On Limited-Memory Subsampling Strategies for Bandits

Dorian Baudry [*1] **Yoan Russac** [*2] **Olivier Cappé** [2]

Abstract

There has been a recent surge of interest in non-parametric bandit algorithms based on subsampling. One drawback however of these approaches is the additional complexity required by random subsampling and the storage of the full history of rewards. Our first contribution is to show that a simple deterministic subsampling rule, proposed in the recent work of Baudry et al. (2020) under the name of "last-block subsampling", is asymptotically optimal in one-parameter exponential families. In addition, we prove that these guarantees also hold when limiting the algorithm memory to a polylogarithmic function of the time horizon. These findings open up new perspectives, in particular for non-stationary scenarios in which the arm distributions evolve over time. We propose a variant of the algorithm in which only the most recent observations are used for subsampling, achieving optimal regret guarantees under the assumption of a known number of abrupt changes. Extensive numerical simulations highlight the merits of this approach, particularly when the changes are not only affecting the means of the rewards.

1. Introduction

In the K-armed stochastic bandit model, the learner repeatedly picks an action among K available alternatives and only observes the rewards associated with her actions. By interacting with the environment, the learner aims at maximizing her expected sum of rewards and needs to sequentially adapt her decision strategy in light of the information gained up to now. In this model, over-confident policies are provably suboptimal and a proper trade-off between exploitation and exploration has to be found.

[*]Equal contribution [1]Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9198-CRIStAL, F-59000 Lille, France [2]DI ENS, CNRS, Inria, ENS, Université PSL, Paris, France. Correspondence to: Dorian Baudry <dorian.baudry@inria.fr>, Yoan Russac <yoan.russac@ens.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Multi-armed bandits models have been used to address a wide range of sequential optimization tasks under uncertainty: online recommendation (Li et al., 2011; 2016), strategic pricing (Bergemann & Välimäki, 1996) or clinical trials (Zelen, 1969; Vermorel & Mohri, 2005) to name a few. In its standard formulation the multi-armed bandit model postulates that the distributions of the rewards obtained when drawing the different arms remain constant over time. However, in some scenarios the stationary assumption is not realistic. In clinical trials, the disease to defeat may mutate and the initially optimal treatment could become suboptimal compared to another candidate (Gorre et al., 2001). In strategic pricing problems, the price maximizing the profit of a given asset can evolve with the introduction of a new product on the market (Eliashberg & Jeuland, 1986). For online recommendation systems, the preferences of the users are likely to evolve (Wu et al., 2018) and collected data becomes progressively obsolete.

During the past ten years, several works have considered non-stationary variants of the multi-armed bandit model, proposing methods that can be grouped into two main categories: they either actively try to detect modifications in the distribution of the arms with changepoint detection algorithms (Liu et al., 2017; Cao et al., 2019; Auer et al., 2019; Chen et al., 2019; Besson et al., 2020) or they passively forget past information (Garivier & Moulines, 2011; Raj & Kalyani, 2017; Trovo et al., 2020). To some extent, all of these methods require some knowledge on the distribution to obtain theoretical guarantees.

To balance exploration and exploitation, the algorithms mentioned so far are based on one of the two standard building blocks introduced in the bandit literature: *Upper Confidence Bound* (UCB) constructions (Auer et al., 2002) or *Thompson Sampling* (TS) (Thompson, 1933). However, there has been a recent surge of interest for alternative non-parametric bandit strategies (Kveton et al., 2019a;b; Riou & Honda, 2020). Instead of using prior information on the reward distributions as in Thompson sampling or of building tailored upper-confidence bounds (Cappé et al., 2013) those methods only use the empirical distribution of the data. These algorithms are non-parametric in the sense that the *exact same* implementation can be used with different probability distributions, while still achieving optimal regret guarantees (in a sense to be defined in Section 2 below).

In particular, subsampling algorithms (Baransi et al., 2014; Chan, 2020; Baudry et al., 2020) have demonstrated their potential thanks to their flexibility and strong theoretical guarantees. From a high level perspective, they all rely on the same two components. **(1)** *subsampling*: the arms that have been pulled a lot are randomized by sampling only a fraction of their history. **(2)** *duels*: the arms are pulled based on the outcomes of duels between the different pairs of arms. Note that the term *duel*, which we will also use in the following, refers to the algorithmic principle of comparing the arms two by two, based on their subsamples. It is totally unrelated to the dueling bandit framework introduced by Yue & Joachims (2009).

Scope and contributions In this paper, we build on the Last-Block Subsampling Duelling Algorithm (LB-SDA) introduced by Baudry et al. (2020) but for which no theoretical guarantees were provided. This approach is of interest because of its simplicity and its computational efficiency compared to other strategies based on randomized subsampling. We first prove that for stationary environments LB-SDA is asymptotically optimal in one-parameter exponential family models and therefore matches the guarantees obtained by Baudry et al. (2020) for randomized subsampling schemes. The main technical challenge is to devise an alternative to the *diversity* condition used in their work, which was specifically designed for randomized subsampling schemes.

Furthermore, we show that, without additional changes, these guarantees still hold for a variant of the algorithm using a *limited memory* of the observations of each arm. We prove that storing $\Omega\left((\log T)^2\right)$ observations instead of T is sufficient to ensure the asymptotic guarantees, making the algorithm more tractable for larger time horizons. To the best of our knowledge, this paper is the first to propose an asymptotically optimal subsampling algorithm with poly-logarithmic storage of rewards under general assumptions.

Building a subsampling algorithm based on the most recent observations makes it an ideal candidate for a passively forgetting policy. Our third contribution is to propose a natural extension of the LB-SDA strategy to non-stationary environments. By limiting the extent of the time window in which subsampling is allowed to occur, one obtains a passively forgetting non-parametric bandit algorithm, which we refer to as Sliding Window Last Block Subsampling Duelling Algorithm (SW-LB-SDA). To analyze the performance of this algorithm, we assume an abruptly changing environment in which the reward distributions change at unknown time instants called *breakpoints*. We show that SW-LB-SDA guarantees a regret of order $\mathcal{O}(\sqrt{\Gamma_T T \log(T)})$ for any abruptly changing environment with at most Γ_T breakpoints, thus matching the lower bounds from Garivier & Moulines (2011), up to logarithmic factors. The only required assumption is that, during each stationary phase, the reward distributions belong to the same one-parameter exponential family for all arms. Due to its non-parametric nature, this algorithm can thus be used in many scenarios of interest beyond the standard bounded-rewards / change-in-the-mean framework. We discuss some of these scenarios in Section 5, where we validate numerically the potential of the approach by comparing it with a variety of state-of-the-art algorithms for non-stationary bandits.

2. Preliminaries

The algorithms to be presented below are designed for the *stochastic K-armed bandit* model, which is the most studied setting in the bandit literature. We introduce in this section the two variants of this basic model that will be considered in the paper: *stationary* and *abruptly changing* environments.

Stationary environments When the environment is stationary, the K arms are characterized by the reward distributions $(\nu_k)_{k\leq K}$ and their associated means $(\mu_k)_{k\leq K}$, with $\mu^\star = \max_{k\in\{1,\dots,K\}} \mu_k$ denoting the highest expected reward. We denote by $(Y_{k,s})_{s\in\mathbb{N}}$ the i.i.d. sequence of rewards from arm k. Following Chan (2020), our algorithm operates in successive rounds, whose length varies between 1 and K time steps. At each round r, the *leader* denoted $\ell(r)$ is defined and $(K-1)$ duels with the remaining arms called *challengers* are performed. Denoting by $N_k(r)$ the number of pulls of arm k up to the round r the leader is the arm that has been most pulled. Namely,

$$\ell(r) = \text{argmax}_{k\in\{1,\dots,K\}} N_k(r) \ . \tag{1}$$

When several arms are candidate for the maximum number of pulls, the one with the largest sum of rewards is chosen. If this is still not sufficient to obtain a unique arm, the leader is chosen at random among the arms maximizing both criteria. At round r, a subset $\mathcal{A}_r \subset \{1,\dots,K\}$ is selected by the learner based on the outcomes of the duels against $\ell(r)$. Next, all arms in $\mathcal{A}_r$ are drawn, yielding $Y_{k,N_k(r)}$ for $k \in \mathcal{A}_r$, where $N_k(r) = \sum_{s=1}^{r} \mathbb{1}(k \in \mathcal{A}_s)$.

The regret is defined as the expected difference between the highest expected reward and the rewards collected by playing the sequence of arms $(A_t)_{t\leq T}$:

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} (\mu^\star - \mu_{A_t})\right] \ .$$

For distributions in one-parameter exponential families, the lower bound of Lai & Robbins (1985) states that no strategy can systematically outperform the following asymptotic regret lower bound

$$\liminf_{T\to\infty} \frac{\mathcal{R}_T}{\log(T)} \geq \sum_{k:\mu_k<\mu^\star} \frac{\mu^\star - \mu_k}{\text{kl}(\mu_k, \mu^\star)} \ .$$

Abruptly changing environments In Section 4, we consider abruptly changing environments. The number of breakpoints up to time T, denoted Γ_T, is defined by

$$\Gamma_T = \sum_{t=1}^{T-1} \mathbb{1}\{\exists k, \nu_{k,t} \neq \nu_{k,t+1}\}.$$

The time instants $(t_1, ..., t_{\Gamma_T})$ associated to these breakpoints define $\Gamma_T + 1$ stationary phases where the reward distributions are fixed. Note that in this model, the change do not need to affect all arms simultaneously. In such environments, letting $\mu_t^\star = \max_{k\in\{1,...,K\}} \mu_{k,t}$ denote the best arm at time t, the performance of a policy is measured through the *dynamic regret* defined as

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_t^\star - \mu_{A_t})\right].$$

We will explain how to extend the notion of leader to this setting in Section 4.

In the non-stationary case, the lower bound for the regret takes a different form: for any strategy, there exists an abruptly changing instance such that $\mathbb{E}[\mathcal{R}_T] = \Omega(\sqrt{T\Gamma_T})$ (Garivier & Moulines, 2011; Seznec et al., 2020). Note that in the bandit literature, there is also another, more general, way of characterizing non-stationary environments based on a variational distance introduced by Besbes et al. (2014). In this work, we however only consider the case of abruptly changing environments.

3. LB-SDA in Stationary Environments

In this section we detail the subsampling strategy used in the LB-DSA algorithm and obtain asymptotically optimal regret guarantees for its performance. In Section 3.3, we consider the variant of LB-SDA in which the memory available to the algorithm is strongly limited.

3.1. Last Block Sampling

Compared to the algorithms analyzed in (Baudry et al., 2020) where the sampler is randomized, we consider a *deterministic sampler*. At round r, the duel between arm $k \neq \ell(r)$ and the leader consists in comparing the average reward from arm k with the average reward computed only from the last $N_k(r)$ observations of the leader. The challenger k thus wins its duel if

$$\bar{Y}_{k,N_k(r)} \geq \bar{Y}_{\ell(r),N_{\ell(r)}(r)-N_k(r)+1:N_{\ell(r)}(r)}\ , \qquad (2)$$

where $\bar{Y}_{k,i:j} = \frac{1}{j-i+1}\sum_{n=i}^{j} Y_{k,n}$ denotes the average computed on the $j-i+1$ observations of arm k between its i-th and j-th pull, and $\bar{Y}_{k,n}$ is a shortcut for $\bar{Y}_{k,1:n}$.

At each round, the set $\mathcal{A}_{r+1}$ includes all of the challengers that have defeated the leader, according to Equation (2), as well as under-explored arms for which $N_k(r) \leq \sqrt{\log(r)}$. If $\mathcal{A}_{r+1}$ is empty, only the leader is pulled. Combining these elements gives LB-SDA detailed below.

Algorithm 1 LB-SDA

Input: K arms, horizon T
Initialization: $t \leftarrow 1$, $r \leftarrow 1$, $\forall k \in \{1, ..., K\}, N_k \leftarrow 0$
while $t < T$ **do**
  $\mathcal{A} \leftarrow \{\}$, $\ell \leftarrow$ leader(N, Y)
  if $r = 1$ **then**
    $\mathcal{A} \leftarrow \{1, \dots, K\}$ (Draw each arm once)
  else
    for $k \neq \ell \in \{1, ..., K\}$ **do**
      if $N_k \leq \sqrt{\log(r)}$ or $\bar{Y}_{k,N_k} \geq \bar{Y}_{\ell,N_\ell-N_k+1:N_\ell}$ **then**
        $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$
      if $|\mathcal{A}| = 0$ **then**
        $\mathcal{A} \leftarrow \{\ell\}$
  for $k \in \mathcal{A}$ **do**
    Pull arm k, observe reward Y_{k,N_k+1}, $N_k \leftarrow N_k + 1$,
    $t \leftarrow t + 1$
  $r \leftarrow r + 1$

Baransi et al. (2014) propose interesting arguments explaining why subsampling methods work. Essentially, if the sampler allows enough *diversity* in the duels, the probability of repeatedly selecting a suboptimal arm is small. On the sampler side, this condition is satisfied when out of a large number of duels between two arms there is a reasonable amount of them with non-overlapping subsamples. We prove that last block sampling satisfies such property. The second requirement concerns the distribution of the arms, and has been formulated by Baransi et al. (2014) who introduced the *balance function* of a family of distributions. In particular, Chan (2020) shows that introducing an asymptotically negligible sampling obligation of $\sqrt{\log r}$ is enough to make subsampling suitable when the arms come from the same one-parameter exponential family of distributions. Namely, if each arm has at least $\sqrt{\log r}$ samples at round r, the *diversity* of duels will guarantee each arm to be pulled enough. This exploration rate does not have to be tuned and is not detrimental in practice : for an horizon of, say, $T = 10^6$ it only forces each arm to be sampled at least 4 times.

3.2. Regret Analysis of LB-SDA

We consider that the arms come from the same one-parameter exponential family of distributions $\mathcal{P}_\Theta$, i.e., that there exists a function $g : \mathbb{R} \times \Theta \mapsto \mathbb{R}$ such that any arm k has a density of the form

$$g_k(x) = g(x, \theta_k) = e^{\theta_k x - \Psi(\theta_k)} g(x, 0)\ ,$$

where $\Psi(\theta_k) = \log\left[\int e^{\theta_k x} g(x,0)\,dx\right]$. This assumption is standard in the literature and covers a broad range of bandits applications. The exact knowledge of the family of distributions of the arms (e.g Bernoulli, Gaussian with known variance, Poisson, etc.) can be used to calibrate algorithms like Thompson Sampling (Kaufmann et al., 2012), KL-UCB (Cappé et al., 2013) or IMED (Honda & Takemura, 2015) in order to reach asymptotic optimality. Recently, subsampling algorithms like SSMC (Chan, 2020) and RB-SDA (Baudry et al., 2020) have been proved to be optimal *without* knowing exactly $\mathcal{P}_\Theta$. This means that the same algorithm can run on Bernoulli or Gaussian distributions and achieve optimality. We first prove that LB-SDA matches these theoretical guarantees. We denote $\mathrm{kl}(\mu,\mu')$ the Kullback-Leibler divergence between two distributions of mean μ and μ' in the exponential family $\mathcal{P}_\Theta$.

Theorem 1 (Asymptotic optimality of LB-SDA)**.** *For any bandit model $\nu = (\nu_1,\dots,\nu_K) \subset \mathcal{P}_\Theta^K$ where $\mathcal{P}_\Theta$ is any one-parameter exponential family of distributions, the regret of LB-SDA satisfies, for all $\varepsilon > 0$,*

$$\mathcal{R}(T) \leq \sum_{k:\mu_k<\mu^\star} \frac{1+\varepsilon}{\mathrm{kl}(\mu_k,\mu^\star)} \log(T) + C(\nu,\varepsilon)\ ,$$

where $C(\nu,\varepsilon)$ is a problem-dependent constant.

Proof sketch We assume without loss of generality that there is a unique optimal arm denoted $k^\star$. The analysis of Chan (2020) and Baudry et al. (2020) shows that for any SDA algorithm the number of pulls of a suboptimal arm may be bounded as follow.

Lemma 1 (Lemma 4.1 in Baudry et al. (2020))**.** *For any suboptimal arm $k \neq k^\star$, the expected number of pulls of k is upper bounded by*

$$\mathbb{E}[N_k(T)] \leq \frac{1+\varepsilon}{\mathrm{kl}(\mu_k,\mu^\star)} \log(T) + C_k(\nu,\varepsilon) + 32\sum_{r=1}^{T} \mathbb{P}(N_{k^\star}(r) \leq (\log r)^2)\ , \quad (3)$$

where $C_k(\nu,\varepsilon)$ is a problem-dependent constant.

The next step consists in upper bounding the probability that the best arm is not pulled "enough" during a run of the algorithm. This part is more challenging and relies on the notion of *diversity* in the subsamples provided by the subsampling algorithm. This notion was introduced by Baransi et al. (2014) to analyze the Best Empirical Sampled Average (BESA) algorithm. Intuitively, random block sampling (Baudry et al., 2020) or sampling without replacement (Baransi et al., 2014) explore different part of the history thus bringing diversity in the duels. Unfortunately, this property is not satisfied by deterministic samplers. Nonetheless, with a careful examination of the relation implied by the deterministic nature of last-block subsampling it is possible to prove that the number of pulls of the optimal arm is large enough with high probability.

Lemma 2. *The probability that the optimal arm is not pulled enough by LB-SDA can be upper bounded as follows*

$$\sum_{r=1}^{+\infty} \mathbb{P}\left(N_{k^\star}(r) \leq (\log r)^2\right) \leq C_{k^\star}(\nu)\ ,$$

for some constant $C_{k^\star}(\nu)$.

Plugging the result of Lemma 2 in Lemma 1 gives the asymptotic optimality of LB-SDA (Theorem 1). The proof of Lemma 2 is reported in Appendix A.

3.3. Memory-Limited LB-SDA

One of our main motivations for studying LB-SDA is its simplicity and efficiency. Yet, all existing subsampling algorithms (Baransi et al., 2014; Chan, 2020; Baudry et al., 2020) as well as the vanilla version of LB-SDA have to store the entire history of rewards for all the arms. In this section, we explain how to modify LB-SDA to reduce the storage cost while preserving the theoretical guarantees.

The fact that LB-SDA is asymptotically optimal means that, when T is large, the arm with the largest mean is most often the leader with all of its challengers having a number of pulls that is of order $O(\log T)$ only. With duels based on the last block, this would mean in particular that only the last $O(\log T)$ observations from the optimal arm should be stored and that previous observations will *never* be used again in practice. Based on this intuition, one might think that keeping only $\log(T)/(\mu^\star-\mu_k)^2$ observations is enough for LB-SDA. However, this could only be done with the knowledge of the gaps that are unknown.

We propose instead to limit the storage memory of each arm at round r to a value of the form

$$m_r = \max\left(M, \left\lceil C(\log r)^2 \right\rceil\right)\ ,$$

where $C > 0$ and $M \in \mathbb{N}$. M ensures that a minimum number of samples are stored during the first few rounds. Following the definition of Agrawal & Goyal (2012), we then define the set of *saturated arms* at a round r as

$$\mathcal{S}_r = \{k \in \{1,\dots,K\} : N_k(r) \geq m_r\}\ .$$

The only modification of LB-SDA is the following: at each round r, if a saturated arm is pulled then the newly collected observation replaces the oldest observation in its history. The pseudo code of LB-SDA with Limited Memory (LB-SDA-LM) is given in Appendix B and the following result shows that it keeps the same asymptotical performance as LB-SDA under general assumptions on m_r.

Theorem 2 (Asymptotic optimality of LB-SDA with Limited Memory)**.** *For any bandit model $\nu = (\nu_1, \dots, \nu_K) \subset \mathcal{P}_\Theta^K$ where $\mathcal{P}_\Theta$ is any one-parameter exponential family of distributions, if $m_r / \log(r) \to \infty$, the regret of memory-limited LB-SDA satisfies, for all $\varepsilon > 0$,*

$$\mathcal{R}_T \leq \sum_{k:\mu_k<\mu^\star} \frac{1+\varepsilon}{\mathrm{kl}(\mu_k, \mu^\star)} \log(T) + C'(\nu, \varepsilon, \mathcal{M}) \ ,$$

where $\mathcal{M} = (m_1, m_2, \dots, m_T)$ denotes the sequence $(m_r)_{r\in\mathbb{N}}$ and $C'(\nu, \varepsilon, \mathcal{M})$ is a problem-dependent constant.

The proof of this theorem is reported in Appendix B, which provides precise estimates of the dependence of $C'(\nu, \varepsilon, \mathcal{M})$ with respect to the parameters, and in particular, with respect to the sequence $\mathcal{M}$. Note that LB-SDA-LM remains an anytime algorithm because the storage constraint does not depend on the time horizon T but only on the current round.

3.4. Storage and Computational Cost

To the best of our knowledge, LB-SDA-LM is the only subsampling bandit algorithm that does not require to store the full history of rewards. We report in Table 1 estimates of the computational cost of LB-SDA-LM and its competitors.

Table 1. Storage and computational cost at round T for existing subsampling algorithms.

Algorithm	Storage	Comp. cost Best-Worst case
BESA (Baransi et al., 2014)	$O(T)$	$O((\log T)^2)$
SSMC (Chan, 2020)	$O(T)$	$O(1)$-$O(T)$
RB-SDA (Baudry et al., 2020)	$O(T)$	$O(\log T)$
LB-SDA **(this paper)**	$O(T)$	$O(1)$-$O(\log T)$
LB-SDA-LM **(this paper)**	$O((\log T)^2)$	$O(1)$-$O(\log T)$

The computational cost can be broken into two parts: (a) the subsampling cost and (b) the computation of the means of the samples. We assume that drawing a sample of size n without replacement has $O(n)$ cost and that computing the mean of this subsample costs another $O(n)$. Furthermore, at round T, each challenger to the best arm has about $O(\log T)$ samples. This gives an estimated cost of $O\left((\log T)^2\right)$ for BESA (Baransi et al., 2014). For RB-SDA (Baudry et al., 2020) the estimated cost is $O(\log(T))$, because the sampling cost for random block sampling is $O(1)$ and only the sample mean has to be recomputed at each round.

For the three deterministic algorithms (namely SSMC (Chan, 2020), LB-SDA, LB-SDA-LM), when the leader arm wins all its duels, its sample mean can be updated sequentially at cost $O(1)$. This is the *best case* in terms of computational cost. However, when a challenger arm is pulled, SSMC requires a full screening of the leader's history, with $O(T)$ cost, while LB-SDA and LB-SDA-LM only need the computation of the mean of the last $O(\log T)$ samples from the leader.

4. LB-SDA in Non-Stationary Environments

In stationary environments, LB-SDA achieves optimal regret rates, even when its decisions are constrained to use at most $O((\log T)^2)$ observations. One might think that this argument itself is sufficient to address non-stationary scenarios as the duels are performed mostly using recent observations. However, the latter is only true for the best arm and in the case where an arm that has been bad for a long period of time suddenly becomes the best arm, adapting to the change would still be prohibitively slow. For this reason, LB-SDA has to be equipped with an additional mechanism to perform well in non-stationary environments.

4.1. SW-LB-SA: LB-SDA with a Sliding-Window

We keep a *round-based* structure for the algorithm, where, at each round r, duels between arms are performed and the algorithm subsequently selects the subset of arms $\mathcal{A}_r$ that will be pulled. In contrast to Section 3.3, where a constraint on storage related to the number of pulls was added, here, we use a sliding window of length τ to limit the historical data available to the algorithm to that of the last τ rounds.

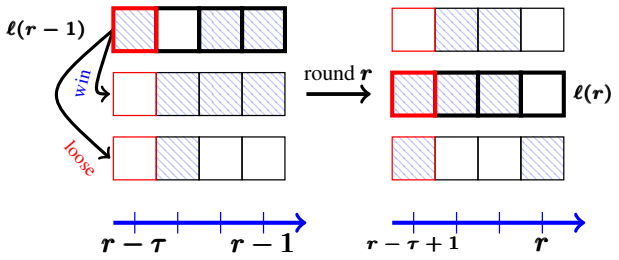

Figure 1. Illustration of a *passive leadership takeover* with a sliding window $\tau = 4$ when the standard definition of leader is used. The bold rectangle correspond to the leader. A blue square is added when an arm has an observation for the corresponding round and the red square correspond to the information that will be lost at the end of the round due to the sliding window.

Modified leader definition The introduction of a sliding window requires a new definition for the *leader*. By analogy with the stationary case, the leader could be defined as the arm that has been pulled the most during the τ last rounds.

Algorithm 2 SW-LB-SDA

Input: K arms, horizon T, τ length of sliding window
Initialization: $t \leftarrow 1$, $r \leftarrow 1$, $\forall k \in \{1, ..., K\}, N_k \leftarrow 0$, $N_k^\tau \leftarrow 0$
while $t < T$ **do**
  $\mathcal{A} \leftarrow \{\}$, $\ell \leftarrow \text{leader}(N, Y, \tau)$
  if $r = 1$ **then**
    $\mathcal{A} \leftarrow \{1, \dots, K\}$ (Draw each arm once)
  else
    for $k \neq \ell \in \{1, ..., K\}$ **do**
      if $N_k^\tau \leq \sqrt{\log(\tau)}$ *or* $D_k^\tau(r) = 1$ **then**
        $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$
      else
        $\widehat{\mu}_k^\tau = \bar{Y}_{k, N_k - N_k^\tau + 1 : N_k}$
        $N = \min(N_k^\tau, N_\ell^\tau)$
        $\widehat{\mu}_{\ell,k}^\tau = \bar{Y}_{N_\ell - N + 1 : N_\ell}$
        if $\widehat{\mu}_k^\tau \geq \widehat{\mu}_{\ell,k}^\tau$ **then**
          $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$
    if $|\mathcal{A}| = 0$ **then**
      $\mathcal{A} \leftarrow \{\ell\}$
  for $k \in \mathcal{A}$ **do**
    Pull arm k, observe reward $Y_{k, N_k + 1}$
    Update $N_k \leftarrow N_k + 1$, $N_k^\tau \leftarrow N_k^\tau + 1$, $t \leftarrow t + 1$
  for $k \in \{1, ..., K\}$ **do**
    if $k \in \mathcal{A}_{r-\tau+1}$ **then**
      $N_k^\tau \leftarrow N_k^\tau - 1$
  $r \leftarrow r + 1$

However, with the inclusion of the sliding window, a new phenomenon, which we call *passive leadership takeover*, can occur. Let us define $N_k^\tau(r) = \sum_{s=r-\tau}^{r-1} \mathbb{1}(k \in \mathcal{A}_{s+1})$, the number of times arm k has been pulled during the last τ rounds and consider a situation with 3 arms $\{1, 2, 3\}$. Assume that the leader is arm 1 and at a round $(r-1)$ we have $N_1^\tau(r-1) = N_2^\tau(r-1)$. If the leader has been pulled τ rounds away and wins its duel against arm 2 but looses against arm 3, only arm 3 will be pulled at round r. Consequently, at round r, arm 2 will have a strictly larger number of pulls than arm 1 without having actually defeated the leader. This situation, illustrated on Figure 1, is not desirable as it can lead to spurious leadership changes. We fix this by imposing that any arm has to defeat the current leader to become the leader itself. Define,

$$\mathcal{B}_r = \{k \in \mathcal{A}_{r+1} \cap \{N_k^\tau(r+1) \geq \min(r, \tau)/K\}\} \ .$$

Then for any $r \in \mathbb{N}$, the leader at round $r+1$ is defined as $\ell^\tau(r+1) = \text{argmax}_{k \in \{1,\dots,K\}} N_k^\tau(r+1)$ if $N_{\ell^\tau(r)}^\tau(r+1) < \min(r, \tau)/(2K)$ and the argmax is taken over $\mathcal{B}_r \cup \{\ell^\tau(r)\}$ otherwise. This modified definition of the leader ensures that an arm can become the leader only after earning at least τ/K samples and winning a duel against the current leader, or if the leader loses a lot of duels and its number of samples falls under a fixed threshold. Thanks to this definition it holds that $N_{\ell^\tau(r)}^\tau(r) \geq \min(r, \tau)/(2K)$. More details are given in Appendix C.

Additional diversity flags As in the vanilla LB-SDA, we use a sampling obligation to ensure that each arm has a minimal number of samples. However, in contrast to the stationary case, this very limited number of forced samples may not be sufficient to guarantee an adequate variety of duels, due to the forgetting window. To this end, the sampling obligation is coupled with a *diversity flag*. We define it as a binary random variable $D_k^\tau(r)$, satisfying $D_k^\tau(r) = 1$ only when, for the last $\lceil (K-1)(\log \tau)^2 \rceil$ rounds the three following conditions are satisfied: 1) some arm $k' \neq k$ has been leader during all these rounds, 2) k' has not been pulled, and 3) k has not been pulled and satisfy $N_k^\tau(r) \leq (\log \tau)^2$. In practice, there is a very low probability that these conditions are met simultaneously but this additional mechanism is required for the theoretical analysis. Note that the diversity flags have no impact on the computational cost of the algorithm as they require only to store the number of rounds since the last draw of the different arms (which can be updated recursively) as well as the last leader takeover. Arms that raise their diversity flag are automatically added to the set of pulled arms.

Bringing these parts together, gives the pseudo-code of SW-LB-SDA in Algorithm 2.

4.2. Regret Analysis in Abruptly Changing Environments

In this section we aim at upper bounding the dynamic regret in abruptly changing environments, as defined in Section 2. Our main result is the proof that the regret of SW-LB-SDA matches the asymptotic lower bound of Garivier & Moulines (2011).

Theorem 3 (Asymptotic optimality of SW-LB-SDA)**.** *If the time horizon T and number of breakpoint Γ_T are known, choosing $\tau = O(\sqrt{T \log(T)/\Gamma_T})$ ensures that the dynamic regret of SW-LB-SDA satisfies*

$$\mathcal{R}_T = O(\sqrt{T \Gamma_T \log T}) \ .$$

To prove this result we only need to assume that, during each stationary period, the rewards come from the same one-parameter exponential family of distributions. In contrast, current state-of-the-art algorithms for non-stationary bandits typically require the assumption that the rewards are *bounded* to obtain similar guarantees. Hence, this result is of particular interest for tasks involving unbounded reward distributions that can be discrete (e.g Poisson) or continuous (e.g Gaussian, Exponential). SW-LB-SDA can also

be used for general bounded rewards with the same performance guarantees by using the *binarization trick* (Agrawal & Goyal, 2013). Note however, that the knowledge of the horizon T and the estimated number of change point Γ_T is still required to obtain optimal rates, which is an interesting direction for future works on this approach (Auer et al., 2019; Besson et al., 2020). We provide a high-level outline of the analysis behind Theorem 3 and the complete proof is given in Appendix C.

Regret decomposition For the $\Gamma_T + 1$ stationary phases $[t_\phi, t_{\phi+1}-1]$ with $\phi \in \{1, \dots, \Gamma_T\}$, we define r_ϕ as the first round where an observation from the phase ϕ was pulled. Introducing the gaps $\Delta_k^\phi = \mu^*_{t_\phi} - \mu_{t_\phi,k}$ and denoting the optimal arm $k^\star_\phi$, we can rewrite the regret as

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{\phi=1}^{\Gamma_T}\sum_{r=r_\phi-1}^{r_{\phi+1}-2}\sum_{k\neq k^\star_\phi}\mathbb{1}\left(k\in\mathcal{A}_{r+1}\right)\Delta_k^\phi\right]$$
$$= \sum_{\phi=1}^{\Gamma_T}\sum_{k\neq k^*_\phi}\mathbb{E}[N_k^\phi]\Delta_k^\phi\,,$$

where we define $N_k^\phi = \sum_{r=r_\phi-1}^{r_{\phi+1}-2}\mathbb{1}(k\in\mathcal{A}_{r+1})$ the number of pulls of an arm k during a phase ϕ when it is suboptimal.

Note that the quantities t_ϕ, r_ϕ and Δ_k^ϕ for the different stationary phases ϕ are only required for the theoretical analysis and the algorithm has no access to those values. We highlight that the sequence $(r_\phi)_{\phi\geq 1}$ is a random variable that depends on the trajectory of the algorithm. However, we show in Appendix C that this causes no additional difficulty for upper bounding the regret. We introduce $\delta_\phi = t_{\phi+1} - t_\phi$ the length of a phase ϕ. Combining elements from the proofs of Garivier & Moulines (2011) and that of Theorem 1, we first provide an upper bound on $\mathbb{E}[N_k^\phi]$ for any suboptimal arm k during the phase ϕ as

$$\mathbb{E}[N_k^\phi] \leq 2\tau + \frac{\delta_\phi A_k^{\phi,\tau}}{\tau} + c_{k,1}^{\phi,\tau} + c_{k,2}^{\phi,\tau} + c_{k,3}^{\phi,\tau}\,.$$

In this decomposition we define $A_k^{\phi,\tau} = b_k^\phi \log(\tau)$ for some constant $b_k^\phi > 0$, along with the terms $c_{k,1}^{\phi,\tau}$, $c_{k,2}^{\phi,\tau}$ and $c_{k,3}^{\phi,\tau}$, which all represents a different technical aspect of the regret decomposition of SW-LB-SDA. Before interpreting them we start with their formal definition,

$$c_{k,1}^{\phi,\tau} = \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2}\mathbb{1}\left(\mathcal{G}_k^\tau(r, A_k^{\phi,\tau})\right)\right]\,,$$

$$c_{k,2}^{\phi,\tau} = \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2}\mathbb{1}\left(\ell^\tau(r)=k^*_\phi, D_k^\tau(r)=1\right)\right]\,,$$

$$c_{k,3}^{\phi,\tau} = \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2}\mathbb{1}\left(\ell^\tau(r)\neq k^*_\phi\right)\right]\,,$$

where $\mathcal{G}_k^\tau(r,n)$ is equal to

$$\{k\in\mathcal{A}_{r+1}, \ell^\tau(r)=k^*_\phi, N_k^\tau(r)\geq n, D_k^\tau(r)=0\}\,.$$

Bounding individual terms The three terms have intuitive interpretation and summarize well the technical contributions behind Theorem 3. To some extent they all rely on the notion of *saturated* arms defined in Section 3.3 and that we refine in Appendix C for the problems considered in this section (mainly by properly tuning $A_k^{\phi,\tau}$ in the theoretical analysis).

First, $c_{k,1}^{\phi,\tau}$ is an upper bound on the expectation of the number of times a *saturated suboptimal arm* can defeat the *optimal* leader (i.e $\ell^\tau(r) = k^*_\phi$). To prove this result we establish a new concentration inequality for Last-Block Sampling in the context of SW-LB-SDA.

The second term $c_{k,2}^{\phi,\tau}$ controls the probability that the *diversity flag* is activated when the optimal arm k^*_ϕ is the leader. We prove that if this event happen, then k^*_ϕ has necessarily lost at least one duel against a saturated *sub-optimal* arm, and that this event has only a low probability.

The term $c_{k,3}^{\phi,\tau}$ is the most difficult to handle, the main challenge is to upper bound the probability that the *optimal arm* is *not saturated* after a large number of rounds.

In Appendix C we provide the complete analysis of each of these terms and a full description of all the technical results that led to Theorem 3.

5. Experiments

Limiting the storage in stationary environments. In our first experiment[1] reported on Figure 3, we compare LB-SDA and LB-SDA-LM on a stationary instance with $K = 2$ arms with Bernoulli distributions for a horizon $T = 10000$. We add natural competitors (Thompson Sampling (Thompson, 1933), kl-UCB (Cappé et al., 2013)), that know ahead of the experiment that the reward distributions are Bernoulli and are tuned accordingly. The arms satisfy $(\mu_1, \mu_2) = (0.05, 0.15)$ with a gap $\Delta = 0.1$. We run LB-SDA-LM with a memory limit $m_r = \log(r)^2 + 50$, which gives a storage ranging from 50 to 150 samples (much smaller than the horizon T = 10000). The regret are averaged on 2000 independent replications and the upper and lower quartiles are reported. In this setup LB-SDA-LM performs similarly to KL-UCB, and the impact of limiting the memory is mild, when compared to LB-SDA. This illustrates that even with relatively small gaps (here 0.1), a substantial reduction of the storage can be done with only minor loss of performance with LB-SDA-LM.

[1]The code for obtaining the different figures reported in the paper is available at https://github.com/YRussac/LB-SDA.

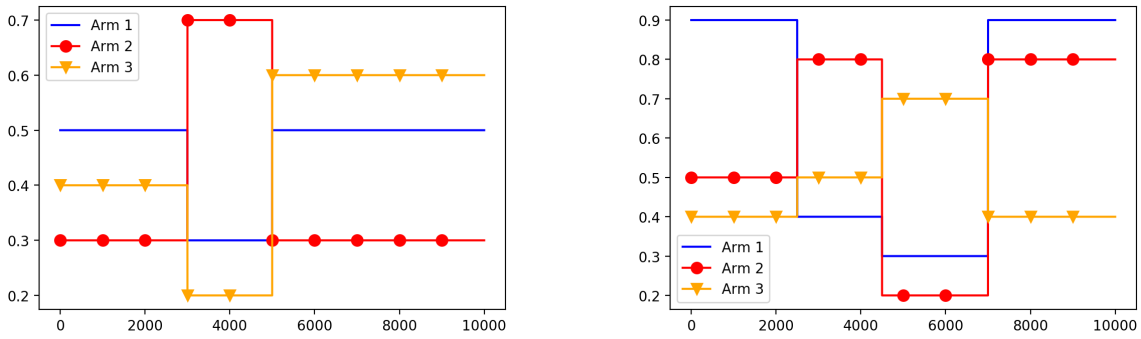

Figure 2. Evolution of the means: Left, Bernoulli arms (Fig. 4); Right, Gaussian arms (Figs. 5 and 6).

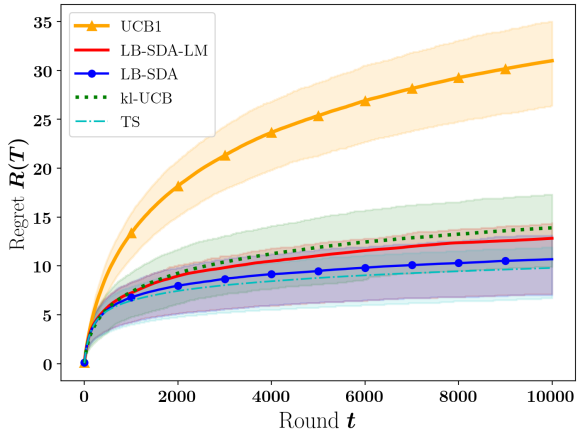

Figure 3. Cost of storage limitation on a Bernoulli instance. The reported regret are averaged over 2000 independent replications.

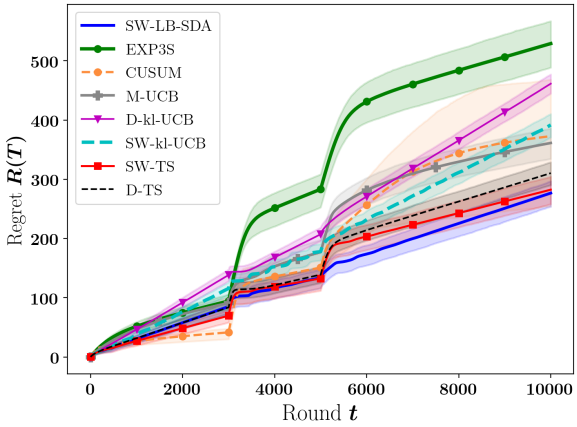

Figure 4. Performance on a Bernoulli instance averaged on 2000 independent replications.

Empirical performance in abruptly changing environments. In the second experiment, we compare different state-of-the-art algorithms on a problem with $K = 3$ Bernoulli-distributed arms. The means of the distributions are represented on the left hand side of Figure 2 and the performance averaged on 2000 independent replications are reported on Figure 4. Two changepoint detection algorithms, CUSUM (Liu et al., 2017) and M-UCB (Cao et al., 2019) are compared with progressively forgetting policies based on upper confidence bound, SW-klUCB and D-klUCB adapted from Garivier & Moulines (2011), or Thompson sampling, DTS (Raj & Kalyani, 2017) and SW-TS (Trovo et al., 2020). We also add EXP3S (Auer et al., 2002) designed for adversarial bandits and our SW-LB-SDA algorithm for the comparison. The different algorithms make use of the knowledge of T and Γ_T.

To allow for fair comparison, we use for SW-LB-SDA, the same value of $\tau = 2\sqrt{T\log(T)/\Gamma_T}$ that is recommended for SW-UCB (Garivier & Moulines, 2011). D-UCB uses the discount factor suggested by Garivier & Moulines (2011), $1/(1-\gamma) = 4\sqrt{T/\Gamma_T}$. The changepoint detection algorithms need extra information such has the minimal gap for a breakpoint and the minimum length of a stationary phase. For M-UCB, we set $w = 800$ and $b = \sqrt{w/2\log(2KT^2)}$ as recommended by Cao et al. (2019) but set the amount of exploration to $\gamma = \sqrt{K\Gamma_T\log(T)/T}$ following Besson et al. (2020). In practice, using this value rather than the theoretical suggestion from Cao et al. (2019) improved significantly the empirical performance of M-UCB for the horizon considered here. For CUSUM, α and h are tuned using suggestions from Liu et al. (2017), namely $\alpha = \sqrt{\Gamma_T/T\log(T/\Gamma_T)}$ and $h = \log(T/\Gamma_T)$. On this specific instance, using $\varepsilon = 0.05$ (to satisfy Assumption 2 of Liu et al. (2017)) and $M = 50$ gives good performance. For the EXP3S algorithm, following (Auer et al., 2002) the parameters α and γ are tuned as follows: $\alpha = 1/T$ and $\gamma = \min(1, \sqrt{K(e+\Gamma_T\log(KT)/((e-1)T)})$.

This problem is challenging because a policy that focuses on arm 1 to minimize the regret in the first stationary phase also has to explore sufficiently to detect that the second arm is the best in the second phase. SW-LB-SDA has performance

comparable to the forgetting TS algorithms and is the best performing algorithm in this scenario. Note that both TS algorithms use the assumption that the arms are Bernoulli whereas SW-LB-SDA does not. SW-klUCB performs better than D-klUCB performance and its performance closely matches the one from the changepoint detection algorithms. By observing the lower and the upper quartiles, one sees that the performance of CUSUM vary much more than the other algorithms depending on its ability to detect the breakpoints. Finally, EXP3S, which can adapt to more general adversarial settings, lags behind the other algorithms in abruptly changing stochastic environments.

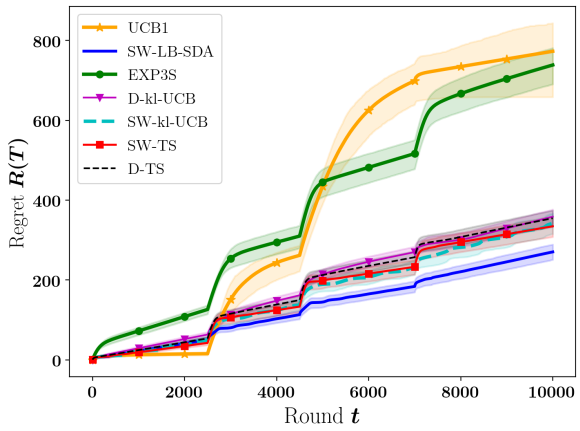

Figure 5. Performance on a Gaussian instance with a constant standard deviation of $\sigma = 0.5$ averaged on 2000 independent replications.

In the third experiment with $\Gamma_T = 3$ breakpoints, the $K = 3$ arms comes from Gaussian distributions with a fixed standard deviation of $\sigma = 0.5$ but time dependent means. The evolution of the arm's means is pictured on the right of Figure 2 and Figure 5 displays the performance of the algorithms. CUSUM and M-UCB can not be applied in this setting because CUSUM is only analyzed for Bernoulli distributions and M-UCB assume that the distributions are bounded. Even if no theoretical guarantees exist for Thompson sampling with a sliding window or discount factors, when the distribution are Gaussian with known variance, we add them as competitors. The analysis of SW-UCB and D-UCB was done under the bounded reward assumption but the algorithms can be adapted to the Gaussian case. Yet, the tuning of the discount factor and the sliding window had to be adapted to obtain reasonable performance, using $\tau = 2(1+2\sigma)\sqrt{T\log(T)/\Gamma_T}$ for D-UCB and $\gamma = 1 - 1/(4(1+2\sigma))\sqrt{\Gamma_T/T}$ for SW-UCB (considering that, practically, most of the rewards lie under $1+2\sigma$). For reference, Figure 5 also displays the performance of the UCB1 algorithm that ignores the non-stationary structure. Clearly, SW-LB-SDA, in addition of being the only algorithm analyzed in this setting with unbounded rewards, also has the best empirical performance.

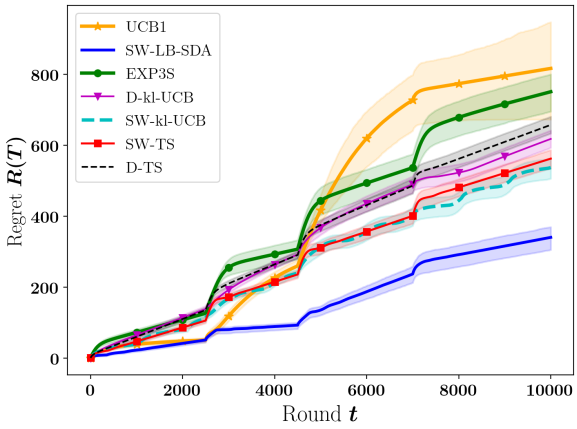

Figure 6. Performance on a Gaussian instance with time dependent standard deviations averaged on 2000 independent replications.

Changes affecting the variance. The last experiment features the same Gaussian means but with different standard errors. The standard error takes the values $0.5, 0.25, 1$ and 0.25, respectively, in the four stationary phases. The algorithms based on upper confidence bound are given the maximum standard error $\sigma = 1$, whereas SW-LB-SDA is not provided with any information of this sort. Figure 6 shows that the non-parametric nature of SW-LB-SDA is effective, with a significant improvement over state-of-the-art methods in such settings.

Acknowledgements

The PhD of Dorian Baudry is funded by a CNRS80 grant.

References

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Agrawal, S. and Goyal, N. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pp. 99–107. PMLR, 2013.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47, 2002.

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

Baransi, A., Maillard, O.-A., and Mannor, S. Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 115–131. Springer, 2014.

Baudry, D., Kaufmann, E., and Maillard, O.-A. Sub-sampling for efficient non-parametric bandit exploration. *Advances in Neural Information Processing Systems*, 33, 2020.

Bergemann, D. and Välimäki, J. Learning and strategic pricing. *Econometrica: Journal of the Econometric Society*, pp. 1125–1149, 1996.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pp. 199–207, 2014.

Besson, L., Kaufmann, E., Maillard, O.-A., and Seznec, J. Efficient change-point detection for tackling piecewise-stationary bandits. Prepint, December 2020.

Cao, Y., Wen, Z., Kveton, B., and Xie, Y. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 418–427. PMLR, 2019.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., Stoltz, G., et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41 (3):1516–1541, 2013.

Chan, H. P. The multi-armed bandit problem: An efficient nonparametric solution. *The Annals of Statistics*, 48(1): 346–373, 2020.

Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pp. 696–726. PMLR, 2019.

Eliashberg, J. and Jeuland, A. P. The impact of competitive entry in a developing market upon dynamic pricing strategies. *Marketing Science*, 5(1):20–36, 1986.

Garivier, A. and Moulines, E. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

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

Gorre, M. E., Mohammed, M., Ellwood, K., Hsu, N., Paquette, R., Rao, P. N., and Sawyers, C. L. Clinical resistance to sti-571 cancer therapy caused by bcr-abl gene mutation or amplification. *Science*, 293(5531):876–880, 2001.

Honda, J. and Takemura, A. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16:3721–3756, 2015.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory - 23rd International Conference, ALT*, 2012.

Kveton, B., Szepesvari, C., Ghavamzadeh, M., and Boutilier, C. Perturbed-history exploration in stochastic multi-armed bandits. *arXiv preprint arXiv:1902.10089*, 2019a.

Kveton, B., Szepesvari, C., Vaswani, S., Wen, Z., Lattimore, T., and Ghavamzadeh, M. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 3601–3610. PMLR, 2019b.

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

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

Li, S., Karatzoglou, A., and Gentile, C. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 539–548, 2016.

Liu, F., Lee, J., and Shroff, N. A change-detection based framework for piecewise-stationary multi-armed bandit problem. *arXiv preprint arXiv:1711.03539*, 2017.

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

Riou, C. and Honda, J. Bandit algorithms based on thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory*, pp. 777–826. PMLR, 2020.

Seznec, J., Menard, P., Lazaric, A., and Valko, M. A single algorithm for both restless and rested rotting bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3784–3794. PMLR, 2020.

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

Trovo, F., Paladino, S., Restelli, M., and Gatti, N. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

Vermorel, J. and Mohri, M. Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pp. 437–448. Springer, 2005.

Wu, Q., Iyer, N., and Wang, H. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 495–504, 2018.

Yue, Y. and Joachims, T. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009.

Zelen, M. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64 (325):131–146, 1969.