Bayesian Change-Point Detection for Bandit Feedback in Non-stationary Environments

Reda Alami REDA.ALAMI@TII.AE
Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, United Arab Emirates

Editors: Emtiyaz Khan and Mehmet Gonen

Abstract

The stochastic multi-armed bandit problem has been widely studied under the stationary assumption. However in real world problems and industrial applications, this assumption is often unrealistic because the distributions of rewards may change over time. In this paper, we consider the piece-wise iid non-stationary stochastic multi-armed bandit problem with unknown change-points and we focus on the change of mean setup. To solve the latter, we propose a change-point based framework where we study a class of change-detection based optimal bandit policies that actively detects change-point using the restarted Bayesian online change-point detector and then restarts the bandit indices. Analytically, in the context of regret minimization, our proposal achieves a $\mathcal{O}(\sqrt{ATK_T})$ regret upper-bound where K_T is the overall number of change-points up to the horizon T and A is the number of arms. The derived bound matches the existing lower bound for abruptly changing environments. Finally, we demonstrate the cumulative regret reduction of the our proposal over synthetic Bernoulli rewards as well as Yahoo! datasets of webpage click-through rates.

Keywords: Non-stationary multi armed bandits, Bayesian Online Change-point detection.

1. Introduction and related work

Multi-Armed Bandit (MAB) problems model sequential allocation in the face of uncertainty and partial feedback on rewards. At each round, the learning agent (decision-maker) resolves to pull one arm amongst a finite number of possible arms. This decision is based on the past observations. At each time t, upon selecting arm $A_t \in \{1, ...A\}$, the agent receives a reward $X_{A_t,t}$, and he aims at building a sequential sampling strategy that maximizes the expected sum of these rewards. This is equivalent to minimizing the regret, defined as the difference between the total reward of the oracle strategy always selecting the arm with largest mean, and that of the agent strategy. The multi-armed bandit problem has been extensively applied in several domains such as communication systems (Thompson (1933)), online recommendation systems (Li et al. (2012)), online advertisement campaign (Schwartz et al. (2017)) and clinical trials (Villar et al. (2015)).

The stationary stochastic multi-armed bandit problem has been well-studied since the work of Lai and Robbins (1985). In the context of regret minimization, several algorithms with $\mathcal{O}(\log T)$ problem-dependent regret upper bound have been proposed `UCB 1` (Auer et al. (2002)), `UCB V` (Audibert et al. (2007)), `CP UCB` (Garivier and Cappé (2011)), Bayes `UCB` (Kaufmann et al. (2012a)), `KL UCB` (Cappé et al. (2013)), `DMED` (Honda and Takemura (2010)), `MOSS` (Audibert and Bubeck (2009)), Thompson Sampling (Korda et al. (2013)) and Maillard Sampling (Bian and Jun (2022)). However these algorithms perform poorly in non-stationary environments where the distributions of rewards change over time. To address this issue, the non-stationary multi-armed bandit problem has been proposed in the literature. Essentially, there are two kinds of strategies for the non-stationary multi-armed bandit: passively adaptive policies (Besbes et al. (2014); Wei et al. (2016)) and actively adaptive policies (Hartland et al. (2006); Mellor and Shapiro (2013a)).

Passively adaptive policy In order to forget the past rewards, the first passively adaptive strategies propose to penalize the past rewards by multiplying them with a discount factor $\gamma \in \big(0,1\big)$ such that the penalization is of γ^s if the arm was not seen since s time steps. The Discounted UCB (`D-UCB`) was first proposed by Kocsis and Szepesvári (2006) and then it has been analyzed by Garivier and Moulines (2011) where they prove a regret upper bound of $\mathcal{O}(\sqrt{K_T T}\log(T))$ if the discount factor $\gamma = 1 - \sqrt{K_T/T}/4$ where K_T is the overall number of change-point up to the horizon T. Another popular mechanism to forget the past rewards is to use a sliding window of fixed size τ, where only the τ last rewards are used for the decision-maker. The sliding Window UCB (`SW-UCB`) has been analysed by Garivier and Moulines (2011) who demonstrates a regret upper bound of $\mathcal{O}(\sqrt{K_T T \log(T)})$ in the case where $\tau = 2\sqrt{T\log(T)/K_T}$. There are also other recent algorithms such as Discounted Thompson Sampling (Raj and Kalyani (2017)), Thompson Sampling with sliding window (Trovo et al. (2020)) and `REXP3` (Besbes et al. (2014)) that use passively adaptive mechanisms.

Actively adaptive policy There is a large literature exploring the idea of monitoring the change in the reward distribution via online change-point detection and triggering the reset of the bandit algorithm. This kind of algorithm aims at localizing the change-point and hence demonstrate better performances than the passive policies. The `Adapte-EvE` algorithm (Hartland et al. (2006)) uses the Page-Hinkley test to detect the change-point and hence restart the UCB1 strategy once an alarm is raised. Then, in Mellor and Shapiro (2013a), the authors design the switching Thompson sampling strategy (`STS`): a combination between the Bayesian online change-point detector Adams and MacKay (2007) and Thompson Sampling. This work has been revisited in Alami et al. (2016) by adding an extra expert aggregation step. A recent and related work Liu et al. (2018) uses CUSUM algorithm for change-point detection. Furthermore, the Monitored UCB algorithm (`M-UCB` Cao et al. (2018)) also combines a CUSUM instance with UCB. However, the change-detection test is much easier and a forced exploration phase is also performed. Moreover, in Besson and Kaufmann (2019b) the authors propose a hybrid combination between KL-UCB algorithm and a Bernoulli Generalized Likelihood Ratio Test for change-point detection. They reach a $\mathcal{O}(K_T\sqrt{T\log(T)})$ as regret upper-bound. There is also the work on combining a GLR instance with the UCB algorithm. Indeed, the authors of Mukherjee and Maillard (2019) have derived a $\mathcal{O}(\log(T))$ regret bounds. Furthermore, in Auer et al. (2019) the authors propose `ADSWITCH` an adaptively tracking algorithm for the best arm with an unknown number of change-point. It has been shown that `ADSWITCH` achieves (nearly) optimal mini-max regret bounds of $\mathcal{O}(\sqrt{ATK_T})$. Finally, in Gopalan et al. (2021) the authors propose a bandit quickest change-point detection framework where they have designed an ε-greedy changepoint detection

Other non-stationary bandits in the literature In the literature of the non-stationary bandit, we essentially find the work of Chen et al. (2019) where the authors propose the first contextual bandit algorithm that is parameter-free, efficient, and optimal in terms of dynamic regret. Specifically, our algorithm achieves $\mathcal{O}\left(\min\left\{\sqrt{AST}, A^{\frac{1}{3}}\Delta^{\frac{1}{3}}T^{\frac{2}{3}}\right\}\right)$ dynamic regret for a contextual bandit problem with T rounds, A actions, S switches and Δ total variation in data distributions. Moreover, in the context of linear bandits, the authors of Zhao et al. (2020) have investigated the problem of non-stationary linear bandits, where the unknown regression parameter is evolving over time. They designed an UCB-type algorithm to balance exploitation and exploration, and restart it periodically to handle the drift of unknown parameters. Finally, in the context of rested rotting bandits where the reward of an action decreases every time it is pulled, the authors in Seznec et al. (2019) propose a nearly optimal algorithm for this setting called the filtering on expanding window average (`FEWA`) algorithm that constructs moving averages of increasing windows to identify arms that are more likely to return high rewards when pulled once more. Also, in the authors introduce a novel algorithm called Routing Adaptive Windows UCB (`RAW-UCB`) to address both the rested and restless bandit for all types of non-stationary environments.

Contributions and outline In this paper, we propose a new framework for piece-wise stationary bandit which consists on a combination between any multi-armed bandit algorithm with the restarted Bayesian online change-point detector Alami et al. (2020). In the case of Bernoulli rewards, we derive a regret upper bound for the framework applied on the Thompson sampling strategy which is of the order $\mathcal{O}(\sqrt{ATK_T})$ for a known number of change-point K_T. This upper bound matches the actual lower bound stated in Garivier and Moulines (2011). Finally, we conduct experiments highlighting the performance of the proposed and existing strategies are validated by both synthetic and real world datasets, and we show that our proposed algorithm is superior to other existing policies in terms of pseudo cumulative regret.

The remainder of the paper is organized as follows: we describe the piece-wise stationary bandit model in Section 2. In Section 3, we describe the framework of Bayesian Change-point Detection for bandit feedback in Bernoulli environment. Then, in section 4 we provide the regret upper bound analysis of the framework applied to the Thompson sampling strategy. Then, we demonstrate experiment results in Section 5. Finally, section 6 concludes the paper. Due to space limitations, we provide the proofs of the analytical results in the appendices.

2. The piece-wise stationary multi-armed bandit problem

A piece-wise stationary multi-armed bandit is a discrete time stochastic control process defined by a 3-tuple $\left(\mathbf{A}, \mathbf{T}, \{\mathcal{F}(\mu_{a,t})\}_{a\in\mathbf{A}, t\in\mathbf{T}}\right)$ where $\mathbf{A} = \{1, ..., A\}$ denotes the discrete set of actions of size A, $\mathbf{T} = \{1, 2, ..., T\}$ a sequence of time-steps going up to the horizon T and $\mathcal{F}(\mu_{a,t})$ the reward probability distribution of arm a at time t (probability density function) whose mean is $\mu_{a,t}$. We assume a local switching model that allows asynchronous changes to happen, i.e. arm switches are independent. We denote the overall number of break-points up to the horizon T by $K_T = \sum_{t=2}^{T} \mathbb{I}\left\{\exists a \in \mathbf{A} : \mu_{a,t} \neq \mu_{a,t-1}\right\} + 1$, where $\mathbb{I}\left\{\bullet\right\}$ denotes the indicator function. Then, we denote the sequence of break-points up to the horizon T by: $(\tau_1 = 1, \tau_2, ..., \tau_{K_T+1} = T + 1)$.

Note that when a breakpoint occurs, we do not assume that all the arms means change, but that there exists an arm which experiences a changepoint, i.e. whose mean satisfies $\mu_{a,t} \neq \mu_{a,t+1}$. Letting C_T denote the total number of changepoints before horizon T, we have

$C_T \in \{K_T, ..., AK_T\}$. By this way, we shall denote by $\tau_{a,k}$ the kth change-point experienced by arm a.

Note that an instant of break-point τ_k corresponds to one or several change-points.

Following this, the environment is now described by K_T piece-wise stationary segment denoted by $\mathcal{T}_k = [\tau_k, \tau_{k+1})$. Then, it is convenient to use the variable $\theta_{a,[k]}$ to denote the constant behavior of $\mu_{a,t}$ for $t \in [\tau_k, \tau_{k+1})$. Moreover, we denote by $\Theta_{[k]} = \left(\mathcal{F}\left(\theta_{1,[k]}\right), ..., \mathcal{F}\left(\theta_{A,[k]}\right)\right)$ the stationary multi-armed bandit on epoch $\mathcal{T}_k$. By the way, a piece-wise stationary bandit is ultimately only a sequence of K_T stationary bandit denoted by $\left(\Theta_{[1]}, ..., \Theta_{[K_T]}\right)$.

A decision maker will sequentially interact with this piece-wise stationary bandit for T times. At each round $t \geqslant 1$, he has to select an arm $A_t \in \mathbf{A}$ based on past observations and receive the corresponding reward $X_{A_t,t} \sim \mathcal{F}\left(\mu_{A_t,t}\right)$. At time t, we let $a_t^\star = \operatorname{argmax}_{a\in\mathbf{A}} \mu_{a,t}$ denotes the optimal arm. For convenience, we will be interested in the optimal arm during the stationary epoch $\mathcal{T}_k$ which we shall denote by $a_{[k]}^\star = \operatorname{argmax}_{a\in\mathbf{A}} \theta_{a,[k]}$. Also, the optimal mean reward on epoch $\mathcal{T}_k$ is denoted by $\theta_{[k]}^\star$. Thus, the *bandit gap* of arm a during epoch $\mathcal{T}_k$ is $\Delta_{a,[k]} = \theta_{[k]}^\star - \theta_{a,[k]}$. Finally, the change magnitude of arm a related to the change-point τ_k is $\Lambda_{a,[k]} = \left|\theta_{a,[k]} - \theta_{a,[k-1]}\right|$.

In addition, we make the following three assumptions for tractability.

Assumption 1 (Bernoulli rewards)**.** *The distributions of all the arms are Bernoulli distributions denoted as* $\mathcal{B}\left(\mu_{a,t}\right) \forall a \in \mathbf{A}, \forall t \in \mathbf{T}$.

Assumption 1 has been widely used in the literature e.g. in Kaufmann et al. (2012b); Mellor and Shapiro (2013b); Besbes et al. (2014). Moreover, working on the Bernoulli distributions is not as restrictive as it may seem. On the first hand, from a concentration point of view, Bernoulli distributions can be seen as a worst case of bounded distributions. Furthermore, Bernoulli distributions are crucially used in many widespread applications of machine learning, for instance in modelling the collisions in cognitive radio, in monitoring the performances of statistical models, in monitoring events in probes for network supervision, in the multi armed bandit problem and finally in experiments in clinical trials and recommender systems.

Assumption 2 (Abrupt switching environments)**.** *There exists a sequence* $(\gamma_1, \gamma_2, ..., \gamma_A) \in (0,1)^A$*, such that the parameter* $\mu_{a,t}$ *follows an abrupt switching behavior driven by the hazard rate* γ_a*:*

$$\mu_{a,t} = \begin{cases} \mu_{a,t-1} & \textit{with probability } 1-\gamma_a \\ \mu_{\text{new}} \in [0,1] & \textit{with probability } \gamma_a \end{cases} \tag{1}$$

Assumption 2 is similar to the one used in Mellor and Shapiro (2013a) and Garivier and Moulines (2011). Moreover, we assume that the hazard rate γ_a is small in the sense that we have the possibility to collect enough samples between two consecutive change-points in order to well estimate the mean of each arm.

Assumption 3 (Change-point detectability)**.** *There exists a threshold* $\lambda > 0$ *such that* $\forall a \in \mathbf{A}$ *and* $\forall t \in \mathbf{T}$*, if* $\mu_{a,t} \neq \mu_{a,t+1}$ *then* $|\mu_{a,t} - \mu_{a,t+1}| \geqslant \lambda$*.*

Assumption 3 excludes infinitesimal mean change, which is reasonable in real world application when detecting abrupt changes bounded from below by a certain threshold.

Moreover, one should note that Assumption 2 and Assumption 3 characterise the hardness of a non-stationary multi armed bandit problem. Indeed, the higher the switching rate γ_a, the harder the detection of change related to arm a. Furthermore, the tighter the threshold λ, the longer the detection of the change.

Regret minimization in a piece-wise stationary model The agent's objective is to build a policy π in order to maximize its expected cumulative reward during T consecutive time steps, i.e. $\max \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t,t}\right]$, which is equivalent to minimizing its T-step pseudo cumulative regret $\mathcal{R}_T$ defined as:

$$\mathcal{R}_T^{\pi} = \sum_{t=1}^{T} \max_{a\in\mathbf{A}} \mathbb{E}\left[X_{a,t}\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t,t}\right] = \sum_{t=1}^{T} \left(\mu_t^{\star} - \mu_{A_t,t}\right)$$

Following Assumption 1, the quantity $\mathcal{R}_T^{\pi}$ is upper bounded as: $\mathcal{R}_T^{\pi} \leqslant \sum_{a\in\mathbf{A}} \mathbb{E}\left[\overline{N}_{a,T}\right]$ where $\overline{N}_{a,T} = \sum_{t=1}^{T} \mathbb{I}\left\{A_t = a \text{ and } a \neq a_t^{\star}\right\}$ denotes the number of draws related to arm a when it is considered as sub-optimal arm.

3. The framework of Bayesian Change-point Detection for Bandit Feedback in Bernoulli environment

The Bayesian change-point for bandit feedback framework consists of two main components: an optimal bandit algorithm and the restarted Bayesian online change-point detector (`RBOCPD`) (Alami et al. (2020)). At each round t and based on the past observations, the bandit outputs a decision $A_t \in \mathbf{A}$. By playing action A_t, the environment reveals a reward $X_{A_t,t} \sim \mathcal{B}\left(\mu_{A_t,t}\right)$ which is observed by both the bandit algorithm and the `RBOCPD` instance. The sequential change-point detector which monitors the distribution of each arm either sends a positive signal to restart the estimated parameters related the played arm A_t when a change-point is detected or sends a negative signal when no change is observed.

The `RBOCPD` algorithm is chosen among all the sequential change-point detector algorithms in the state of the art for three main reasons.

- *Well adaptability to unknown priors.* Indeed, the `RBOCPD` algorithm has been designed to solve the problem of sequential change-point detection in a setting where both the change-points and the distributions before and after the change are assumed to be unknown. This setting corresponds exactly to the situation of an agent facing a multi armed bandit whose distributions are unknown and may change abruptly at some unknown instants.
- *Minimum detection delay.* This corresponds to the first criteria assessing the performance of a sequential change-point detector. The detection delay is defined as the number of samples needed to detect a change. In Alami et al. (2020), the authors have shown that the detection delay of the `RBOCPD` strategy is asymptotically optimal in the sense that it reaches the existing lower bound stated in Theorem 3.1 in Lai and Xing (2010).

- *Well controlled false alarm rate.* The false alarm rate corresponds to the probability of detecting a change at some instant where there is no change. Again, in Alami et al. (2020), the authors have demonstrated that $\forall \delta \in (0,1)$ `RBOCPD` doesn't make any false alarm with a probability at least $1-\delta$.

In the following, we briefly describe the subroutine bandit used in the framework as well as the restarted Bayesian change-point detector strategy.

3.1. Subroutine bandit for the stationary environment

A subroutine bandit denoted as `BANDIT` is a policy that takes at each time step t, the number of times $N_{a,t}$ arm a has been pulled since $t=0$ and the actual success counter $S_{a,t} = \sum_{s=1}^{t} \mathbb{I}\big\{X_{a,s}=1\big\}$ in order to compute the index $\mathfrak{I}_{a,t}^{\texttt{BANDIT}}$ of arm a at time t. The bandit chooses to pull the arm $A_t = \arg\max\limits_{a\in\mathbf{A}} \mathfrak{I}_{a,t}^{\texttt{BANDIT}}$ whose index is the highest one. The computation of the arm index is usually an exploration-exploitation dilemma implementation that takes either the form of a posterior distribution sampling Kaufmann et al. (2012a,b) or an upper confidence bound computation Auer et al. (2002); Garivier and Cappé (2011). For instance, in the Thompson sampling (TS) strategy Kaufmann et al. (2012b), the index of arm a at time t denoted as $\mathfrak{I}_{a,t}^{\mathrm{TS}} \sim \mathrm{Beta}\left(S_{a,t}+s_0, N_{a,t}-S_{a,t}+f_0\right)$ is a sample from the posterior Beta distribution of the arm where $s_0>0, f_0>0$ are the prior hyperparameters for arm a. For the Bayes UCB strategy Kaufmann et al. (2012a), the index of arm a at time t denoted as $\mathfrak{I}_{a,t}^{\mathrm{BayesUCB}} = \mathbf{Q}\Big(1-\frac{1}{(t\log t)^c}, \mathrm{Beta}\left(S_{a,t}+s_0, N_{a,t}-S_{a,t}+f_0\right)\Big)$ is the quantile of order $(t\log t)^c$ of the posterior Beta distribution related to arm a, for some constant $c \geqslant 1$.

3.2. The restarted Bayesian online change-point detector

The authors in Alami et al. (2020) have designed a variant of the original Bayesian online change-point detector introduced by Adams and MacKay (2007). The resulting strategy is named restarted Bayesian online change-point detector `RBOCPD`. It is a pruning version of the original algorithm reinterpreted from the standpoint of forecasters aggregation and expressed as a restart procedure pruning the useless forecasters.

More formally, for a binary sequence $\big(x_r,...,x_n\big) \in \{0,1\}$, the final formulation of the `RBOCPD` strategy takes the following form:

$$\texttt{RBOCPD\_Restart}(x_r,...,x_t) = \mathbb{I}\big\{\exists s \in (r,t] : \vartheta_{r,s,t} > \vartheta_{r,r,t}\big\} \tag{2}$$

where the weight of the forecasters $\vartheta_{r,s,t}$ are computed in a recursive way as follows (assuming an initial weight $\vartheta_{r,1,1}=1$):

$$\vartheta_{r,s,t} = \begin{cases} \frac{\eta_{r,s,t}}{\eta_{r,s,t-1}} \exp\left(-l_{s,t}\right) \vartheta_{r,s,t-1} & \forall s<t, \\ \eta_{r,t,t} \times \mathcal{V}_{r:t} & s=t\,. \end{cases} \tag{3}$$

such that the initial weight of the forecaster takes the form of $\mathcal{V}_{r:t} := \exp\Big(-\sum_{s'=r}^{t-1} l_{s',t-1}\Big)$ and the instantaneous loss $l_{s,t} := -\log \mathtt{Lp}\left(x_t|x_s...x_{t-1}\right)$ is computed based on the Laplace predictor $\mathtt{Lp}\left(x_t|x_s...x_{t-1}\right) := \begin{cases} \frac{\sum_{i=s}^{t-1} x_i+1}{t-s+2} & \text{if } x_t=1 \\ \frac{\sum_{i=s}^{t-1}(1-x_i)+1}{t-s+2} & \text{if } x_t=0 \end{cases}$. The hyper-parameter $\eta_{r,s,t}$ is tuned as a decreasing function in t and depends also on the probability of false alarm δ.

3.3. Application of the framework

In order to resolve a piece-wise stationary multi-armed bandit, we propose the Bayesian Change-Point Detection for bandit framework BAYESIAN-CPD-BANDIT, that combines any multi-armed bandit algorithm (BANDIT) with the restarted Bayesian online change-point detector (RBOCPD) running on each arm $a \in \mathbf{A}$. At some time t, BAYESIAN-CPD-BANDIT re-initializes the parameters related to arm A_t when the BOCPD associated to arm A_t has raised an alarm.

Forced exploration In the majority of cases where the environment is described by several change-points, these change-point can affect sub-sampled arms. Thus, for local changes, it is not enough to combine (even) an optimal bandit algorithm with an optimal online change point detector strategy like RBOCPD. A third ingredient is requested. It is a question of adding some forced exploration parameterized by $\alpha \in (0, 1)$ to ensure each arm is sampled enough and changes can also be detected on arms currently under-sampled by the bandit algorithm. By this way, the bandit will play the arm whose current index is maximal with high probability or sample uniformly the arms set with low probability.

We formally state the BAYESIAN-CPD-BANDIT framework for the Bernoulli case in Algorithm 1 and for the simplicity of notations we adopt the following useful notations.

Notations 1. *Let $\tau_a(t)$ denotes the last restart related to arm a that happened before time t. Then, let $N_{a,t} = \sum_{i=\tau_a(t)}^{t} \mathbb{I}\{A_s = a\}$ denotes the number of time arm a has been drawn from the last restart until the current time t. For convenience, we shall use $Y_{a,N_{a,t}}$: a re-shifted version of the observation $X_{a,t}$.*

Algorithm 1 Bayesian Change-Point Detection for Bandit feedback (`BAYESIAN-CPD-BANDIT`)

Require: $\mathbf{A}$: Arm set, BANDIT: Multi-Armed Bandit strategy as subroutine, $\alpha \in (0,1)$: forced exploration rate, $s_0 > 0, n_0 > 0$: parameters for initialization, T: Horizon.

1: **Initialization:**

$\forall a \in \mathbf{A} \quad N_{a,0} = n_0$ and $S_{a,0} = s_0$

2: **Define:**

$$\begin{aligned} &\forall a \in \mathbf{A}, \forall t \in \mathbf{T}: \quad \mathfrak{I}_{a,t}^{\text{BANDIT}} \text{ is defined following the MAB strategy BANDIT.} \\ &\text{(it can be a Thompson Sampling or Bayes UCB strategy)} \end{aligned} \tag{4}$$

3: **For** $t = 1, \ldots, T$

4: Choose action $A_t = \begin{cases} \operatorname{argmax}_a \mathfrak{I}_{a,t}^{\text{BANDIT}} & \text{with probability } 1-\alpha \\ a & \forall a \in \mathbf{A} \text{ with probability } \frac{\alpha}{A} \end{cases}$.

5: Observe $X_{A_t,t} \sim \mathcal{B}\left(\mu_{A_t,t}\right)$.

6: Re-shift observation $Y_{A_t,N_{A_t,t}} = X_{A_t,t}$.

7: Update $N_{A_t,t+1} = N_{A_t,t} + 1$ and $S_{A_t,t+1} = S_{A_t,t} + X_{A_t,t}$.

8: Perform change-point detection using `RBOCPD` on the sequence $\left(Y_{A_t,1}, ..., Y_{A_t,N_{A_t,t}}\right)$.

9: **If** `RBOCPD_Restart`$(Y_{A_t,1}, ..., Y_{A_t,N_{A_t,t}}) = 1$ **then** $N_{A_t,t+1} = n_0$ and $S_{A_t,t+1} = s_0$

10: Update $\mathfrak{I}_{A_t,t+1}^{\text{BANDIT}}$ according to Eq.(4).

4. Performance Analysis

In this section, we provide a mathematical analysis of the regret upper bound related to the application of the framework on the Thompson sampling algorithm as bandit. The analysed strategy is by the way called `BAYESIAN-CPD-TS`. First, in Theorem 1 we start by upper bounding the expected number of pulls related to arm $a \in \mathbf{A}$ when acting as sub-optimal arm. To do so, we introduce the quantity $\mathbb{E}\left[F_T\right]$ which denotes the expected number of false alarm raised up to horizon T. We also introduce the quantity $\mathbb{E}\left[D_{a,k}\right]$ denoting the expected detection delay related to the change-point $\tau_{a,k}$. We also denote by $\mathrm{NC}_{a,T} := \sum_{t=1}^{T-1} \mathbb{I}\left\{\mu_{a,t} \neq \mu_{a,t+1}\right\}$. Then, in Theorem 2 we state the upper bound control regarding the expected number of the false alarms and the expected detection delay. Finally, we combine the results of Theorem 1 and Theorem 2 to state the regret upper bound of the `BAYESIAN-CPD-TS` strategy. Due to space limitations, the proofs are presented in the supplementary material.

Theorem 1 (Bounding the number of samples related to sub-optimal arms)**.** *Under Assumptions 1 and 3, for any $\alpha \in (0,1)$ and any arm $a \in \mathbf{A}$, the* `BAYESIAN-CPD-TS` *strategy* *achieves:*

$$\forall \varepsilon \in [0,1], \exists C(\theta^\star_{[1]}, \theta_{a,[1]}, ..., \theta^\star_{[K_T]}, \theta_{a,[K_T]}) > 0 :$$
$$\mathbb{E}\left[\overline{N}_{a,T}\right] \leqslant \frac{\alpha T}{A} + \sum_{k=1}^{NC_{a,T}} \mathbb{E}\left[D_{a,k}\right] + \left(NC_{a,T} + \mathbb{E}\left[F_T\right]\right) \times (1+\varepsilon) \times \frac{\log T + \log\log T}{\min\limits_{k\in[1,K_T], a\neq a^\star_k} \boldsymbol{kl}\left(\theta_{a,[k]}, \theta^\star_{[k]}\right)} + C$$

where $\boldsymbol{kl}(\bullet,\bullet)$ *stands for the Kullback-Leibler divergence for Bernoulli distributions.*

Remark 1. *The problem dependant constant* $C\left(\theta^\star_{[1]}, \theta_{a,[1]}, ..., \theta^\star_{[K_T]}, \theta_{a,[K_T]}\right)$ *comes directly from the analysis of the Thompson sampling in Kaufmann et al. (2012b).*

Theorem 2 (False alarm and detection delay control)**.** *Under Assumptions 1 and 2 and for some* $\delta \in (0,1)$*, the control of the expected number of false alarm* $\mathbb{E}[F_T]$ *as well as the expected detection delays* $\{\mathbb{E}[D_{a,k}], k \in [1, NC_{a,T}]\}$ *take the following form.*

$$\forall \delta \in (0,1) \quad \mathbb{E}\left[F_T\right] \leqslant \delta \quad \text{and} \quad \forall k \in [1, NC_{a,T}] \;\; \mathbb{E}\left[D_{a,k}\right] = \mathcal{O}\left(\frac{o\left(\log \frac{K_T}{\delta}\right)}{2\alpha \times \min\limits_{a:\Lambda_{a,[k]}\neq 0} \Lambda^2_{a,[k]}}\right)$$

Corollary 1 (Regret upper bound for a known number of change-points)**.** *Under Assumption 1 and 2, assuming that the horizon* T *and the number of change points* K_T *are known in advance, by choosing* $\alpha = \sqrt{\frac{AK_T}{T}}$*, the regret upper bound of the strategy Bayesian Change-point detection using Thompson Sampling takes the following form:*

$$\mathcal{R}_T^{\texttt{Bayesian-CPD-TS}} = \mathcal{O}\left(\frac{K_T \log T}{\min\limits_{k\in[1,K_T], a\neq a^\star} \boldsymbol{kl}\left(\theta_{a,[k]}, \theta^\star_{[k]}\right)} + \sqrt{AK_T T}\right)$$

Discussion 1 (Knowledge of the number of break-points K_T)**.** *One should note that the optimal tuning of the exploration rate* α *requires a prior knowledge on the number of change-points* K_T *which is a common way to tune the hyper-parameters of the majority of the non-stationary multi-armed bandit algorithms. For instance, the classical discount factor in* `D-UCB` *(Garivier and Moulines (2011)) depends on* K_T*, the sliding window size in* `SW-UCB` *(Garivier and Moulines (2011)) depends also on* K_T*. Moreover, the exploration rate used in* `GLR-KLUCB` *(Besson and Kaufmann (2019a)) is chosen with respect to* K_T*. Finally, the* γ *parameter used in the* `M UCB` *strategy (Cao et al. (2018)) is also tuned with respect to the number of change-points.*

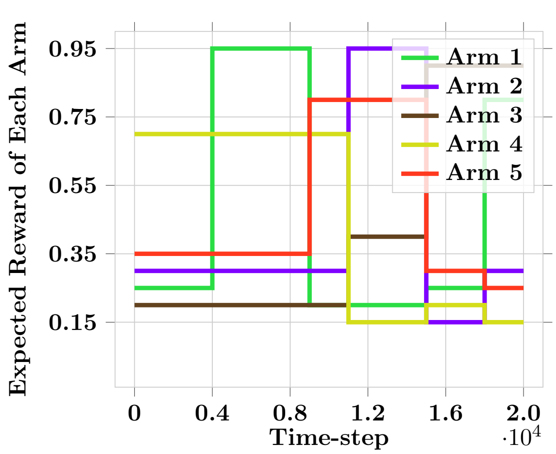
Discussion 2 (Optimality of the regret upper bound)**.** *In the sense of the current lower bound computed for abruptly changing environment which is* $\Omega(T)$ *and stated in Corollary 14 of Garivier and Moulines (2011), the* `Bayesian-CPD-TS` *strategy reaches the order optimal regret rate.*

5. Simulation Results

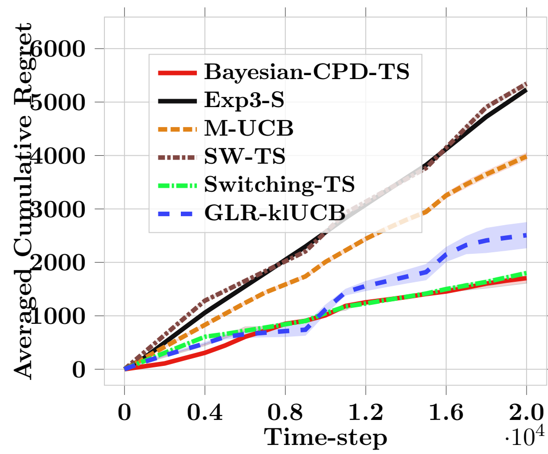
We evaluate the Bayesian change-point detection framework applied on the Thompson sampling algorithm (Bayesian-CPD-TS) in two non-stationary environments: a synthetic dataset (where a switching scenario is simulated) and one real-world dataset from Yahoo!. In both experiments, we compare the performance of Bayesian-CPD-TS against 4 multi-armed bandits algorithms designed for the non-stationary case: Exp3S (Auer et al. (2003)), Sliding Window Thompson Sampling (SW-TS Trovo et al. (2020)), Switching Thompson Sampling (Switching-TS Mellor and Shapiro (2013b)) and Monitored UCB (M-UCB Cao et al. (2018)). For the M-UCB algorithm, we tune the hyper-parameters based on Remark 1 in Cao et al. (2018). Namely, we choose $w = 4\tilde{\delta}^2 \left[(\log(2AT^2))^{1/2} + (\log(2T))^{1/2}\right]^2$, $b = \left[w \log(2AT^2)/2\right]^{1/2}$ and $\gamma = \sqrt{A(K_T - 1) \times (2b + 3\sqrt{w})/(2T)}$, where $\tilde{\delta}$ designates the minimal amplitude of change defined in Cao et al. (2018) Section 5. We choose $\tau = 2\sqrt{T \log T / K_T}$ for SW-TS (same as the tuning of sliding window UCB in Garivier and Moulines (2011)). For Exp-3S, we use $\alpha = 1/T$ and $\gamma = \min\left\{1, \sqrt{A \log(AT)/T}\right\}$. Finally, for the Thompson Sampling bandit used in Bayesian-CPD-TS, we use $s_0 = f_0 = 1$ which corresponds to a uniform prior. Finally, the exploration rate α is tuned following Corollary 1

5.1. Synthetic environment

In this first setting, we generate a piece-wise stationary Bernoulli environment, with a horizon $T = 20000$, $A = 5$ arms and $K_T = 6$ local break-points at time-steps 4000, 9000, 11000, 15000 and 18000 as shown in Figure 1a. We test the above strategies in 50 simulations and record the mean and std. deviation of the cumulative regrets as indicated in Figure 1b.



(a) Generated piece-wise stationary Bernoulli environment with $T = 20000$, $A = 5$ and $K_T = 6$.

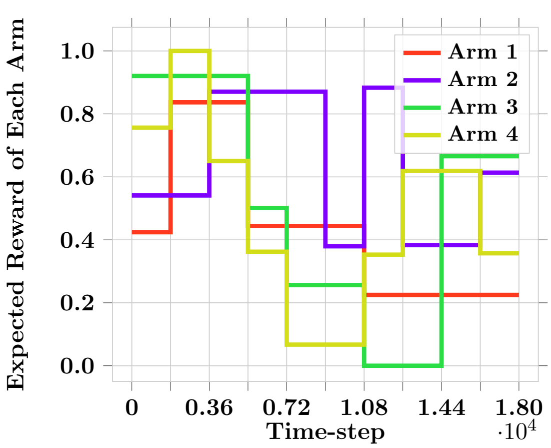


(b) Averaged cumulative regrets for different algorithms in the piece-wise stationary scenario shown in Figure 1a over 50 runs.

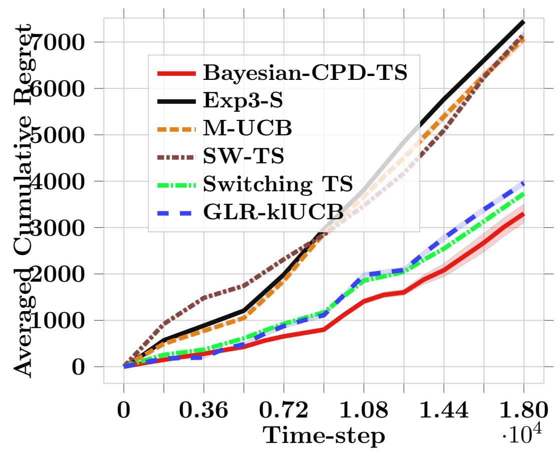
Figure 1: Generated environment and cumulative regrets of MAB strategies from the synthetic dataset.

5.2. Real world environment: Yahoo! Dataset

We apply the previous strategies to a Yahoo! Front Page Today Module dataset [1]. The dataset contains a set of recommended articles, each associated with a binary value, representing whether the user chooses to click the article. We randomly pick $A = 4$ articles, among a pool of 50 articles which have been recommended together the most. Each article is associated with an arm, and we assume a piece-wise stationary Bernoulli process with $K_T = 10$ local break points, by evaluating the mean click-through rates every 1800 seconds, for a total of $T = 18000$ seconds (which is equivalent to five hours). Unlike Cao et al. (2018), we don't set a minimum amplitude of change, but we scale the click-through rates in 0-1 range to obtain greater mean changes. For each strategy, we evaluate the hyper-parameters setting described above, using the obtained environment shown in Figure 2a. In Figure 2b, we observe the mean, and std. deviation of the cumulative regrets over 50 simulations.



(a) Click-through rates computed with $T = 18000$, $A = 4$ and $K_T = 10$.



(b) Averaged cumulative regrets for different algorithms in the piece-wise stationary scenario shown in Figure 2a over 50 runs.

Figure 2: Generated environment and cumulative regrets of MAB strategies from the Yahoo! Dataset.

Discussion 3 (Analysis of the simulation results)**.** *The* `BAYESIAN-CPD-TS` *compares favorably against the state-of-the-art non-stationary MAB strategies, whether on the synthetic or the real-world dataset experiment. Another limitation for the other strategies is that they take a parametric approach to change-point detection, which requires an extra step for hyper-parameters tuning.* `M-UCB` *for example, does not perform well enough in the Yahoo! Dataset because Assumption 1 in Cao et al. (2018) is not verified.*

Discussion 4 (Extension to other distributions)**.** *This work can naturally be extended to other distributions since Thompson Sampling has also been designed for the non-Bernoulli case. To do so, we should consider an extension of the* `RBOCPD` *algorithm to handle non-binary*

1. R6B - Yahoo! Front Page Today Module User Click Log Dataset, available on : https://webscope.sandbox.yahoo.com

observations by replacing the Laplace predictor with a more suitable predictor for general observations.

6. Conclusion and Future Works

We have proposed a class of algorithms for the piece-wise stationary bandit problem; the Bayesian Change-Point detector for bandit framework, Bayesian-CPD-Bandit, which combines the any multi armed bandit algorithm with a optimal restarted Bayesian change-point detector RBOCPD. In the case where Thompson sampling is chosen as bandit algorithm, we have derived a regret upper bound of the order of $\mathcal{O}(\sqrt{ATK_T})$ matching the existing lower bound. From the experiments, the application of this framework using Thompson sampling as bandit algorithm compares favorably against the most popular strategies designed for the non-stationary bandit setting. This comes directly from the powerful RBOCPD test: its detection delay is optimal and the false alarm rate probability is well controlled. As future works, we plan to extend the framework for non-Bernoulli distributions which requires the adaptation of restarted Bayesian online change-point detection for these distributions.

References

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Réda Alami, Odalric Maillard, and Raphael Féraud. Memory bandits: a bayesian approach for the switching bandit problem. In *Neural Information Processing Systems: Bayesian Optimization Workshop.*, 2016.

Reda Alami, Odalric Maillard, and Raphael Feraud. Restarted Bayesian online change-point detector achieves optimal detection delay. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 211–221. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/alami20a.html.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Variance estimates and exploration function in multi-armed bandit. In *CERTIS Research Report 07–31*. Citeseer, 2007.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The non-stochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003. ISSN 0097-5397. doi: 10.1137/S0097539701398375. URL http://dx.doi.org/10.1137/S0097539701398375.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158, 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 199–207, Cambridge, MA, USA, 2014. MIT Press. URL http://dl.acm.org/citation.cfm?id=2968826.2968849.

Lilian Besson and Emilie Kaufmann. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits, 2019a.

Lilian Besson and Emilie Kaufmann. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *arXiv preprint arXiv:1902.01575*, 2019b.

Jie Bian and Kwang-Sung Jun. Maillard sampling: Boltzmann exploration done optimally. In *International Conference on Artificial Intelligence and Statistics*, pages 54–72. PMLR, 2022.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. *arXiv preprint arXiv:1802.03692*, 2018.

Olivier Cappé, Aurélien Garivier, Rémi Maillard, Odalric-Ambrym Munos, and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The 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. *arXiv preprint arXiv:1902.00980*, 2019.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

Aditya Gopalan, Braghadeesh Lakshminarayanan, and Venkatesh Saligrama. Bandit quickest changepoint detection. *Advances in Neural Information Processing Systems*, 34, 2021.

Cédric Hartland, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michèle Sébag. Multi-armed bandit, dynamic environments and meta-bandits. 2006.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer, 2010.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*, pages 592–600, 2012a.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *ALT*, volume 12, pages 199–213. Springer, 2012b.

Levente Kocsis and Csaba Szepesvári. Discounted ucb. In *2nd PASCAL Challenges Workshop*, volume 2, 2006.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

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

Tze Leung Lai and Haipeng Xing. Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential analysis*, 29(2):162–175, 2010.

Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, pages 19–36, 2012.

Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Joseph Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change point detection. *CoRR, abs/1302.3721*, 2013a.

Joseph Charles Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change point detection. *CoRR*, abs/1302.3721, 2013b. URL http://arxiv.org/abs/1302.3721.

Subhojyoti Mukherjee and Odalric-Ambrym Maillard. Distribution-dependent and time-uniform bounds for piecewise iid bandits. *arXiv preprint arXiv:1905.13159*, 2019.

Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/seznec19a.html.

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

Francesco Trovo, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. In *Advances in neural information processing systems*, pages 3972–3980, 2016.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 2020, 2020.