 \pm 37}$	0	$\mathbf{1269 \pm 74}$	0
D-klUCB	216 ± 14	0	355 ± 19	0	581 ± 26	0	1731 ± 55	0
AdSwitch	1339 ± 96	2.2	–	–	–	–	–	–
M-klUCB ($w = 200$)	273 ± 32	1.8	408 ± 50	1.9	612 ± 89	1.9	1848 ± 532	1.8
M-klUCB ($w = 800$)	280 ± 34	1.2	415 ± 60	1.4	623 ± 92	1.8	1811 ± 343	2
CUSUM-klUCB ($M = 200, \varepsilon = 0.1$)	280 ± 50	3.2	422 ± 79	4.1	646 ± 123	5	1776 ± 265	5.8
CUSUM-klUCB ($M = 400, \varepsilon = 0.05$)	321 ± 89	4.6	485 ± 119	7.1	752 ± 179	11.1	2018 ± 457	28.5
GLR-klUCB Global (α_k)	$\mathbf{199 \pm 38}$	2	$\mathbf{290 \pm 59}$	2	$\mathbf{460 \pm 95}$	2.1	1420 ± 340	2.8
GLR-klUCB Global (constant α)	268 ± 33	2	403 ± 59	2.2	609 ± 91	2.5	1804 ± 348	2
GLR-klUCB Local (α_k)	$\mathbf{200 \pm 39}$	2.1	$\mathbf{295 \pm 59}$	2.2	$\mathbf{465 \pm 103}$	2.4	1403 ± 354	3
GLR-klUCB Local (constant α)	269 ± 32	2.3	399 ± 52	2.6	602 ± 82	2.9	1786 ± 334	3

Figure 1: Results for **Problem 1** (displayed in the top left corner). The table shows the final regret R_T (left column) and the average number of restarts (right column) for several algorithms run for different values of the horizon T. The top right corner displays the cumulative regret of the top 8 algorithms for $T = 20000$. The regret is estimated based on $N = 1000$ independent repetitions (except for $T = 100000$ where $N = 100$ and for the AdSwitch algorithm for which $N = 50$).

Local restart is competitive with that of the SW-klUCB which performs best for the different time horizons T. However, recall that this algorithm is tuned using the knowledge of Υ_T, unlike GLR-klUCB. In Problem 2, the regret of GLR-klUCB is the smallest for large horizons ($T = 20000, 100000$) whereas for shorter horizons ($T = 5000, 10000$) klUCB and SW-klUCB have (slightly) smaller regret. Regarding other passively adaptive approaches, we see that D-klUCB is competitive with (sometimes even better than) actively adaptive algorithms, whereas Exp3.S only manages to outperform klUCB for large horizons. GLR-klUCB largely outperforms AdSwitch for $T = 5000$, which is the largest horizon for which we could implement this algorithm. We now turn our attention to CPD-based algorithms.

The tests used by CUSUM-UCB and M-UCB depend on two sets of parameters that should in principle be chosen according to some prior knowledge of the problem, and we tried for each algorithm two different tunings of these parameters. For CUSUM-UCB, the two sets of parameters yield similar regret on Problem 2, but one is much better than the other on Problem 1. For M-UCB, the two sets of parameters yield similar regret on Problem 1, but one is much better than the other



Algorithm	$T = 5000$		$T = 10000$		$T = 20000$		$T = 100000$	
	$(\alpha \simeq 0.142)$		$(\alpha \simeq 0.105)$		$(\alpha \simeq 0.077)$		$(\alpha \simeq 0.019)$	
Oracle-klUCB	65 ± 14	4	80 ± 17	4	95 ± 18	4	127 ± 24	4
klUCB	$\mathbf{162 \pm 52}$	0	$\mathbf{293 \pm 82}$	0	541 ± 125	0	2425 ± 440	0
EXP3.S	304 ± 30	0	448 ± 42	0	662 ± 56	0	1587 ± 137	0
SW-klUCB	182 ± 15	0	$\mathbf{288 \pm 22}$	0	$\mathbf{457 \pm 29}$	0	1258 ± 65	0
D-klUCB	218 ± 15	0	361 ± 21	0	591 ± 29	0	1787 ± 61	0
AdSwitch	1418 ± 113	1.3	–	–	–	–	–	–
M-klUCB ($w = 200$)	485 ± 89	0	927 ± 163	0	1813 ± 285	0	9674 ± 1082	0
M-klUCB ($w = 800$)	333 ± 112	1	609 ± 272	0.8	1237 ± 617	0.6	7726 ± 3331	0.4
CUSUM-klUCB ($M = 200, \varepsilon = 0.1$)	238 ± 25	6	352 ± 39	7.7	520 ± 64	8.6	1249 ± 191	12.1
CUSUM-klUCB ($M = 400, \varepsilon = 0.05$)	238 ± 27	5.7	353 ± 51	9.1	534 ± 92	15.1	1270 ± 198	38.5
GLR-klUCB Global (α_k)	209 ± 13	3	$\mathbf{303 \pm 16}$	3.5	$\mathbf{443 \pm 16}$	4	$\mathbf{989 \pm 20}$	4
GLR-klUCB Global (constant α)	252 ± 13	3	350 ± 14	3.5	501 ± 15	4	1115 ± 17	4
GLR-klUCB Local (α_k)	206 ± 17	3.2	$\mathbf{298 \pm 23}$	4.1	$\mathbf{447 \pm 28}$	5.8	$\mathbf{1058 \pm 96}$	5.5
GLR-klUCB Local (constant α)	249 ± 19	3.1	349 ± 26	4.5	506 ± 43	5.2	1230 ± 243	4.9

Figure 2: Results for **Problem 2** (displayed in the top left corner). The table shows the final regret R_T (left column) and the average number of restarts (right column) for several algorithms run for different values of the horizon T. The top right corner displays the cumulative regret of the top 8 algorithms for $T = 20000$. The regret is estimated based on $N = 1000$ independent repetitions (except for $T = 100000$ where $N = 100$ and for the AdSwitch algorithm for which $N = 50$).

on Problem 2. This sheds light on the fact that tuning these parameters may be difficult. On the contrary, the tuning of the Bernoulli GLR test used in GLR-klUCB only requires to specify the error probability δ, and setting it to $\delta = 1/\sqrt{T}$ as suggested by Corollary 6 yield good performance on both Problem 1 and Problem 2.

The table shows the final regret R_T (left column) and the average number of restarts (right column) for several algorithms run for different values of the horizon T. The top right corner displays the cumulative regret of the top 8 algorithms for $T = 20000$. The regret is estimated based on $N = 1000$ independent repetitions (except for $T = 100000$ where $N = 100$ and for the AdSwitch algorithm for which $N = 50$).

To understand the behavior of the CPD-based algorithms, we analyze their average number of restarts, reported in the tables in Figure 1 and 2. In an asymptotic regime (i.e. for T such that Assumption 4 or Assumption 11 is satisfied), GLR-klUCB should detect all breakpoints with Global restart and all change-points with Local restart. As can be seen, the asymptotic regime is not met in our experiments, except for $T \geq 20000$ on Problem 2 in which GLR-klUCB with Global restart

performs exactly $\Upsilon_T = 4$ restarts. Besides this case, GLR-klUCB typically detects fewer changes than expected, for example between 2 and 3 on Problem 1. Note that M-UCB tends to detect fewer changes than GLR-klUCB, whereas CUSUM-UCB tend to detect more. Especially, when the parameter ε (giving the minimal amount of change the CUSUM test should detect) is $\varepsilon = 0.05$, we observe that CUSUM-UCB experiences false-alarms, especially for large horizons (yet this does not prevent the algorithm from having a regret smaller than that of klUCB). Overall, we remark that GLR-klUCB is among the best algorithms on both problems for all the horizon values, including the smallest ones: it shows that GLR-klUCB is competitive in practice even when the Assumptions 4 and 11 are violated.

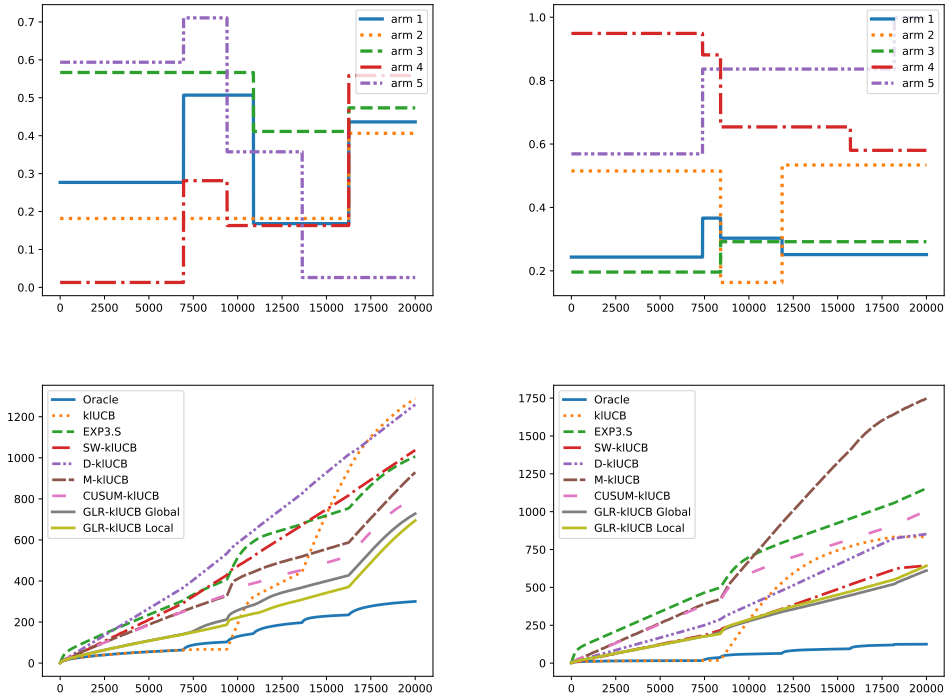
In these experiments, we tried four variants of GLR-klUCB: we investigate Global and Local restarts and two exploration sequences: a constant exploration probability $\alpha = \sqrt{\Upsilon_T A \ln(T)/T}$ and the exploration sequence $\alpha_k = \sqrt{kA\ln(T)/T}$ that does not require to know the number of breakpoints. We observe that the two types of restarts yield comparable performance (with a slight advantage for Global restarts), and thus investigate the two variants further on a wider benchmark. As for the exploration sequences, we observe that the time-varying one (agnostic to Υ_T) always performs best. The reason is that it performs less forced exploration in the first episodes, and as we shall see in our next experiments, scaling down the exploration probability (or exploration sequence) for CPD-based algorithms can lead to better empirical performance. Still, GLR-klUCB with a constant exploration probability α also outperforms most of the time other CPD-based algorithms using the exact same α.

6.3 Robustness on more diverse benchmarks

We now investigate further the performance of the best algorithms for Problem 1 and Problem 2 on a large number of randomly generated piecewise stationary bandit models, with $T = 20000$. To generate a random instance, we specify the number of arms A, the maximal number of breakpoints Υ, a change-point probability p, a minimal distance $d_{\min}$, a minimal and maximal amount of change, $\Delta_{\min}$ and $\Delta_{\max}$. Then, we sample the breakpoints uniformly at random under the constraint that $\tau^k - \tau^{k-1} \geq d_{\min}$. For each breakpoint k, each arm has a probability p to experience a change-point, whose magnitude is chosen uniformly at random in $[\Delta_{\min}, \Delta_{\max}]$.

First, we sample two problems from this procedure with $A = 5$ arms, $\Upsilon = 5$ breakpoints, with spacing larger than $d_{\min} = 1000$, a magnitude in $[0.05, 0.4]$, and a change-point probability $p = 0.5$. Results for these two problems are displayed in Figure 3. On **Problem 3**, there are important changes of the optimal arm as the initial worse arm ends up being the best, with three changes of optimal arms. On **Problem 4**, the best two arms remain the same but are switched around the middle of the budget T. As can be seen in Figure 3, on these two instances the two versions of GLR-klUCB attain the smallest regret (with SW-klUCB that is very competitive in Problem 4). We note that the Local version performs slightly more restarts (as can already be observed on Problems 1 and 2), but there is again no clear winner between global and local restarts. These experiments also confirm our conclusions regarding the lack of robustness of M-UCB (with $w = 800$) and CUSUM-UCB (with $\varepsilon = \Delta_{\min}$) and their tendency to under-detect or over-detect, respectively.

Finally, in Table 1, we report the regret of the different algorithms averaged over $N = 2000$ different instances with $K = 5$, $\Upsilon = 6$, $\Delta_{\min} = 0.05$, $\Delta_{\max} = 0.3$, $d_{\min} = 1000$ and $p = 0.5$. We can remark that even in the most favorable situation in which the amount of change in each breakpoint is maximal ($\Delta^{c,(k)} = \Delta_{\max}$ for all k) and the breakpoints are evenly spaced ($\tau^{(k)} - \tau^{(k-1)} = T/7$),



Algorithm	Problem 3		Problem 4	
Oracle-klUCB	300 ± 37	5	125 ± 15	5
klUCB	1289 ± 206	0	834 ± 68	0
EXP3.S	1006 ± 91	0	1154 ± 78	0
SW-klUCB	1037 ± 39	0	$\mathbf{643 \pm 29}$	0
D-klUCB	1260 ± 39	0	852 ± 31	0
M-klUCB ($w = 800$)	929 ± 53	1.1	1747 ± 695	0.5
CUSUM-klUCB ($M{=}400, \varepsilon{=}0.05$)	801 ± 87	20.7	1006 ± 102	6.277
GLR-klUCB Global (α_k)	$\mathbf{728 \pm 65}$	3.8	$\mathbf{611 \pm 43}$	3.1
GLR-klUCB Local (α_k)	$\mathbf{695 \pm 59}$	4.7	$\mathbf{642 \pm 44}$	3.6

Figure 3: 9 algorithms run with $T = 20000$ on **Problem 3** (left) and **Problem 4** (right). The middle row displays the cumulative regret as a function of time and the bottom table displays the final regret and number of restarts averaged over $N = 1000$ runs.

the detection delay $d^{(k)}$ defined in Assumption 4 is lower bounded by

$$d^{(1)} \geq 4\sqrt{\frac{AT}{\ln(T)}}\left(\frac{\frac{3}{2}\ln\left(\frac{3T}{14}\right) + \frac{1}{2}\ln(T)}{\Delta_{\max}^2} + 1\right),$$

which is larger than 70000 for $T = 20000$ and $\Delta_{\max} = 0.3$, and is in particular much larger than the length of any stationary period. As a consequence, Assumption 4 cannot be satisfied in any of the $N = 2000$ instances that were drawn at random[6]. Hence, this last experiment also

6. A similar reasoning explains was Assumption 11 cannot be satisfied for such instances either.

investigates the performance of GLR-klUCB beyond the asymptotic regime in which it is studied in theory. In this experiment, we also study the influence of using a smaller exploration probability, $\alpha = \alpha_0\sqrt{\Upsilon_T A \ln(T)/T}$ with some constant $\alpha_0 \in (0, 1]$ (and multiplying α_k by the same α_0 for GLR-klUCB). We see that for $\alpha_0 = 1$, GLR-klUCB outperforms all non-stationary bandit algorithms, but not klUCB. This can be explained by the fact that some of the random instances may contain no change of optimal arm, and klUCB is expected to be good in this setting, without paying the price of forced exploration. However, when we reduce the amount of forced exploration by setting α_0 to 0.5 and 0.1, klUCB is outperformed by actively adaptive algorithms, out of which GLR-klUCB has the best performance, with a slight advantage for Global restart.

Algorithm	$\alpha_0 = 1$		$\alpha_0 = 0.5$		$\alpha_0 = 0.1$	
klUCB	**515**	0	515	0	515	0
EXP3.S	1083	0	1083	0	1083	0
SW-klUCB	826	0	826	0	826	0
D-klUCB	1022	0	1022	0	1022	0
M-klUCB ($w = 800$)	1070	0.9	716	0.8	415	0.8
CUSUM-klUCB ($M = 400, \varepsilon = 0.05$)	973	12.4	640	10.9	347	10.9
GLR-klUCB Global (α_k)	**611**	3.1	**404**	3.0	**252**	2.9
GLR-klUCB Local (α_k)	**691**	4.4	**462**	3.8	**284**	3.5

Table 1: Regret and number of restarts for several algorithm run with $T = 20000$, averaged over $N = 20000$ random problem instances with $K = 5$, $\Upsilon = 6$, $\Delta_{\min} = 0.05$, $\Delta_{\max} = 0.3$, $d_{\min} = 1000$ and $p = 0.5$.

6.4 GLR-klUCB beyond piecewise stationary models

We summarize in this paragraph the experimental results of (Seznec et al., 2020), who performed experiments including GLR-klUCB[7] on a restless rotting bandit problem, in which the reward function $\mu_a(t)$ of each arm is assumed to be a non-increasing function of t. They designed 9 bandit games with arms mean rewards learned from the R6A - Yahoo! Front Page Today Module. In this setup, each arm corresponds to a news article and its mean reward at a certain time step (the probability that the news is clicked on) is learned from the dataset with a sliding window average of 30 000 samples. Each game corresponds to a 12 hours time frame between 6 p.m. and 6 a.m. EST during one day of May 2009 and the number of rounds is equal to the number of users visiting the Yahoo! Front Page on the period. This time frame was chosen because the mean rewards (click probabilities) are mostly decaying. However, besides this high-level data selection, they do not enforce any of their theoretical assumptions (the reward functions are neither piecewise stationary, nor strictly non-increasing).

They studied GLR-klUCB (with Gaussian confidence intervals) with no forced exploration ($\alpha = 0$) and local restart together with Exp3.S and two of their algorithms designed for the rotting case, namely RAW-UCB and FEWA. GLR-klUCB recovers consistently the best performance on the 9 games, significantly outperforming FEWA and Exp3.S. RAW-UCB performs almost the same as GLR-klUCB, though the former does not need the knowledge of T to be tuned. Interestingly, the regret of GLR-klUCB has the same logarithmic shape as FEWA or RAWUCB when one arm is

7. A preliminary version of this paper appeared on arXiv before the paper (Seznec et al., 2020), which explains why GLR-klUCB could be used in this work

significantly above the others. This is not the case for Exp3.S which keeps a small linear trend in regret due to random exploration. These algorithms were proved to have logarithmic regret on each stationary part of a non-increasing piecewise bandit problem. This suggests that $\mathrm{GLR\text{-}klUCB}$ with no forced exploration and local restart could also enjoy logarithmic regret in a similar setting.

7. Conclusion

We proposed a new algorithm based on a change-point detector which empirically outperforms other CPD-based approaches designed for the general piecewise-stationary bandit problems, while attaining similar regret without the knowledge of the number of breakpoints and without any prior knowledge of the arms means. We proved that $\mathrm{GLR\text{-}klUCB}$ attains a $\mathcal{O}(\sqrt{\Upsilon_T A T \ln(T)}/(\Delta^{\text{change}})^2)$ regret for "easy" instances in which the breakpoints are far enough from each other. When the smallest magnitude of a breakpoint Δ^{change} is not too small, this is comparable to the regret of recently proposed algorithms that are agnostic to Υ_T but whose implementation is much more intricate than that of $\mathrm{GLR\text{-}klUCB}$.

The presence of Δ^{change} in the regret bound comes from the fact that our analysis (as that of other CPD-based algorithms) assumes that all the breakpoints are detected by the algorithm. But our experiments reveal that on "harder" instances, the algorithm actually does not detect all the breakpoints and still attain small regret. Hence, in future work, we intend to work on an improved analysis of $\mathrm{GLR\text{-}klUCB}$ introducing some notion of "meaningful changes" that need to be detected by the algorithm to ensure a small regret. We will also investigate whether $\mathrm{GLR\text{-}klUCB}$ can be used without forced exploration under some extra assumption on the types of breakpoints encountered, such as global changes (Mukherjee and Maillard, 2019) or restless rotting bandits (Seznec et al., 2020).

Acknowledgments

The research presented was supported by European CHIST-ERA project DELTA and the French ANR projects BADASS (ANR-16-CE40-0002) and BOLD (ANR-19-CE23-0026-04).

References

Reda Alami, Odalric-Ambrym Maillard, and Raphael 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.

J-Y. Audibert and S. Bubeck. Regret Bounds And Minimax Policies Under Partial Monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

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.

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Peter Auer, Yifang Chen, Pratik Gajane, Chung-Wei Lee, Haipeng Luo, Ronald Ortner, and Chen-Yu Wei. Achieving optimal dynamic regret for non-stationary bandits without prior information. In *Conference on Learning Theory (COLT)*, 2019a.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Congerence on Learning Theory (COLT)*, 2019b.

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.

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.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. In *Congerence on Learning Theory (COLT)*, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *AISTATS*, 2019.

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

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.

Chen Jie and AK Gupta. Parametric statistical change point analysis. *Birkh User*, 2000.

E. Kaufmann and W.M. Koolen. Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

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.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*. Cambridge University Press, 2019.

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.

Subhojyoti Mukherjee and Odalric-Ambrym Maillard. Distribution-dependent and time-uniform bounds for piecewise i.i.d bandits. In *ICML Workshop on Reinforcement Learning for Real Life*, 2019.

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.

Julien Seznec, Pierre Ménard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. In *The 23rd International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2020.

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. In *ACC*, 2018.

S.S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Yanhong Wu. *Inference for change point and post change means after a CUSUM test*, volume 180. Springer, Lecture Notes in Statistics, 2007.

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.

A. Proof of Proposition 3

We consider one arm $a \in \{1, \dots, A\}$, and when the GLR-klUCB algorithm is running, we consider two time steps $s \leq t \in \mathbb{N}^\star$, chosen between two restart times for that arm a. All time steps $u \in \{s, \dots, t\}$ belong to the same episode, that is $k_u = k_t$. Lines 3-4 of Algorithm 1 imply that for all $u \in \{s+1, \dots, t\}$,

$$\left\{ u \mod \left\lceil \frac{A}{\alpha_{k_t}} \right\rceil = a \right\} \subset \{A_u = a\}$$

Thus we have

$$\begin{aligned} n_a(t) - n_a(s) &= \sum_{u=s+1}^{t} \mathbb{1}(A_u = a) \\ &\geq \sum_{u=s+1}^{t} \mathbb{1}\left(u \mod \left\lceil \frac{A}{\alpha_{k_t}} \right\rceil = a \right) \\ &= \big(t - (s+1) + 1\big) / \left\lceil \frac{A}{\alpha_{k_t}} \right\rceil \geq \left\lfloor \frac{\alpha_{k_t}}{A}(t-s) \right\rfloor, \end{aligned}$$

which proves Proposition 3.

B. Concentration Inequalities

B.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

$$\mathbb{E}[e^{\lambda X}] \leq e^{\phi_\mu(\lambda)} \quad \text{with} \quad \mu = \mathbb{E}[X], \tag{12}$$

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 (1). 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 9 below and from the definition of $\beta(n,\delta)$:

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

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 9 $(X_i)_{i\in\mathbb{N}}$ *and* $(Y_i)_{i\in\mathbb{N}}$ *two independent* i.i.d. *processes with resp. means* μ *and* μ' *such that*

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

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* $\hat{\mu}_s = \frac{1}{s}\sum_{a=1}^{s} X_i$ *and* $\hat{\mu}'_r = \frac{1}{r}\sum_{a=1}^{r} Y_i$*, it holds that for every* $s, r \in \mathbb{N}^\star$*,*

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

where $\mathcal{T}$ *is the function defined in* (3)*.*

Proof of Lemma 9 Using the same construction as in the proof of Theorem 14 in Kaufmann and Koolen (2021), 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 $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ that satisfies $\mathbb{E}[M^\lambda(s)] \leq 1$ and

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

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

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

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 $\mathcal{F}'_r = \sigma(Y_1, \ldots, Y_r)$ such that

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

for the same function $g(\lambda)$. In the terminology of Kaufmann and Koolen (2021), 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-VCC for Ville-Cramér-Chernoff, as Ville'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's maximal inequality still works). One can write

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

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

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

for any function g_ξ defined above. The conclusion follows by optimizing for both λ and ξ, using Lemma 18 in Kaufmann and Koolen (2021).

B.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 10 *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*

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

Proof of Lemma 10 We first note that

$$\mathbb{P}\left(\frac{sr}{s+r}\left(\hat{\mu}_{a,s}-\hat{\mu}_{b,r}-\Delta\right)^2\geq u\right)$$
$$\leq \mathbb{P}\left(\hat{\mu}_{a,s}-\hat{\mu}_{b,r}\geq \Delta+\sqrt{\frac{s+r}{sr}u}\right)+\mathbb{P}\left(\hat{\mu}_{b,r}-\hat{\mu}_{a,s}\geq -\Delta+\sqrt{\frac{s+r}{sr}u}\right), \tag{13}$$

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

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

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

$$\mathbb{P}\left(\hat{\mu}_{1,n_1}-\hat{\mu}_{2,n_2}\geq \mu_1-\mu_2+x\right)\leq \exp\left(-\frac{n_1n_2}{n_1+n_2}\frac{x^2}{2\sigma^2}\right).$$

Using this inequality twice in the right hand side of (13) concludes the proof.

C. Elements of the analysis of GLR-klUCB with Global Restarts

We present in this section the detailed proof of the two crucial lemmas in the analysis of GLR-klUCB with Global Restarts.

C.1 Proof of Lemma 7

Lemma 7 follows from summing over k and a upper bounds on the quantities

$$E_{k,a}:=\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=\tau^{(k)}+1}^{\tau^{(k+1)}}\Delta_a^{(k)}\mathbb{1}(A_t=a,\mathrm{UCB}_a(t-1)\geq \mathrm{UCB}_{k^\star}(t-1))\right],$$

for each $k \in \{0, \dots, \Upsilon_T\}$ and each arm a such that $\Delta_a^{(k)} > 0$.

Using that on $\mathcal{E}_T$, $\hat{\tau}^{(k)} \leq \tau^{(k)} + d^{(k)}$, one can write $E_{k,a} \leq \Delta_a^{(k)} d^{(k)} + R_{k,a}$ where

$$\begin{aligned} R_{k,a} &= \mathbb{E}\left[\mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \Delta_a^{(k)} \mathbb{1}(A_t = a, \mathrm{UCB}_a(t-1) \geq \mathrm{UCB}_{k^\star}(t-1))\right] \\ &= \underbrace{\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \mathbb{1}\left(\mathrm{UCB}_{k^\star}(t-1) \leq \mu_{k^\star}^{(k)}\right)\right]}_{:=(A)} + \underbrace{\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \Delta_a^{(k)} \mathbb{1}\left(A_t = a, \mathrm{UCB}_a(t-1) \geq \mu_{k^\star}^{(k)}\right)\right]}_{:=(B)} \end{aligned}$$

To upper bound (A) and (B), we recall the following notation. We let $\hat{\tau}(t)$ be the last time before t that the algorithm restarted. Moreover, we denote by $n_a(t) = \sum_{s=\tau(t)+1}^{t} \mathbb{1}(A_s = a)$ the number of selections of arm a since the last (global) restart, and $\hat{\mu}_a(t) = \frac{1}{n_a(t)} \sum_{s=\tau(t)+1}^{t} X_{a,s} \mathbb{1}(A_s = a)$ their empirical average (if $n_a(t) \neq 0$).

Upper bound on Term (A) By definition of the Upper Confidence Bound, one can write

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

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

$$\mathcal{C}^{(k)} = \left\{\forall j \leq k, \hat{\tau}^{(j)} \in \{\tau^{(j)} + 1, \dots, \tau^{(j)} + d^{(j)}\}\right\}. \tag{14}$$

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, $\hat{\mu}_{k^\star}(t)$ is the average of samples that have all mean $\mu_{k^\star}^{(k)}$. Thus, introducing $\hat{\mu}_s$ as a sequence of *i.i.d.* random variables with mean $\mu_{k^\star}^{(k)}$, one can write

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

Using the concentration inequality given in Lemma 2 of Cappé et al. (2013) and the fact that $f(t) = \ln(t) + 3\ln(\ln(t))$ allow to upper bound the probability corresponding to term t' by $1/(t' \ln(t'))$.

Using the law of total expectation yields

$$(A) \leq 2 + \sum_{t=3}^{\tau^{(k+1)}-\tau^{(k)}} \frac{1}{t\ln(t)} \leq 3 + \ln\left(\ln\left(\tau^{(k+1)} - \tau^{(k)}\right)\right).$$

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

$$\begin{aligned}
(B) &= \Delta_a^{(k)} \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \mathbb{1}\left(A_t = a, \mathrm{UCB}_a(t-1) \geq \mu_{k^\star}^{(k)}\right)\Big] \\
&= \Delta_a^{(k)} \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \mathbb{1}\left(A_t = a, n_a(t-1)\mathrm{kl}\left(\hat{\mu}_a(t-1), \mu_{k^\star}^{(k)}\right) \leq f\left(t - \hat{\tau}^{(k)}\right)\right)\Big] \\
&\leq \Delta_a^{(k)} \mathbb{E}\left[\mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}^{(k)}+1}^{\tau^{(k+1)}} \sum_{s=1}^{t-\hat{\tau}^{(k)}} \mathbb{1}\left(A_t = a, n_a(t-1) = s\right) \mathbb{1}\left(s\,\mathrm{kl}(\tilde{\mu}_{a,s}^{(k)}, \mu_{k^\star}^{(k)}) \leq f\left(\tau^{(k+1)} - \tau^{(k)}\right)\right)\right] \\
&\leq \Delta_a^{(k)} \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \sum_{s=1}^{n_a(\tau^{(k+1)})} \mathbb{1}\left(s\,\mathrm{kl}(\tilde{\mu}_{a,s}^{(k)}, \mu_{k^\star}^{(k)}) \leq f(\tau^{(k+1)} - \tau^{(k)})\right)\Big] \\
&\leq \Delta_a^{(k)} \mathbb{E}\Big[\mathbb{1}(\mathcal{C}_k) \sum_{s=1}^{n_a(\tau^{(k+1)})} \mathbb{1}\left(s\,\mathrm{kl}(\tilde{\mu}_{a,s}^{(k)}, \mu_{k^\star}^{(k)}) \leq f(\tau^{(k+1)} - \tau^{(k)})\right)\Big],
\end{aligned}$$

where $\mathcal{C}^{(k)}$ is the event already defined in (14). Conditionally to $\mathcal{F}_{\hat{\tau}^{(k)}}$, when $\mathcal{C}^{(k)}$ holds, for $s \in \{1, \ldots, n_a(\tau^{(k+1)})\}$, $\tilde{\mu}_{a,s}^{(k)}$ is the empirical mean from *i.i.d.* observations of mean $\mu_a^{(k)}$. Therefore, introducing $\hat{\mu}_s$ as a sequence of *i.i.d.* random variables with mean $\mu_a^{(k)}$, it follows from the law of total expectation that

$$(B) \leq \Delta_a^{(k)} \sum_{s=1}^{\tau^{(k+1)}-\tau^{(k)}} \mathbb{P}\left(s \times \mathrm{kl}(\hat{\mu}_s, \mu_{k^\star}^{(k)}) \leq f\left(\tau^{(k+1)} - \tau^{(k)}\right)\right).$$

As $\hat{\mu}_s$ is the empirical mean of i.i.d. observation of mean $\mu_a^{(k)}$ and $\mu_{k^\star}^{(k)} > \mu_a^{(k)}$, and upper bound on this sum of probabilities can be found in Appendix A.2 of Cappé et al. (2013), which yields

$$(B) \leq \frac{\Delta_a^{(k)}}{\mathrm{kl}(\mu_a^{(k)}, \mu_{k^\star}^{(k)})} \ln\left(\tau^{(k+1)} - \tau^{(k)}\right) + \mathcal{O}\left(\sqrt{\ln\left(\tau^{(k+1)} - \tau^{(k)}\right)}\right).$$

Conclusion Combining the upper bound on (A) and (B) yields

$$E_{k,a} \leq \Delta_a^{(k)} d^{(k)} + \frac{\Delta_a^{(k)} \ln\left(\tau^{(k+1)} - \tau^{(k)}\right)}{\mathrm{kl}\left(\mu_a^{(k)}, \mu_{k^\star}^{(k)}\right)} + \mathcal{O}\left(\sqrt{\ln\left(\tau^{(k+1)} - \tau^{(k)}\right)}\right).$$

The bound in Lemma 7 follows from the observation that $E_{k,a}$ is also trivially upper bounded by $\Delta_a^{(k)}\left(\tau^{(k+1)} - \tau^{(k)}\right)$.

C.2 Proof of Lemma 8

Recall that $\mathcal{C}^{(k)}$ defined in (14) is the event that all the breakpoints up to the k-th have been correctly detected. Using a union bound, one can write

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

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

The final result follows by proving that $(a) \leq A\delta$ 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. Under the bandit algorithm, the change-point detector associated with each arm a is based on (possibly much) less than $t-\tau_a(t)$ samples from arm a, which makes false alarm even less likely to occur. More precisely, we upper bound term (a) by

$$\begin{aligned}(a) &\leq \mathbb{P}\left(\exists a, \exists s < t \leq n_a(\tau^{(k)}) : s\,\mathrm{kl}\left(\tilde{\mu}_{a,1:s}^{(k-1)}, \tilde{\mu}_{a,1:t}^{(k-1)}\right) + (t-s)\,\mathrm{kl}\left(\tilde{\mu}_{a,s+1:t}^{(k-1)}, \tilde{\mu}_{a,1:t}^{(k-1)}\right) > \beta(t,\delta) \mid \mathcal{C}^{(k-1)}\right)\\ &\leq \sum_{a=1}^{A} \mathbb{P}\left(\exists s < t : s\,\mathrm{kl}(\hat{\mu}_{1:s}, \mu_a^{(k-1)}) + (t-s)\,\mathrm{kl}(\hat{\mu}_{s+1:t}, \mu_a^{(k-1)}) > \beta(t,\delta)\right),\end{aligned}$$

with $\hat{\mu}_{s:s'} = \sum_{r=s}^{s'} Z_{i,r}$ where $Z_{i,r}$ is an *i.i.d.* sequence with mean $\mu_a^{(k-1)}$. Indeed, conditionally to $\mathcal{C}^{(k-1)}$, the $n_a(\tau^{(k)})$ successive observations of arm a arm starting from $\hat{\tau}^{(k)}$ are *i.i.d.* with mean $\mu_a^{(k-1)}$. Using Lemma 9, term (a) is upper bounded by $A\delta$.

Upper bound on the term (b): controlling the delay From the definition of $\Delta^{c,(k)}$, there exists an arm a such that $\Delta^{c,(k)} = |\mu_a^{(k)} - \mu_a^{(k-1)}|$. We shall prove that it is unlikely that the change-point detector associated with this arm a does not trigger within the delay $d^{(k)}$. Controlling the detection delay for arm a under the adaptive sampling scheme of GLR-klUCB is tricky and 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_k of the current episode.

First step: upper bound Assume that $\mathcal{C}^{(k-1)}$ holds. It follows from Proposition 3 that there exists $\bar{t} \in \{\tau^{(k)},\dots,\tau^{(k)}+d^{(k)}\}$ such that $n_a(\bar{t}) - n_a(\tau^{(k)}) = \bar{r}$ where $\bar{r} = \lfloor \frac{\alpha_k}{A} d^{(k)} \rfloor$. This is because the mapping $t \mapsto n_a(t) - n_a(\tau^{(k)})$ is non-decreasing, is 0 at $t = \tau^{(k)}$ and its value at $\tau^{(k)}+d^{(k)}$ is larger than $\bar{r}$ as $k_t \geq k$. Using that

$$\left(\hat{\tau}^{(k)} \geq \tau^{(k)}+d^{(k)}\right) \cap \mathcal{C}^{(k-1)} \subseteq \left(\hat{\tau}^{(k)} \geq \bar{t}\right) \cap \mathcal{C}^{(k-1)}$$

further implies that $(b) = \mathbb{P}\left(\hat{\tau}^{(k)} \geq \tau^{(k)}+d^{(k)} | \mathcal{C}^{(k-1)}\right)$ is upper bounded as follows:

$$(b) \leq \mathbb{P}\left(n_a(\tau^{(k)})\,\mathrm{kl}\left(\tilde{\mu}_{i,n_a(\tau^{(k)})}^{k-1}, \tilde{\mu}_{i,n_a(\bar{t})}^{k-1}\right) + \bar{r}\,\mathrm{kl}\left(\tilde{\mu}_{i,n_a(\tau^{(k)}):n_a(\bar{t})}^{k-1}, \tilde{\mu}_{i,n_a(\bar{t})}^{k-1}\right) \leq \beta(n_a(\tau^{(k)})+\bar{r},\delta)\middle|\mathcal{C}^{(k-1)}\right),$$

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

Second step: controlling $n_a(\tau^{(k)})$ Thanks to Proposition 3, we know that $n_a(\tau^{(k)})$ lies in the interval $\left[\left\lfloor \frac{\alpha_k}{A}\left(\tau^{(k)}-\hat{\tau}^{(k-1)}\right)\right\rfloor, \left(\tau^{(k)}-\hat{\tau}^{(k-1)}\right)\right]$. 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

$$n_a(\tau^{(k)}) \quad \in \quad \mathcal{I}_k \ \text{ where } \ \mathcal{I}_k := \left\{ \left\lfloor \frac{\alpha_k}{2A}(\tau^{(k)}-\tau^{(k-1)}) \right\rfloor, \ldots, \tau^{(k)}-\tau^{(k-1)} \right\}.$$

Introducing $\hat{\mu}_{a,s}$ (resp. $\hat{\mu}_{b,s}$) the empirical mean of s *i.i.d.* observations with mean $\mu_a^{(k-1)}$ (resp. $\mu_a^{(k)}$), 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 $\tilde{\mu}_{a,n_a(\bar{t})}^{k-1} = \left(n_a(\tau^{(k)})\tilde{\mu}_{a,n_a(\tau^{(k)})}^{k-1} + \overline{r}\tilde{\mu}_{a,n_a(\tau^{(k)}):n_a(\bar{t})}^{k-1}\right)/(n_a(\tau^{(k)})+\overline{r})$.

Third step: concluding with concentration inequalities Using Pinsker's inequality and introducing the gap $\Delta_a^{c,(k)} = \mu_a^{(k-1)} - \mu_a^{(k)}$ (which is such that $\Delta^{c,(k)} = |\Delta_a^{c,(k)}|$), one can write

$$\begin{aligned}
(b) &\leq \mathbb{P}\left(\exists s \in \mathcal{I}_k : \frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s}-\hat{\mu}_{b,\overline{r}}\right)^2 \leq \beta(s+\overline{r},\delta)\right) \\
&\leq \mathbb{P}\left(\exists s \in \mathbb{N} : \frac{2sr}{s+r}\left(\hat{\mu}_{a,s}-\hat{\mu}_{b,s}-\Delta_a^{c,(k)}\right)^2 \geq \beta(s+r,\delta)\right) \qquad (15) \\
&+ \mathbb{P}\left(\exists s \in \mathcal{I}_k : \frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s}-\hat{\mu}_{b,\overline{r}}-\Delta_a^{c,(k)}\right)^2 \leq \beta(s+\overline{r},\delta), \frac{2s\overline{r}}{s+\overline{r}}\left(\hat{\mu}_{a,s}-\hat{\mu}_{b,\overline{r}}\right)^2 \leq \beta(s+\overline{r},\delta)\right) \qquad (16)
\end{aligned}$$

Using Lemma 10 (given above in Appendix B.2) and a union bound, the first term (15) is upper bounded by δ (as $\beta(r+s,\delta) \geq \beta(s,\delta) \geq \ln(3s\sqrt{s}/\delta)$). For the second term (16) we use the observation that

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

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

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

Let $s_{\min} = \left\lfloor \frac{\alpha_k}{A}(\tau^{(k)}-\tau^{(k-1)})/2 \right\rfloor$. 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 can further write

$$\begin{aligned}
(b) \quad &\leq \quad \delta + \mathbb{P}\left(\exists s \in \mathcal{I}_k : \left(\Delta^{c,(k)}\right)^2 \leq 2\frac{s_{\min}+\overline{r}}{s_{\min}\overline{r}}\beta(\tau^{(k)}-\tau^{(k-1)}+\overline{r},\delta)\right) \\
&\leq \quad \delta + \mathbb{P}\left(\exists s \in \mathcal{I}_k : \left(\Delta^{c,(k)}\right)^2 \leq \frac{4}{\overline{r}}\beta\left(\frac{3}{2}(\tau^{(k)}-\tau^{(k-1)}),\delta\right)\right), \qquad (17)
\end{aligned}$$

where in the last step we use that by Assumption 4, it holds that $\overline{r} \leq s_{\min} \leq (\tau^{(k)} - \tau^{(k-1)})/2$. To conclude the proof, it remains to observe that by definition of the delay $d^{(k)}$,

$$\overline{r} = \left\lfloor \frac{\alpha_k}{A} d^{(k)} \right\rfloor > \frac{4}{\left(\Delta^{c,(k)}\right)^2} \beta\left(\frac{3}{2}(\tau^{(k)} - \tau^{(k-1)}), \delta\right)$$

hence the probability in the right hand side of (17) is equal to zero, which yields $(b) \leq \delta$.

D. GLR-klUCB with Local Restarts

Rather than featuring the number of *breakpoints* Υ_T, our analysis for local restarts features the number of *change-points* C_T defined below. We first define the number of change-points on arm a as

$$\text{NC}_a := \sum_{t=1}^{T-1} \mathbb{1}\left(\mu_t(a) \neq \mu_{t+1}(a)\right).$$

Clearly, $\text{NC}_a \leq \Upsilon_T$, but there can be an arbitrary difference between NC_a and Υ_T for some arms. We denote by $C_T := \sum_{a=1}^{A} \text{NC}_a$ the total number of change-points on the arms. Observe C_T can take all the values in $\{\Upsilon_T, \dots, A\Upsilon_T\}$.

We further denote by $\tau_a^{(\ell)}$ the position of the ℓ-th change-point for arm a:

$$\tau_a^{(\ell)} = \inf\{t > \tau_a^{(\ell-1)} : \mu_a(t) \neq \mu_a(t+1)\},$$

with the convention $\tau_a^{(0)} = 0$, and let $\overline{\mu}_a^{(\ell)}$ be the ℓ-th value for the mean of arm a, such that $\forall t \in [\tau_a^{(\ell-1)} + 1, \tau_a^{(\ell)}],\ \ \mu_a(t) = \overline{\mu}_a^{(\ell)}$. We also introduce the gap of the ℓ-th change-point on arm a which is $\Delta_a^{c,(\ell)} = \overline{\mu}_a^{\ell} - \overline{\mu}_a^{\ell-1} > 0$.

Assumption 11 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 12 provides a counterpart to Theorem 5 for GLR-klUCB based on Local Restart.

Assumption 11 *Define the delay* $d_a^{(\ell)} = d_a^{(\ell)}(\alpha, \delta) = \left\lceil \frac{4A}{\alpha_\ell \left(\Delta_a^{c,(\ell)}\right)^2} \beta\left(\frac{3}{2}(\tau_a^{(\ell)} - \tau_a^{(\ell-1)}), \delta\right) + \frac{A}{\alpha_\ell} \right\rceil$, *we assume that for all arm* a *and all* $\ell \in \{1, \dots, NC_a\}$, $\tau_a^{(\ell)} - \tau_a^{(\ell-1)} \geq 2\max(d_a^{(\ell)}, d_a^{(\ell-1)})$.

Theorem 12 *For* $\alpha = (\alpha_1, \alpha_2, \dots)$ *and* $\delta \in (0,1)$ *for which Assumption 11 is satisfied, the regret of* GLR-klUCB *with parameters* α *and* δ *based on* ***Local*** *Restart satisfies*

$$R_T \leq (A + 2C_T)\delta T + \alpha_{C_T+1} T$$
$$+ 2\sum_{a=1}^{A}\sum_{\ell=1}^{NC_a} \frac{4A}{\alpha_\ell \left(\Delta_a^{c,(\ell)}\right)^2} \beta\left(\tfrac{3}{2}(\tau_a^{(\ell)} - \tau_a^{(\ell-1)}), \delta\right) + \sum_{a=1}^{A}\sum_{\ell=0}^{NC_a} \frac{\ln(\tau_a^{(\ell+1)} - \tau_a^{\ell})}{\text{kl}\left(\overline{\mu}_a^{(\ell)}, \mu_a^{(\ell),*}\right)} + \mathcal{O}\left(\sqrt{\ln(T)}\right),$$

where $\mu_a^{(\ell),*} = \inf\left\{\mu_{a_t^\star}(t) : \mu_{a_t^\star}(t) \neq \overline{\mu}_a^{(\ell)}, t \in [\tau_a^{(\ell)} + 1, \tau_a^{(\ell+1)}]\right\}$, *with* $\inf \emptyset = 0$.

Corollary 13 *Recall* Δ^{opt} *defined in Section 5 and let* Δ *be the minimum value of* $\Delta_a^{c,(\ell)}$ *for all* a *and* $\ell \in \{1, \dots, NC_a\}$. *For problem instances satisfying the corresponding Assumption 11, for any* $\gamma > 1/2$,

1. *Choosing* $\alpha_k = \sqrt{\frac{kA\ln(T)}{T}}$ *and* $\delta = \frac{1}{T^\gamma}$ *yields*

$$R_T = \mathcal{O}\left((1+\gamma) \left(\sum_{a=1}^{A} \sqrt{NC_a} \right) \frac{\sqrt{AT\ln(T)}}{\Delta^2} + \frac{C_T \ln(T)}{(\Delta^{\text{opt}})^2} \right).$$

2. *If* C_T *is known, choosing* $\alpha_k = \sqrt{\frac{C_T A\ln(T)}{T}}$ *and* $\delta = \frac{1}{T^\gamma}$ *yields*

$$R_T = \mathcal{O}\left((1+\gamma) \frac{\sqrt{\mathrm{C}_T AT\ln(T)}}{\Delta^2} + \frac{\mathrm{C}_T \ln(T)}{(\Delta^{\text{opt}})^2} \right).$$

Corollary 13 specifies possible choices for the exploration sequence α and the parameter δ that yield $\sqrt{T}$ regret. If the number of change-point C_T is known, one can achieve $\mathcal{O}\Big(\sqrt{C_T AT\ln(T)}/\Delta^2\Big)$, where Δ is the minimal magnitude of any change-point (which can be smaller than Δ^{change} defined in Section 5). Without the knowledge of C_T, one achieves a slightly worse regret as the $\sqrt{C_T}$ factor is replaced by the larger $\sum_{a=1}^{A}\sqrt{\mathrm{NC}_a}$. As C_T can moreover take any value in $\{\Upsilon_T, \dots, A\Upsilon_T\}$, the guarantees obtained for GLR-klUCB with Local Restart are essentially worse than those obtained for Global Restart.

For particular instances such that $\Upsilon_T = \mathrm{C}_T$, *i.e.*, at each breakpoint only one arm changes (like in Problem 1 in Section 6), one obtains however competitive results for GLR-klUCB with Local Restart. Indeed, in that case $\Delta = \Delta^{\text{change}}$ and when Υ_T is known the two versions of GLR-klUCB achieve the same $\mathcal{O}(\sqrt{\Upsilon_T AT\ln(T)}/(\Delta^{\text{change}})^2)$ regret. Yet when Υ_T is unknown, GLR-klUCB with Local restart is only guaranteed to have a regret of $\mathcal{O}(A\sqrt{\Upsilon_T T\ln(T)}/(\Delta^{\text{change}})^2)$ when each arm has the same number of change-points, which is sub-optimal by a factor $\sqrt{A}$. Still, observe that these similar (or slightly worse) regret guarantees under Local Restart hold for a wider variety of problems as Assumption 11 is less stringent than Assumption 4.

We highlight that the results of Theorem 5 and Theorem 12 provide only upper bounds on the regret, which can be viewed as a sanity-check for using both variants of GLR-klUCB. The fact that our current results are worse for Local Restart is not in contradiction with our observation in Section 6 that the empirical performance of the two algorithms is very close.

D.1 Proof of Corollary 13

Choosing $\delta = T^{-\gamma}$ and $\alpha_k = \sqrt{k\ln(T)A/T}$, Theorem 5 upper bound the regret of GLR-klUCB by

$$(A+2C_T)T^{1-\gamma} + \sqrt{(C_T+1)AT\ln(T)} + \sum_{a=1}^{A}\sum_{\ell=1}^{\mathrm{NC}_a} \frac{4A}{\alpha_\ell(\Delta_a^{(\ell)})^2}\beta(\tfrac{3}{2}T, T^{-\gamma}) + \sum_{a=1}^{A}\sum_{\ell=1}^{\mathrm{NC}_a} \frac{\ln(T)}{\mathrm{kl}\left(\bar{\mu}_a^{(\ell)}, \mu_a^{(\ell),*}\right)} + \mathcal{O}\Big(\sqrt{\ln(T)}\Big).$$

For $\gamma > 1/2$, the leading term in this expression is

$$\sqrt{(C_T+1)AT\ln(T)} + \sum_{a=1}^{A}\sum_{\ell=1}^{\mathrm{NC}_a} \frac{4A}{\alpha_\ell(\Delta_a^{(\ell)})^2}\beta(\tfrac{3}{2}T, T^{-\gamma}) + \sum_{a=1}^{A}\sum_{\ell=1}^{\mathrm{NC}_a} \frac{\ln(T)}{\mathrm{kl}\left(\bar{\mu}_a^{(\ell)}, \mu_a^{(\ell),*}\right)}.$$

Using further that there exists some absolute constant C' such that $\beta(n,\delta) \le C'\ln(n/\delta)$, one obtains

$$R_T = \mathcal{O}\left(\sqrt{(C_T+1)AT\ln(T)} + (1+\gamma)\sum_{a=1}^{A}\sum_{\ell=1}^{\mathrm{NC}_a} \frac{A}{\alpha_\ell \Delta_a^{(\ell)}}\ln(T) + \frac{C_T\ln(T)}{(\Delta^{\text{opt}})^2} \right) \tag{18}$$

Finally, the first statement of the corollary follows from the fact that, for all i,

$$\sum_{\ell=1}^{\mathrm{NC}_a} \frac{1}{\alpha_\ell} = \sqrt{\frac{T}{A\ln(T)}} \sum_{\ell=1}^{\mathrm{NC}_a} \frac{1}{\sqrt{\ell}} \leq \sqrt{\frac{\mathrm{NC}_a T}{A\ln(T)}}.$$

Choosing $\alpha_\ell = \sqrt{\frac{AC_T \ln(T)}{T}}$ in (18) yields the second statement.

D.2 Proof of Theorem 12

We first introduce some notation for the proof. Recall that $\tau_a^{(\ell)}$ denotes the ℓ-th change-point for arm a. We use the convention $\tau_a^{(\mathrm{NC}_a+1)} = T$. We denote by $\hat{\tau}_a^{(\ell)}$ the ℓ-th change-point detected for arm a by GLR-klUCB, leading to a restart for this arm.

Distinguishing the exploration steps and the steps in which GLR-klUCB uses the UCBs to select the next arm to play, one can upper bound the regret as

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(t\left\lceil\frac{A}{\alpha_{k_t}}\right\rceil \in \{1,\dots,A\}\right) + \sum_{t=1}^{T}(\mu_{a_t^\star}(t) - \mu_{A_t}(t))\mathbb{1}\big(\mathrm{UCB}_{A_t}(t-1) \geq \mathrm{UCB}_{a_t^\star}(t-1)\big)\right] \tag{19}$$

We now introduce some high-probability event in which all the change-points are detected within a reasonable delay for all arms. With $d_a^{(\ell)} = d_a^{(\ell)}(\alpha,\delta)$ in Assumption 11, we define

$$\mathcal{E}_T = \mathcal{E}_T(\alpha,\delta) = \left(\forall a \in \{1,\dots,A\}, \forall \ell \in \{1,\dots,\mathrm{NC}_a\}, \hat{\tau}_a^{(\ell)} \in \left[\tau_a^{(\ell)}+1, \tau_a^{(\ell)} + d_a^{(\ell)}\right], \hat{\tau}_a^{(\mathrm{NC}_a+1)} > T\right).$$

From Assumption 11, as the period between two change-points are long enough, if $\mathcal{E}_T$ holds, then for all arm a and all change-point ℓ, one has $\tau_a^{(\ell)} \leq \hat{\tau}_a^{(\ell)} \leq \tau_a^{(\ell+1)}$ for all $\ell \in \{1,\dots,\mathrm{NC}_a\}$. Also, when $\mathcal{E}_T$ holds, GLR-klUCB experiences a total of C_T restarts (on different arms), which each yield an update of the exploration parameter. Letting $\hat{\sigma}^{(k)}$ be the instant of the k-th restart (on any arm) with the convention that $\hat{\sigma}^{(C_T+1)} = T$, one can write, when $\mathcal{E}_T$ holds:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}\left(t\left\lceil\frac{A}{\alpha_{k_t}}\right\rceil \in \{1,\dots,A\}\right) &\leq \sum_{k=0}^{C_T} \sum_{t=\hat{\sigma}^{(k)}+1}^{\hat{\sigma}^{(k+1)}} \mathbb{1}\left(t\left\lceil\frac{A}{\alpha_{k+1}}\right\rceil \in \{1,\dots,A\}\right) \\
&\leq \sum_{k=0}^{C_T} \alpha_{k+1}\left(\hat{\sigma}^{(k+1)} - \hat{\sigma}^{(k)}\right) \leq \alpha_{C_T+1} \sum_{k=0}^{C_T}\left(\hat{\sigma}^{(k+1)} - \hat{\sigma}^{(k)}\right) \\
&= \alpha_{C_T+1} T.
\end{aligned}$$

On $\mathcal{E}_T$, the second term in (19) can be further decomposed as follows, according to whether the upper confidence bound of the current optimal arm is small, leading to

$$\begin{aligned}
R_T \leq\ & T\mathbb{P}\left(\mathcal{E}_T^c\right) + \alpha_{C_T+1}T + \\
& \underbrace{\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^{T}\mathbb{1}\left(\mathrm{UCB}_{a_t^\star}(t) \leq \mu_{a_t^\star}(t)\right)\right]}_{(C)} + \underbrace{\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^{T}(\mu_{a_t^\star}(t) - \mu_{A_t}(t))\mathbb{1}\left(\mathrm{UCB}_{A_t}(t) \geq \mu_{a_t^\star}(t)\right)\right]}_{(D)}.
\end{aligned}$$

As in the previous proof, the conclusion follows from two lemmas. Lemma 14 controls terms (C) and (D) using some elements from the analysis of klUCB while Lemma 15 upper bounds the probability of the "bad" event $\mathcal{E}_T^c$. The proofs of these two results are presented in the next sections.

Lemma 14 *It holds that*

$$(C) \quad \leq \quad \sum_{a=1}^{A}\sum_{\ell=0}^{NC_a} d_a^{(\ell)}(\alpha,\delta) + C_T \ln(\ln(T)) \ , \tag{20}$$

$$(D) \quad \leq \quad \sum_{a=1}^{A}\sum_{\ell=0}^{NC_a} d_a^{(\ell)}(\alpha,\delta) + \sum_{a=1}^{A}\sum_{\ell=1}^{NC_a}\left[\frac{\ln(\tau_a^{(\ell+1)}-\tau_a^{(\ell)})}{\mathrm{kl}(\overline{\mu}_a^{(\ell)},\mu_a^{(\ell),*})} + O\left(\sqrt{\ln(T)}\right)\right] \ . \tag{21}$$

Lemma 15 *Under Assumption 11, it holds that* $\mathbb{P}(\mathcal{E}_T^c) \leq (A + 2\mathrm{C}_T)\delta$.

D.3 Proof of Lemma 14

Upper bound on the term (C)

$$\begin{aligned}
(C) &\leq \mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^{T}\mathbb{1}\left(n_{a_t^\star}(t)\mathrm{kl}\left(\hat{\mu}_{a_t^\star}(t),\mu_{a_t^\star}(t)\right) \geq f(t-\tau_{a_t^\star}(t))\right)\right] \\
&\leq \sum_{a=1}^{A}\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=1}^{T}\mathbb{1}(a_t^\star = a)\mathbb{1}\left(n_a(t)\mathrm{kl}\left(\hat{\mu}_a(t),\mu_a(t)\right) \geq f(t-\tau_a(t))\right)\right] \\
&\leq \sum_{a=1}^{A}\sum_{\ell=0}^{\mathrm{NC}_a}\mathbb{E}\left[\mathbb{1}(\mathcal{E}_T)\sum_{t=\tau_a^{(\ell)}+1}^{\tau_a^{(\ell+1)}}\mathbb{1}\left(n_a(t)\mathrm{kl}\left(\hat{\mu}_a(t),\bar{\mu}_a^{(\ell)}\right) \geq f(t-\hat{\tau}_a(t))\right)\right] \\
&\leq \sum_{a=1}^{A}\sum_{\ell=0}^{\mathrm{NC}_a} d_a^{(\ell)}(\alpha,\delta) + \sum_{a=1}^{A}\sum_{\ell=0}^{\mathrm{NC}_a}\mathbb{E}\left[\mathbb{1}(\mathcal{C}_a^{(\ell)})\sum_{t=\hat{\tau}_a^{(\ell)}+1}^{\tau_a^{(\ell+1)}}\mathbb{1}\left(n_a(t)\mathrm{kl}\left(\hat{\mu}_a(t),\overline{\mu}_a^{(\ell)}\right) \geq f(t-\hat{\tau}_a^{(\ell)})\right)\right],
\end{aligned}$$

where we introduce the event $\mathcal{C}_a^{(\ell)}$ that all the change-points on arm a up to the ℓ-th have been detected:

$$\mathcal{C}_a^{(\ell)} = \left\{\forall j \leq \ell, \hat{\tau}_a^{(j)} \in \left[\tau_a^{(j)}+1, \tau_a^{(j)}+d_a^{(\ell)}\right]\right\}. \tag{22}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}(\mathcal{C}_a^{(\ell)})\sum_{t=\hat{\tau}_i^{(\ell)}+1}^{\tau_a^{(\ell+1)}}\mathbb{1}\left(n_a(t)\mathrm{kl}\left(\hat{\mu}_a(t),\bar{\mu}_a^{(\ell)}\right) \geq f(t-\hat{\tau}_i^{(\ell)})\right)\middle| \mathcal{F}_{\hat{\tau}_a^{(\ell)}}\right] \\
&= \mathbb{E}\left[\mathbb{1}(\mathcal{C}_a^{(\ell)})\sum_{t=\hat{\tau}_i^{(\ell)}+1}^{\tau_a^{(\ell+1)}}\mathbb{E}\left[\mathbb{1}\left(n_a(t)\mathrm{kl}\left(\hat{\mu}_a(t),\bar{\mu}_a^{(\ell)}\right) \geq f(t-\hat{\tau}_i^{(\ell)})\right) \mid \mathcal{F}_{\hat{\tau}_a^{(\ell)}}\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{1}(\mathcal{C}_a^{(\ell)})\sum_{t'=1}^{\tau_a^{(\ell+1)}-\hat{\tau}_i^{(\ell)}}\mathbb{P}\left(\exists s \leq t' : s \times \mathrm{kl}(\hat{\mu}_s,\bar{\mu}_a^{(\ell)}) \geq f(t')\right)\right] \leq 2 + \sum_{t=3}^{T}\frac{1}{t\ln(t)} \leq 3 + \ln(\ln(T)),
\end{aligned}$$

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

Upper bound on the term (D) Recall that $\mu_a^{(\ell),*}$ is defined in the statement of Theorem 12 as the smallest value of $\mu_{a_t^\star}(t)$ when arm a is sub-optimal on the interval $[\tau_a^{(\ell)}+1, \tau_a^{(\ell+1)}]$. We let $\tilde{\mu}_{a,s}^{(\ell)}$ denote the empirical mean of the first s observations of arm a made after time $t = \hat{\tau}_a^{(\ell)}+1$. To upper bound Term (D), we introduce a sum over all arms and rewrite the sum in t as a sum of consecutive intervals $[\tau_a^{(\ell)}+1, \tau_a^{(\ell+1)}]$.

$$
\begin{aligned}
(D) &\leq \sum_{a=1}^{A} \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \sum_{\ell=0}^{\mathrm{NC}_a} \sum_{t=\tau_a^{(\ell)}+1}^{\tau_a^{(\ell+1)}} \left(\mu_{a_t^\star}(t) - \overline{\mu}_a^{(\ell)}\right) \mathbb{1}\left(A_t = a, \mathrm{UCB}_a(t) \geq \mu_{a_t^\star}(t)\right)\Big] \\
&\leq \sum_{a=1}^{A} \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T) \sum_{\ell=0}^{\mathrm{NC}_a} \sum_{t=\tau_a^{(\ell)}+1}^{\tau_a^{(\ell+1)}} \left(\mu_{a_t^\star}(t) - \overline{\mu}_a^{(\ell)}\right) \mathbb{1}\left(A_t = a, \mathrm{UCB}_a(t) \geq \mu_a^{(\ell),*}\right)\Big] \\
&\leq \sum_{a=1}^{A} \sum_{\ell=0}^{\mathrm{NC}_a} \mathbb{E}\Big[\mathbb{1}(\mathcal{E}_T)\hat{\tau}_i^{(\ell)} + \mathbb{1}(\mathcal{E}_T) \sum_{t=\hat{\tau}_a^{(\ell)}+1}^{\tau_a^{(\ell+1)}} \overline{\Delta}_a^{(\ell)} \mathbb{1}\left(A_t = a, \mathrm{UCB}_a(t) \geq \mu_a^{(\ell),*}\right)\Big]
\end{aligned}
$$

with $\overline{\Delta}_a^{(\ell)} := \max_{t\in[\tau_a^{(\ell)}+1, \tau_a^{(\ell+1)}]} \left(\mu_{a_t^\star}(t) - \overline{\mu}_a^{(\ell)}\right)$. Introducing a sum over $\mathbb{1}(n_a(t) = s)$ and swapping the sums yields

$$
(D) \leq \sum_{a=1}^{A} \sum_{\ell=0}^{\mathrm{NC}_a} d_a^{(\ell)}(\alpha, \delta) + \sum_{a=1}^{A} \sum_{\ell=1}^{\mathrm{NC}_a} \overline{\Delta}_a^{(\ell)} \mathbb{E}\Big[\mathbb{1}(\mathcal{C}_a^{(\ell)}) \sum_{s=1}^{n_a(\tau_a^{(\ell+1)})} \mathbb{1}\left(s \times \mathrm{kl}(\tilde{\mu}_{a,s}^{(\ell)}, \mu_a^{(\ell),*}) \leq f(\tau_a^{(\ell+1)} - \tau_a^{(\ell)})\right)\Big].
$$

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

$$
(D) \leq \sum_{a=1}^{A} \sum_{\ell=0}^{\mathrm{NC}_a} d_a^{(\ell)}(\alpha, \delta) + \sum_{a=1}^{A} \sum_{\ell=1}^{\mathrm{NC}_a} \overline{\Delta}_i^{(\ell)} \times \sum_{s=1}^{\tau_a^{(\ell+1)} - \tau_a^{(\ell)}} \mathbb{P}\left(s \times \mathrm{kl}(\hat{\mu}_s, \mu_a^{(\ell),*}) \leq f(\tau_a^{(\ell+1)} - \tau_a^{(\ell)})\right).
$$

If $\overline{\Delta}_a^{(\ell)} \neq 0$, then $\mu_a^{(\ell),*} > \overline{\mu}_a^{(\ell)}$ and 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

$$
\sum_{s=1}^{\tau_a^{(\ell+1)} - \tau_a^{(\ell)}} \mathbb{P}\left(s \times \mathrm{kl}(\hat{\mu}_s, \mu_a^{(\ell),*}) \leq f(\tau_a^{(\ell+1)} - \tau_a^{(\ell)})\right) \leq \frac{\ln(\tau_a^{(\ell+1)} - \tau_a^{(\ell)})}{\mathrm{kl}\left(\overline{\mu}_a^{(\ell)}, \mu_a^{(\ell),*}\right)} + O\left(\sqrt{\ln(T)}\right).
$$

If $\overline{\Delta}_a^{(\ell)} = 0$, then $\mu_a^{(\ell),*} = 0$ and $\mathrm{kl}\left(\overline{\mu}_a^{(\ell)}, \mu_a^{(\ell),*}\right) = +\infty$, thus it also trivially holds that

$$
\overline{\Delta}_a^{(\ell)} \sum_{s=1}^{\tau_a^{(\ell+1)} - \tau_a^{(\ell)}} \mathbb{P}\left(s \times \mathrm{kl}(\hat{\mu}_s, \mu_a^{(\ell),*}) \leq f(\tau_a^{(\ell+1)} - \tau_a^{(\ell)})\right) \leq \frac{\ln(\tau_a^{(\ell+1)} - \tau_a^{(\ell)})}{\mathrm{kl}\left(\overline{\mu}_a^{(\ell)}, \mu_a^{(\ell),*}\right)} + O\left(\sqrt{\ln(T)}\right)
$$

and the proof of (21) is complete.

D.4 Proof of Lemma 15

With the event $\mathcal{C}_a^{(\ell)}$ defined in (22) and the convention $\tau_a^{(\mathrm{NC}_a+1)} = T$, a simple union bound yields

$$\mathbb{P}(\mathcal{E}_T^c) \leq \sum_{a=1}^{A} \sum_{\ell=1}^{\mathrm{NC}_a+1} \underbrace{\mathbb{P}\left(\hat{\tau}_a^{(\ell)} \leq \tau_a^{(\ell)} \;\middle|\; \mathcal{C}_a^{(\ell-1)}\right)}_{(a)} + \sum_{i=1}^{A} \sum_{\ell=1}^{\mathrm{NC}_a} \underbrace{\mathbb{P}\left(\hat{\tau}_a^{(\ell)} \geq \tau_a^{(\ell)} + d_a^{(\ell)} \;\middle|\; \mathcal{C}_a^{(\ell-1)}\right)}_{(b)}.$$

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 $\hat{\tau}_a^{(\ell)} \leq \tau_a^{(\ell)}$ implies that there is a false alarm for the detection of the ℓ-th change on arm a, which is not likely:

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

with $\hat{\mu}_{s:s'} = \sum_{r=s}^{s'} Z_{i,r}$ where $Z_{i,r}$ is an *i.i.d.* sequence with mean $\mu_a^{(\ell-1)}$. Indeed, conditionally to $\mathcal{C}_a^{(\ell-1)}$, the $n_a(\tau_a^{(\ell)})$ successive observations of arm a arm starting from $\hat{\tau}_a^{(\ell)}$ are *i.i.d.* with mean $\mu_a^{(\ell-1)}$. Using Lemma 9, term (a) is upper bounded by δ.

Upper bound on term (b): controlling the delay Assume that $\mathcal{C}_a^{(\ell-1)}$ holds. For $t > \hat{\tau}_a^{(\ell-1)}$, we now that at least the first $\ell-1$ change-points on arm a have been detected, hence the current number of episodes k_t is larger than ℓ. It follows from Proposition 3 that there exists $\bar{t} \in \left\{\tau_a^{(\ell)}, \ldots, \tau_a^{(\ell)} + d_a^{(\ell)}\right\}$ such that $n_a(\bar{t}) - n_a(\tau_a^{(\ell)}) = \bar{r}$ where $\bar{r} = \lfloor \frac{\alpha_\ell}{A} d_a^{(\ell)} \rfloor$. This is because the mapping $t \mapsto n_a(t) - n_a(\tau_a^{(\ell)})$ is non-decreasing, is 0 at $t = \tau_a^{(\ell)}$ and its value at $\tau_a^{(\ell)} + d_a^{(\ell)}$ is larger than $\bar{r}$ by Proposition 3 as $\alpha_{k_t} \geq \alpha_\ell$. Using that

$$(\hat{\tau}_a^{(\ell)} \geq \tau_a^{(\ell)} + d_a^{(\ell)}) \cap \mathcal{C}_a^{(\ell-1)} \subseteq (\hat{\tau}_a^{(\ell)} \geq \bar{t}) \cap \mathcal{C}_a^{(\ell-1)},$$

the probability $(b) = \mathbb{P}(\hat{\tau}_a^{(\ell)} \geq \tau_a^{(\ell)} + d_a^{(\ell)} | \mathcal{C}_a^{(\ell-1)})$ is further upper bounded as follows:

$$(b) \leq \mathbb{P}\left(n_a(\tau_a^{(\ell)}) \,\mathrm{kl}\left(\tilde{\mu}_{i,n_a(\tau_a^{(\ell)})}^{\ell-1}, \tilde{\mu}_{i,n_a(\bar{t})}^{\ell-1}\right) + \bar{r}\,\mathrm{kl}\left(\tilde{\mu}_{i,n_a(\tau_a^{(\ell)}):n_a(\bar{t})}^{\ell-1}, \tilde{\mu}_{i,n_a(\bar{t})}^{\ell-1}\right) \leq \beta(n_a(\tau_a^{(\ell)}) + \bar{r}, \delta) \middle| \mathcal{C}_a^{(\ell-1)}\right),$$

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

Then, conditionally to $\mathcal{C}_a^{(\ell-1)}$, $n_a(\tau_a^{(\ell)}) \in \left\{\left\lfloor \frac{\alpha_\ell}{A}\left(\tau_a^{(\ell)} - \hat{\tau}_a^{(\ell-1)}\right)\right\rfloor, \ldots, \tau_a^{(\ell)} - \hat{\tau}_a^{(\ell-1)}\right\}$ due to Proposition 3 and to the fact that $d_a^{(\ell-1)} \leq (\tau_a^{(\ell)} - \tau_a^{(\ell-1)})/2$ by Assumption 11. Hence

$$\begin{aligned} n_a(\tau_a^{(\ell)}) &\in \left\{\left\lfloor \frac{\alpha_\ell}{A}\left(\tau_a^{(\ell)} - \tau_a^{(\ell-1)} - d_a^{(\ell-1)}\right)\right\rfloor, \ldots, \left(\tau_a^{(\ell)} - \tau_a^{(\ell-1)}\right)\right\} \\ n_a(\tau_a^{(\ell)}) &\in \left\{\left\lfloor \frac{\alpha_\ell}{2A}\left(\tau_a^{(\ell)} - \tau_a^{(\ell-1)}\right)\right\rfloor, \ldots, \left(\tau_a^{(\ell)} - \tau_a^{(\ell-1)}\right)\right\} := \mathcal{I}_\ell. \end{aligned}$$

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

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

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

Using Pinsker's inequality and the expression of the gap $\Delta_a^{c,(\ell)} = \bar{\mu}_a^{(\ell-1)} - \bar{\mu}_a^{(\ell)}$, one can write

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

Using Lemma 10 stated in Appendix B.2 and a union bound, the first term in the right hand side is upper bounded by δ (as $\beta(r+s,\delta) \geq \beta(s,\delta) \geq \ln(3s\sqrt{s}/\delta)$). For the second term, we use the observation

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

and finally get

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

Let $s_{\min} = \left\lfloor \frac{\alpha_\ell}{A}(\tau_a^{(\ell)} - \tau_a^{(\ell-1)})/2 \right\rfloor$. 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 can further write

$$\begin{aligned}
(b) &\leq \delta + \mathbb{P}\left(\exists s \in \mathcal{I}_\ell : \left(\Delta_a^{(\ell)}\right)^2 \leq 2\frac{s_{\min}+\overline{r}}{s_{\min}\overline{r}}\beta(\tau_a^{(\ell)} - \tau_a^{(\ell-1)} + \overline{r},\delta)\right) \\
&\leq \delta + \mathbb{P}\left(\exists s \in \mathcal{I}_\ell : \left(\Delta_a^{(\ell)}\right)^2 \leq \frac{4}{\overline{r}}\beta\left(\frac{3}{2}(\tau_a^{(\ell)} - \tau_a^{(\ell-1)}),\delta\right)\right), \qquad (23)
\end{aligned}$$

where in the last step we use that by Assumption 11, it holds that $\overline{r} \leq s_{\min} \leq (\tau_a^{(\ell)} - \tau_a^{(\ell-1)})/2$. To conclude the proof, it remains to observe that by definition of the delay $d_a^{(\ell)}$,

$$\overline{r} = \left\lfloor \frac{\alpha_\ell}{A} d_a^{(\ell)} \right\rfloor > \frac{4}{\left(\Delta_a^{c,(\ell)}\right)^2}\beta\left(\frac{3}{2}(\tau_a^{(\ell)} - \tau_a^{(\ell-1)}),\delta\right)$$

hence the probability in the right hand side of (23) is equal to zero, which yields $(b) \leq \delta$.