Journal of Artificial Intelligence Research 68 (2020) 311-364 Submitted 03/2019; published 05/2020

Sliding-Window Thompson Sampling for Non-Stationary Settings

Francesco Trovò FRANCESCO1.TROVO@POLIMI.IT
Stefano Paladino STEFANO.PALADINO@POLIMI.IT
Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
Nicola Gatti NICOLA.GATTI@POLIMI.IT
Politecnico di Milano,
Dipartimento di Elettronica, Informazione e Bioingegneria,
Piazza Leonardo da Vinci 32,
Milano, 20133, Italy

Abstract

Multi-Armed Bandit (MAB) techniques have been successfully applied to many classes of sequential decision problems in the past decades. However, *non-stationary* settings—very common in real-world applications—received little attention so far, and theoretical guarantees on the regret are known only for some *frequentist* algorithms. In this paper, we propose an algorithm, namely Sliding-Window Thompson Sampling (SW-TS), for non-stationary stochastic MAB settings. Our algorithm is based on Thompson Sampling and exploits a sliding-window approach to tackle, in a unified fashion, two different forms of non-stationarity studied separately so far: *abruptly changing* and *smoothly changing*. In the former, the reward distributions are constant during sequences of rounds, and their change may be arbitrary and happen at unknown rounds, while, in the latter, the reward distributions smoothly evolve over rounds according to unknown dynamics. Under mild assumptions, we provide regret upper bounds on the dynamic pseudo-regret of SW-TS for the abruptly changing environment, for the smoothly changing one, and for the setting in which both the non-stationarity forms are present. Furthermore, we empirically show that SW-TS dramatically outperforms state-of-the-art algorithms even when the forms of non-stationarity are taken separately, as previously studied in the literature.

1. Introduction

The Multi-Armed Bandit (MAB) setting, introduced by Auer, Cesa-Bianchi, and Fischer (2002), models the sequential decision-making problem, addressing the well-known exploration-exploitation trade-off. In this setting, at each round of a finite time horizon, the learner selects an action from a finite set of actions (also known as *arms*), and she only observes the reward of the chosen action. The goal of the learner is to play the optimal arm, maximizing the expected reward while minimizing the loss incurred during the learning process. This loss is usually addressed as *regret*, defined as the difference between the expected reward collected by a *clairvoyant* algorithm, selecting the optimal arm through the whole-time horizon, and the expected reward achieved by the used MAB algorithm. We focus on the Non-Stationary stochastic MAB (NS-MAB) setting, where, differently from the classical stochastic MAB setting, the expected reward of each arm may change over time, thus potentially changing the optimal arm.

Non-stationarity behaviours are common in real-world applications. We recall that the former motivation for MAB settings argued by Thompson (1933) was the study of clinical trials, where different treatments are available, and a learner aims at selecting the treatment to use for the next patient. Although the clinical trial scenario was assumed stationary over time in its original formulation, it may not be in the real world. Indeed, in a scenario in which the trial takes place over periods, the disease to defeat may mutate. Thus, as showed by Gorre, Mohammed, Ellwood, Hsu, Paquette, Rao, and Sawyers (2001), a treatment that initially was optimal might subsequently slowly decrease its effectiveness, and another treatment, which initially was ineffective, might become the best option. Similarly, non-stationarity plays a prominent role in Internet economics. For instance, in optimal pricing problems, a non-stationarity may be due to a new product invading the market. For instance, Eliashberg and Jeuland (1986) show that the price maximizing the expected profit of a product already present in the market may change abruptly when a newer product enters. Furthermore, in untruthful auction mechanisms for search advertising where advertisers try to learn the best bid to obtain their ad displayed in some profitable slot, non-stationarity may be due to the arrival and departure of advertisers, which change the profitability of the slots, as studied by Kitts and Leblanc (2004).[1] Finally, Lai, El Gamal, Jiang, and Poor (2011) study a cognitive medium radio access problem, in which a user aims to opportunistically exploit the availability of an empty channel in a multiple channel system. The reward expresses the binary/fractional availability of the channel, whose distribution is unknown to the users. In particular, the probability that a given channel is available to a user changes over time as the other users' behavior changes.

As stressed by Hartland, Gelly, Baskiotis, Teytaud, and Sebag (2006), general-purpose classical MAB algorithms are not suitable when tackling NS-MAB settings, their regret bounds not holding anymore. In non-stationary settings, two main approaches are studied: *passive*, *e.g.*, the works by Combes and Proutiere (2014) and Garivier and Moulines (2008), and *active*, *e.g.*, the works by Liu, Lee, and Shroff (2018), Besson and Kaufmann (2019), Auer, Gajane, and Ortner (2019). The former ones can deal with non-stationarity without detecting explicitly that a change of the reward distributions occurred, while the latter ones exploit techniques coming from the change detection field to deal with reward distributions varying over time. In our work, we focus on passive approaches, since they require fewer assumptions on the change characteristics than those required by active approaches. For instance, as argued by Liu et al. (2018), active approaches commonly require the knowledge of the minimum magnitude of the change, to avoid excessively long delays in its detection. However, in practice, this knowledge is rarely available to the learner, making active approaches less appealing for real-world applications.

Among the techniques following the passive approach, some frequentist algorithms have been proposed showing order optimal theoretical guarantees. We mention the works by Besbes, Gur, and Zeevi (2014), Combes and Proutiere (2014), Garivier and Moulines (2008), Kocsis and Szepesvári (2006), and Wei, Hong, and Lu (2016). To the best of our knowledge,

1. Generalized Second Price (GSP) is an untruthful auction mechanism used by Google and BING. Examples of problems on the publisher side are discussed by Farina and Gatti (2017) and Gatti, Lazaric, Rocco, and Trovò (2015), while examples of problems on the advertiser side are discussed by Nuara, Trovò, Gatti, and Restelli (2018), Gasparini, Nuara, Trovò, Gatti, and Restelli (2018), and Nuara, Sosio, Trovò, Zaccardi, Gatti, and Restelli (2019).

all the known Bayesian methods are only based on heuristics, *e.g.*, the works by Granmo and Berg (2010) and Mellor and Shapiro (2013), while we recall that, in stationary settings, some Bayesian algorithms with theoretical guarantees are known, *e.g.*, Thompson Sampling (TS) by Thompson (1933), and these algorithms empirically outperform frequentist algorithms in most scenarios. Some examples are also discussed by Chapelle and Li (2011), Granmo (2010), Kaufmann, Korda, and Munos (2012b), May, Korda, Lee, and Leslie (2012), Paladino, Trovò, Restelli, and Gatti (2017). Notably, heuristic algorithms might outperform algorithms with theoretical guarantees in specific settings, but, on the other side, they may provide an arbitrarily large regret in others. In the present paper, we provide a Bayesian MAB algorithm for non-stationary settings with theoretical guarantees in terms of dynamic pseudo-regret. Remarkably, our algorithm tackles in a unified fashion two forms of non-stationarity—the abruptly changing one and the smoothly changing one—that have been studied separately so far in the literature. In the former, the reward distributions are constant during sequences of rounds, and their change may be arbitrary and happen at unknown rounds, while, in the latter, the reward distributions smoothly evolve over rounds according to unknown dynamics. More precisely, our original contributions are as follows:

- we design a novel Bayesian MAB algorithm, named Sliding-Window Thompson Sampling (SW-TS), working when two different forms of non-stationarity coexist;
- we derive some an bound over the dynamic pseudo-regret for SW-TS of order $\tilde{O}(N^{\frac{1+\alpha}{2}})$ when abruptly changing non-stationarity is present, and an upper bound of order $\tilde{O}(N^{\beta})$ when smoothly changing non-stationarity is present.[2] Parameters α and β provide a measure of the number of abrupt changes and the number of rounds the expected rewards of the arms are close enough, respectively, over a horizon of N rounds. Finally, we derive a bound on the dynamic pseudo-regret in the setting in which both the non-stationarity forms are present;
- we empirically show the superior performance of SW-TS over state-of-the-art frequentist MAB passive algorithms even when the forms of non-stationarity are taken separately. Finally, we provide a sensitivity analysis of SW-TS for the parameters α and β.

2. Related Works

Non-stationary MAB settings have received attention in the scientific community only in the last few years. When rewards may change arbitrarily over time, the problem of NS-MAB is intractable, *i.e.*, one can only derive trivial bounds on the dynamic pseudo-regret. For this reason, the literature mainly focuses on non-stationary MAB settings with some specific structure in the attempt to design algorithms with better regret bounds. Garivier and Moulines (2008) study abruptly changing MAB settings and present the SW-UCB algorithm achieving an $\tilde{O}(\sqrt{N})$ bound on the dynamic pseudo-regret. The same setting is tackled by Allesiardo, Féraud, and Maillard (2017), who present the SER4 algorithm, which empirically outperforms the SW-UCB algorithm. Combes and Proutiere (2014) present SW-KL-UCB, which is a policy working in a smoothly changing MAB setting. In this

2. With the notation $\tilde{O}(\cdot)$ we disregard logarithmic terms in the computation of the order.

non-stationary setting, the authors provide a bound of $\tilde{O}(\sigma^{1/4}N)$ on the dynamic pseudo-regret, being σ the Lipschitz constant of the process. This result implies that the per-round pseudo-regret of SW-KL-UCB vanishes as the speed at which the expected rewards evolve decreases to 0.

Besbes et al. (2014) study a non-stationary MAB setting under the assumption that the total variation of the expected rewards over the time horizon is bounded by a budget that is *a priori* fixed. They provide a distribution-independent lower bound. Furthermore, they propose the REXP3 algorithm, a near-optimal frequentist algorithm with a dynamic pseudo-regret of order $O(N^{2/3})$. Slivkins and Upfal (2008) focus on the *dynamic bandit* setting—a special case of the *restless bandits*—, in which the reward distribution of the arms changes at each round according to Brownian motion. The authors propose algorithms that minimize the per-round pseudo-regret over an infinite time horizon. We also mention the work by Trovò, Paladino, Restelli, and Gatti (2018), who provide some bandit algorithms for dynamic pricing in non-stationary settings. Finally, the problem of non-stationarity with bounded per-round variation is tackled using contextual bandit techniques by Slivkins (2011), who designs the Contextual Zooming algorithm, and by Luo, Wei, Agarwal, and Langford (2018), for which they use a variant of the classic EXP4 algorithm.

The MAB literature also provides some works that exploit MAB techniques as heuristics on application scenarios without providing theoretical guarantees. To cite a few, Granmo and Berg (2010) propose a Bayesian algorithm for the specific case of non-stationary bandit settings with *normally distributed rewards*. Mellor and Shapiro (2013) analyze an NS-MAB where the probabilities according to which the expected value of the arms change are *a priori* fixed and propose the CTS algorithm that combines Thompson Sampling with a change point detection mechanism. St-Pierre and Jialin (2014) present an evolutionary algorithm to deal with generic non-stationary environments which empirically outperforms classical solutions.

Other settings, closely related to the MAB one, are also studied in the presence of non-stationarity. For instance, Wei et al. (2016) present a study of the regret in the case of non-stationary stochastic experts, providing an upper bound of order $O(N^{1/3})$ in the case we assume a constant number of switches and limited variance of the expected rewards over time.

3. Problem Formulation

We model our problem as a stochastic NS-MAB setting, in which, at each round t over a finite horizon N, the learner selects an arm a_{i_t} among a finite set of K arms $\mathcal{A} := \{a_1, \dots, a_K\}$. At each round t the learner observes a realization of the reward $x_{i_t,t}$ obtained from the chosen arm a_{i_t}. The reward for each arm a_i at round t is modeled by a sequence of independent random variables $X_{i,t}$ from a distribution unknown to the learner. We denote by $\mu_{i,t} := \mathbb{E}[X_{i,t}]$ the expected value of the reward of the arm a_i at round t. As is customary in the MAB literature, here we consider Bernoulli distributed rewards, *i.e.*, $X_{i,t} \sim Be(\mu_{i,t})$.[3] A *policy* $\mathfrak{U}$ is a function $\mathfrak{U}(h_t) = a_{i_t}$ that chooses the arm a_{i_t} to play at round t according to history h_t, defined as the sequence of past plays and obtained rewards.

3. The extension to other bounded distributions is straightforward. Bernoulli variables are considered here for the sake of simplicity.

The goal of the learner is to design a policy $\mathfrak{U}$ that minimizes the loss w.r.t. the optimal decision in terms of reward. This loss, usually addressed as cumulative *dynamic pseudo-regret*, is defined as:

$$\overline{R}_N(\mathfrak{U}) := \mathbb{E}\left[\sum_{t=1}^{N}\left(\mu_{i_t^*,t} - \mu_{i_t,t}\right)\right], \tag{1}$$

where $\mu_{i_t^*,t} = \max_{i\in\{1,\dots,K\}} \mu_{i,t}$ is the expected reward of the optimal arm $a_{i_t^*}$ at round t and $\mathbb{E}[\,\cdot\,]$ is the expectation w.r.t. the stochasticity of the policy. Differently from the classical (stationary) stochastic MAB setting, where an arm (unique unless degeneracy) is optimal for the whole-time horizon ($a_{i_t^*} = a_{i^*}, \forall t$), in the NS-MAB setting the arms that are optimal might change over time. We recall that when the optimal expected value of the arm can change without any restriction, the NS-MAB setting has only trivial bounds on the dynamic pseudo-regret $\overline{R}_N(\mathfrak{U})$. One of the focus of the MAB research is the design of algorithms that guarantee sublinear pseudo-regret, *i.e.*, $\overline{R}_N(\mathfrak{U}) = \tilde{O}(N^\omega)$ with $0 \leq \omega < 1$. When this is not possible, an alternative performance metric is the *average pseudo-regret*, defined as:

$$\overline{AR}_N(\mathfrak{U}) := \lim\sup_{N\to\infty} \frac{\overline{R}_N(\mathfrak{U})}{N}. \tag{2}$$

In what follows, we will discuss two different settings where the evolution over time of the reward distributions of the arms is constrained to change according to specific schemes.

3.1 Abruptly Changing Setting

The Abruptly Changing MAB (AC-MAB) setting is introduced, for the first time, by Garivier and Moulines (2008). In this scenario, the reward distributions are constant during sequences of rounds, namely *phases*, and change at unknown rounds, namely *breakpoints*. Thus, the expected value $\mu_{i,t}$ of the reward of arm a_i at round t only changes at the beginning of each phase and, therefore, the best arm $a_{i_t^*}$ remains constant during the phase. The change of the expected rewards at the breakpoints may be arbitrary and is unknown.

Let us define a breakpoint as a round $b \in \{1,\dots,N\}$ s.t. $\exists i \mid \mu_{i,b-1} \neq \mu_{i,b}$, *i.e.*, a round b in which the expected reward of at least one arm a_i changes w.r.t. the one at round $b-1$. In an AC-MAB setting with horizon N, we have a set of breakpoints $\mathcal{B} := \{b_1,\dots,b_{B_N}\}$ of cardinality B_N (for sake of notation we define $b_0 = 1$), which determines a set of phases $\{\mathcal{F}_1,\dots,\mathcal{F}_{B_N}\}$, where each phase is a set of rounds between two consecutive breakpoints, namely, $\mathcal{F}_\phi = \{t \in \{1,\dots,N\} \text{ s.t. } b_{\phi-1} \leq t < b_\phi\}$. In order to have sublinear dynamic pseudo-regret, we upper bound the number of breakpoints B_N over the time horizon N. We do that by making the following assumption.

Assumption 1. *There exists $\alpha \in [0,1]$, independent of N, s.t. the number of breakpoints B_N is of order $O(N^\alpha)$. That is, there exist $\alpha \in [0,1)$ and $B \in \mathbb{R}^+$ such that: $B_N \leq BN^\alpha$.*

During phase $\mathcal{F}_\phi$ of an AC-MAB setting, with abuse of notation, we denote with $\mu_{i,\phi}$ the expected value of the reward of arm a_i, where $a_{i_\phi^*}$ is the optimal arm and $\mu_{i^*,\phi}$ is the corresponding expected reward. By defining the length of a phase as $N_\phi := |\mathcal{F}_\phi|$, a more

compact formulation of the dynamic pseudo-regret of a generic policy $\mathfrak{U}$ over an AC-MAB is available:

$$\overline{R}_N(\mathfrak{U}) = \sum_{i=1}^{K} \sum_{\phi=1}^{B_N} \Delta_{i,\phi} \mathbb{E}[T_i(\mathcal{F}_\phi)],$$

where $T_i(\mathcal{F}_\phi) = \sum_{t \in \mathcal{F}_\phi} \mathbb{1}\{i_t = i\}$ is the number of times arm a_i has been pulled during phase $\mathcal{F}_\phi$, $\Delta_{i,\phi} := \mu_{i^*,\phi} - \mu_{i,\phi}$ is the difference between the expected reward $\mu_{i^*,\phi}$ of the optimal arm $a_{i^*_\phi}$ of phase $\mathcal{F}_\phi$ and the expected reward $\mu_{i,\phi}$ of arm a_i, and $\mathbb{E}[\cdot]$ is the expectation w.r.t. the stochasticity of the policy.[4] This alternative formulation highlights that the dynamic pseudo-regret in this setting can be decomposed over the different phases such that, in each phase, the dynamic pseudo-regret takes the form of the classic expected pseudo-regret.

3.2 Smoothly Changing Setting

The Smoothly Changing MAB (SC-MAB) setting we study is similar to that one studied by Combes and Proutiere (2014), where the expected value $\mu_{i,t}$ of each arm varies no more than σ at each round, and the evolution of the dynamics is unknown to the learner. More formally, we make the following Lipschitz assumption.

Assumption 2. *There exists $\sigma > 0$, such that $\left|\mu_{i,t} - \mu_{i,t'}\right| \leq \sigma |t - t'|$ for all $t, t' \in \{1, \dots, N\}$ and all $i \in \{1, \dots, K\}$.*

Furthermore, in such a setting, a suboptimal arm a_i might be arbitrarily close to the optimal one $a_{i^*_t}$ in terms of expected reward. Identifying the best arm among those with similar expected rewards is known to be hard, as showed by Lai and Robbins (1985). Indeed, it is known that a learner takes a time of the order of $\frac{1}{(\mu_{i^*_t,t} - \mu_{i,t})^2}$. Thus, to prevent the dynamic pseudo-regret from being linearly dependent on the horizon N, we assume also that the separation between the expected rewards of two arms is arbitrarily small only for a limited number of rounds. More formally, consider $0 < \Delta < 1$, we define:

$$\mathcal{F}_{\Delta,N} := \{t \in \{1, \dots, N\} \text{ s.t. } \exists i \neq j, |\mu_{i,t} - \mu_{j,t}| < \Delta\}$$

and we assume the following.

Assumption 3. *There exist $\beta \in [0, 1]$, $F \in \mathbb{R}^+$, and $\Delta_0 \in (0, 1)$, all independent of N, s.t. for all $\Delta < \Delta_0$ it holds:*

$$|\mathcal{F}_{\Delta,N}| \leq F \Delta N^\beta.$$

We remark that the assumption used by Combes and Proutiere (2014) is a particular case of the above assumption when $\beta = 1$.

4. From now on, we denote with $|\cdot|$ the cardinality operator and with $\mathbb{1}\{\cdot\}$ the indicator function of a generic event.

Table 1: Summary of the notation used in the paper.

Parameter	Description
$\mathcal{A} = \{a_1, \dots a_K\}$	Set of the K available arms a_i.
$X_{i,t}$	Random variable corresponding to the reward for arm a_i at round t.
$\mu_{i,t}$	Expected value of arm a_i at round t.
$a_{i_t^*}$	Optimal arm at round t, *i.e.*, the one providing the largest expected reward $\mu_{i_t^*,t}$
$\mathfrak{U}(h_t)$	Policy selecting an arm in A for round t, given history h_t.
$\mathcal{B} := \{b_1, \dots, b_{B_N}\}$	Set of the B_N breakpoints b_ϕ.
$\mathcal{F}_\phi$	Set of rounds between two consecutive breakpoints $b_{\phi-1}$ and b_ϕ
$T_i(\mathcal{F}_\phi)$	Number of times arm a_i is pulled during phase $\mathcal{F}_\phi$.
$\Delta_{i,\phi} := \mu_{i^*,\phi} - \mu_{i,\phi}$	Difference between the expected reward $\mu_{i^*,\phi}$ of the optimal arm $a_{i_\phi^*}$ of phase $\mathcal{F}_\phi$ and the expected reward $\mu_{i,\phi}$ of arm a_i.
$\mathcal{F}_{\Delta,N}$	Set of rounds over a time horizon of N in which the optimal arm expected reward has a distance of at least Δ from the second best arm.

3.3 Abruptly and Smoothly Changing Setting

Finally, in a quite straightforward way, it is possible to study also a scenario, from now on addressed as Abruptly and Smoothly Changing MAB (ASC-MAB) setting, in which the two forms of non-stationarity introduced above (abrupt changes and smooth ones) simultaneously occur over a finite time period. In this setting, in addition to Assumptions 1 and 3, we also require the following assumption.

Assumption 4. *There exists $\sigma > 0$ and a set of phases $\{\mathcal{F}_\phi, \dots, \mathcal{F}_{B_N}\}$ that, for each $\mathcal{F}_\phi$ with $\phi \in \{1, \dots, B_N\}$, it holds:*

$$\left|\mu_{i,t} - \mu_{i,t'}\right| \leq \sigma \left|t - t'\right|,$$

for all $i \in \{1, \dots, K\}$ and for all $t, t' \in \mathcal{F}_\phi$, i.e.*, the expected value of the reward function is Lipschitz continuous w.r.t. the rounds belonging to a single phase.*

This newly defined assumption is the natural extension of Assumption 2 to this new setting, in which the smoothness assumption might be violated if the process is at breakpoints.[5]

4. Sliding-Window Thompson Sampling Algorithm

We propose an algorithm that exploits a Sliding-Window (SW) approach to forget past information during the learning, which could provide a bias to the estimation process. More precisely, we use a sliding window of length $\tau \in \mathbb{N}$ such that the algorithm, at every

5. A summary of the notation defined in the previous section and used in the following sections is provided in Table 1.

Algorithm 1 SW-TS

1: **Input:** $\{\pi_{i,0}\}_i$ prior distributions, N time horizon, A arm set, τ sliding window size
2: **for** $t \in \{1, \dots, N\}$ **do**
3: **for** $i \in \{1, \dots, K\}$ **do**
4: Compute $\pi_{i,t} = \text{Beta}(S_{i,t,\tau} + 1, T_{i,t,\tau} - S_{i,t,\tau} + 1)$
5: Sample $\vartheta_{i,t}$ from $\pi_{i,t}$
6: Play arm a_{i_t} s.t.: $i_t = \arg\max_{i \in \{1,\dots,K\}} \vartheta_{i,t}$ and observe $x_{i_t,t+1}$

round t, takes into account only the rewards obtained in the last τ rounds. Based on these realizations, we apply a TS-based algorithm to decide which is the arm to pull in the next round. In particular, the expected value of each arm is coupled with a posterior distribution from which we draw samples, and the arm with the highest value is the next arm to play. At first, we describe the algorithm and provide theoretical results about the finite-time analysis of its dynamic pseudo-regret for Bernoulli distributed rewards separately for the AC-MAB and SC-MAB. After that, we study the ASC-MAB setting in which both the non-stationary forms (abrupt and smoothly changing) are present at the same time.[6]

4.1 The Algorithm

The pseudocode of SW-TS for Bernoulli distributed rewards is presented in Algorithm 1. Assume to have, for each arm a_i, a prior $\pi_{i,0}$ on the reward expected value $\mu_{i,t}$ and let $\pi_{i,t}$ be the posterior distribution for the parameter $\mu_{i,t}$ after t rounds. In the case we do not have further information on the expected value of the arm, we use an uninformative prior, *i.e.*, $\pi_{i,0} := \text{Beta}(1, 1)$, where we denote with $\text{Beta}(a, b)$ the Beta distribution with parameters a and b. The posterior of the expected reward of arm a_i at round t is $\pi_{i,t} := \text{Beta}(S_{i,t,\tau} + 1, T_{i,t,\tau} - S_{i,t,\tau} + 1)$, where $T_{i,t,\tau} := \sum_{s=\max\{t-\tau+1,1\}}^{t} \mathbb{1}\{i_s = i\}$ is the number of times arm a_i has been selected in the last $\min\{t, \tau\}$ rounds, and $S_{i,t,\tau} := \sum_{s=\max\{t-\tau+1,1\}}^{t} x_{i,s}\mathbb{1}\{i_s = i\}$ is the cumulative reward collected by arm a_i in the last $\min\{t, \tau\}$ rounds.[7] Once computed the distributions $\pi_{i,t}$ (Line 4), from each one of them, we draw a random sample $\vartheta_{i,t}$, also known as *Thompson sample* (Line 5). Finally, we select arm a_i with the highest sample $\vartheta_{i,t}$ for this round (Line 6). The extension of the SW-TS algorithm to the case where the rewards $X_{i,t}$ comes from other bounded distributions is similar to what proposed for the classical TS algorithm, as showed by Chapelle and Li (2011) and Agrawal and Goyal (2012), given that conjugate prior/posterior distributions for the expected rewards of the arms are available.

4.2 Finite-Time Analysis in the Abruptly Changing Setting

We provide a finite-time analysis of the dynamic pseudo-regret achieved by the SW-TS algorithm, in the AC-MAB setting introduced in Section 3.1.

6. We report the proofs of our theoretical results in Appendix B, Appendix C, and Appendix D, respectively.
7. To avoid an excessively cumbersome notation, we omit the subscript τ in all the terms depending on the choice of τ, *e.g.*, $\pi_{i,t}$.

Theorem 1. *If the SW-TS policy is run over an AC-MAB setting with $X_{i,t} \sim Be(\mu_{i,t})$, for every $\tau \in \mathbb{N}$, the dynamic pseudo-regret after N rounds is at most:*

$$\overline{R}_N(\mathfrak{U}) \leq \sum_{i=1}^{K} \left[\tau B N^{\alpha} + \sum_{\phi=1}^{B_N} \Delta_{i,\phi} \frac{N_\phi}{\tau} \left(\frac{52 \log \tau}{\Delta_{i,\phi}^2} + \log \tau + 5 + \frac{19}{\log \tau} \right) \right],$$

where B and α are defined in Assumption 1 and $\Delta_{i,\phi} := \mu_{i^*,\phi} - \mu_{i,\phi}$ is the difference between the expected reward $\mu_{i^*,\phi}$ of the best arm $a_{i^*_\phi}$ and the expected reward $\mu_{i,\phi}$ of arm a_i during phase $\mathcal{F}_\phi$. By defining:

$$\Delta_i := \min_{\phi \in \{1,\dots,B_N\}} \Delta_{i,\phi} \mathbb{1}\{i \neq i^*_\phi\},$$

for all $i \in \{1, \dots, K\}$, i.e.*, the minimum over all the phases $\mathcal{F}_\phi$ of the difference of the expected rewards $\Delta_{i,\phi}$, the dynamic pseudo-regret can be written as:*

$$\overline{R}_N(\mathfrak{U}) \leq \tau K B N^{\alpha} + \frac{N}{\tau} \sum_{i=1}^{K} \left(\frac{52 \log \tau}{\Delta_i} + \log \tau + 5 + \frac{19}{\log \tau} \right).$$

From this general result we derive two corollaries for the cases in which we have a number of breakpoints B_N which is either sublinear or linear w.r.t. the time horizon N.

Corollary 1. *If the SW-TS policy is run over an AC-MAB setting in which Assumption 1 holds with $\alpha \in [0, 1)$ and using a sliding window $\tau \propto N^{\frac{1-\alpha}{2}}$, the dynamic pseudo-regret is:*

$$\overline{R}_N(\mathfrak{U}) = \tilde{O}(N^{\frac{1+\alpha}{2}}).$$

Notice that the size of the sliding window prescribed by Corollary 1 decreases as the parameter α increases, meaning that, in settings in which we have a large number of breakpoints, we should use a short sliding window. In particular, if Assumption 1 holds for $\alpha = 0$, meaning that the number of breakpoints is constant w.r.t. the time horizon, and we use a sliding window $\tau \propto \sqrt{N}$, the order of the dynamic pseudo-regret is $\tilde{O}(\sqrt{N})$. Interestingly, even in the basic setting with a single breakpoint ($\alpha = 0$ and $B_N = 1$) and two arms, the sliding window approach outperforms classical MAB algorithms for stationary settings, *e.g.*, UCB1. Indeed, MAB algorithms for stationary settings would suffer from $\Omega(\sqrt{N})$ dynamic pseudo-regret in the second phase, in addition to the regret due to the first phase.[8]

Conversely, if Assumption 1 holds with $\alpha = 1$, and consequently with $B < 1$, the above bound would provide a linear upper bound on the dynamic pseudo-regret over the time horizon. In this case, the interest is in bounding the average pseudo-regret, as stated in the following corollary.

Corollary 2. *If the SW-TS policy is run over an AC-MAB setting in which Assumption 1 holds with $\alpha = 1$, $0 < B < 1$, and using a sliding window $\tau \propto \sqrt{\frac{\log\left(\frac{1}{B}\right)}{B}}$, the average pseudo-regret is:*

$$\overline{AR}_N(\mathfrak{U}) = O(\sqrt{-B \log B}).$$

8. For instance, be given two arms a_1, a_2 and a breakpoint $b_1 = N/2$ in which the expected values of the arms switch (*e.g.*, a_1 is better than a_2 before $N/2$ and worse after). After $N/2$, $O(\sqrt{N})$ pulls of a_1 are required before the upper confidence bound of a_2 is higher than that one of a_1. It is easy to see that, in this situation, the algorithm suffers from a dynamic pseudo-regret of at least $\Omega(\sqrt{N})$.

Finally, we remark that the bound provided in Theorem 1 has the order of $O(\log N)$ when we have a single phase and $\tau = T$, *i.e.*, in such a setting we have $N_\phi = T$, $B_N = 0$ and Assumption 1 holds when $B = 0$. Such a result is consistent with the upper bounds on the expected pseudo-regret of Thompson Sampling algorithm provided by Kaufmann et al. (2012b) and Agrawal and Goyal (2012).

4.3 Finite-Time Analysis in the Smoothly Changing Setting

We provide a finite-time analysis of the dynamic pseudo-regret achieved by the SW-TS algorithm, in the SC-MAB setting introduced in Section 3.2.

Theorem 2. *If the SW-TS policy is run over a SC-MAB setting with $X_{i,t} \sim Be(\mu_{i,t})$, Lipschitz constant $\sigma > 0$ and there exists $\Delta_0 \in (0,1)$ as in Assumption 3, for any $\tau \in \mathbb{N}$ s.t. $2\sigma\tau < \Delta \leq \Delta_0$, the dynamic pseudo-regret after N rounds is at most:*

$$\overline{R}_N(\mathfrak{U}) \leq F\Delta N^\beta + \frac{NK}{\tau}\left[\frac{52\log\tau}{(\Delta - 2\sigma\tau)^2} + \log\tau + 5 + \frac{19}{\log\tau}\right] + \\ + K\left[\frac{52\log\tau}{(\Delta - 2\sigma\tau)^2} + 3 + \frac{19}{\log\tau}\right].$$

The dependence of the dynamic pseudo-regret on the factor $\frac{N}{\tau}$ is similar to the result obtained by Combes and Proutiere (2014) and Garivier and Moulines (2008) for frequentist algorithms. In what follows, depending on the value of the parameter β in Assumption 3, we can show that either the dynamic pseudo-regret of the SW-TS algorithm is sublinear in the time horizon N or it is linear.

Depending on the characteristic of the specific SC-MAB setting, we can provide different results in terms of dynamic pseudo-regret once we fix the sliding window size. More specifically, one of the parameter characterizing the dynamic pseudo-regret is P, *i.e.*, the number of times that the expected rewards of a couple of arms switch over the time horizon, or formally:

$$P = |\{t \in \{1, \dots, N-1\} \text{ s.t. } \exists i \neq j (\mu_{i,t} - \mu_{j,t})(\mu_{i,t+1} - \mu_{j,t+1}) < 0\}|.$$

If we do not have any switch between the expected rewards of the arms, we can ensure the following.

Corollary 3. *If the SW-TS policy is run over an SC-MAB setting with no switches between expected rewards of the arms ($P = 0$), in which Assumption 3 holds with $\beta \in [1 - \log_N\left(\frac{\Delta}{2\sigma}\right), 1]$, and using a sliding window $\tau := N^{1-\beta}$, for each $\Delta \leq \Delta_0$ the dynamic pseudo-regret is at most:*

$$\overline{R}_N(\mathfrak{U}) = \tilde{O}(N^\beta).$$

Notice that, as the parameter β increases, the sliding window size prescribed by Corollary 3 reduces. Intuitively, this is due to the fact that settings with a large value of β present a large number of rounds in which two arms are hard to be distinguished, *i.e.*, the difference between their expected rewards is smaller than Δ. This fact, in its turn, also shortens the phases in which the bandit algorithm is capable of properly operating and implies the use of a shorter sliding window. In particular, it is easy to prove that, if Assumption 3 holds

with $\beta = 0$, meaning that we have $\sigma \leq \frac{\Delta}{2N}$, we would have an SC-MAB setting in which the expected rewards do not switch over time. Therefore, using the prescribed time window of $\tau = N$ provides a logarithmic dynamic pseudo-regret.

Conversely, if we have at least one switch between the expected reward of the arms, the result on the dynamic pseudo-regret upper bound requires a further condition on the values of β to hold. More formally, we have the following:

Corollary 4. *If the SW-TS policy is run over an SC-MAB setting with $P \in \mathbb{N}$ switches between expected rewards of the arms, in which Assumption 3 holds with:*

$$\beta \in \left[\max\left\{1 - \log_N\left(\frac{\Delta}{2\sigma}\right), \frac{1}{2} - \log_N\sqrt{\frac{F\Delta}{P}}\right\}, 1\right],$$

where $\max\{a, b\}$ denotes the maximum between a and b, and using a sliding window $\tau := N^{1-\beta}$, F is defined in Assumption 3, for each $\Delta \leq \Delta_0$ the dynamic pseudo-regret is at most:

$$\overline{R}_N(\mathfrak{U}) = \tilde{O}(N^{\beta}).$$

If Assumption 1 holds with $\beta = 1$, the above corollaries provide an upper bound on the dynamic pseudo-regret, which is linear in the time horizon N, similarly to what provided by Combes and Proutiere (2014). Nonetheless, we can bound the average pseudo-regret as follows.

Corollary 5. *If the SW-TS policy is run over an SC-MAB setting in which Assumption 3 holds with $\beta = 1$ and using a sliding window $\tau \propto \sigma^{-\frac{3}{4}}$, the average pseudo-regret is:*

$$\overline{AR}_N(\mathfrak{U}) = \tilde{O}(\sigma^{\frac{1}{2}}).$$

Finally, if we have a stationary environment, *i.e.*, $\sigma = 0$, we are able to find Δ s.t. $|\mathcal{F}_{\Delta,N}| = 0$ and, therefore, we have that $\beta = 0$. Choosing a sliding window $\tau = N$, Theorem 2 provides an upper bound over the expected pseudo-regret of order $O(\log N)$, as it happens in the classical MAB literature, as discussed by Auer et al. (2002) and Auer et al. (2012b).

4.4 Finite-Time Analysis in the Abruptly and Smoothly Changing Setting

In this section, we provide theoretical guarantees of SW-TS in both the AC-MAB and SC-MAB settings. The main result follows.

Theorem 3. *If the SW-TS policy is run over an ASC-MAB setting with $X_{i,t} \sim Be(\mu_{i,t})$, Lipschitz constant $\sigma > 0$ as in Assumption 4 and there exists $\Delta_0 \in (0, 1)$ as in Assumption 3, for any $\tau \in \mathbb{N}$ s.t. $2\sigma\tau < \Delta \leq \Delta_0$, the dynamic pseudo-regret after N rounds is at most:*

$$\overline{R}_N(\mathfrak{U}) \leq F\Delta N^{\beta} + \tau B N^{\alpha} + \frac{NK}{\tau}\left[\frac{52\log\tau}{(\Delta - 2\sigma\tau)^2} + \log\tau + 5 + \frac{19}{\log\tau}\right],$$

where B and α are defined in Assumption 1 and F and β are defined in Assumption 3.

Similarly to what has been done in the other two scenarios, we provide the order of the derived upper bounds when the sliding window length τ has been set properly, depending on the values of the parameters α and β and the number of times the expected rewards switch P.

Corollary 6. *If the SW-TS policy is run over an ASC-MAB setting with no switches between expected rewards of the arms ($P = 0$) and Assumption 1 and Assumption 3 hold with $\alpha \in (1 - 2\log_N\left(\frac{\Delta}{2\sigma}\right), 1)$ and $\beta \in (0, 1)$, respectively, for each $\Delta \leq \Delta_0$, using a sliding window of $\tau := N^{\frac{1-\alpha}{2}}$, the dynamic pseudo-regret is at most:*

$$\overline{R}_N(\mathfrak{U}) = \begin{cases} \tilde{O}\left(N^{\frac{1+\alpha}{2}}\right) & \text{if } \beta \leq \frac{1+\alpha}{2} \\ \tilde{O}\left(N^{\beta}\right) & \text{if } \beta > \frac{1+\alpha}{2} \end{cases}.$$

Conversely, in the case we have some switches between the expected rewards over time, we require further conditions on the parameters α and β.

Corollary 7. *If the SW-TS policy is run over an ASC-MAB setting with $P \in \mathbb{N}$ switches between expected rewards of the arms, and Assumption 1 and Assumption 3 hold with $\alpha \in (1 - 2\log_N\left(\frac{\Delta}{2\sigma}\right), 1)$ and $\beta \in (0, 1)$, respectively, for each $\Delta \leq \Delta_0$, using a sliding window of $\tau := N^{\frac{1-\alpha}{2}}$, if $\beta + \frac{\alpha}{2} \geq \frac{1}{2} - \log_N\left(\frac{F\Delta}{P}\right)$ holds, the dynamic pseudo-regret is at most:*

$$\overline{R}_N(\mathfrak{U}) = \begin{cases} \tilde{O}\left(N^{\frac{1+\alpha}{2}}\right) & \text{if } \beta \leq \frac{1+\alpha}{2} \\ \tilde{O}\left(N^{\beta}\right) & \text{if } \beta > \frac{1+\alpha}{2} \end{cases}.$$

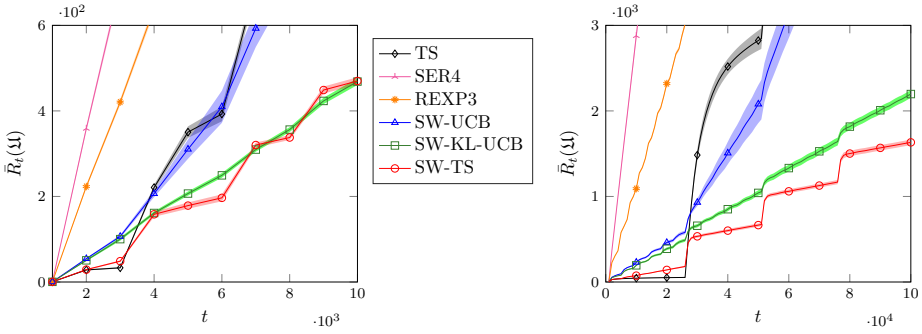
Intuitively, the results in Corollaries 6 and 7 state that, depending on which form of non-stationarity dominates, we have two different orders of dynamic pseudo-regret dependent on either α or β.

Similarly to the other two settings (AC-MAB and SC-MAB), if $\alpha = 1$ and $\beta = 1$, Theorem 3 provides an upper bound over the dynamic pseudo-regret over the time horizon N. Conversely, it is possible to bound the average pseudo-regret as follows:

Corollary 8. *If the SW-TS policy is run over an SC-MAB setting in which Assumption 1 holds with $\alpha = 1$, Assumption 3 holds with $\beta = 1$, and using a sliding window $\tau \propto B^{-\frac{1}{4}}\sigma^{-\frac{3}{4}}$, the average pseudo-regret is:*

$$\overline{AR}_N(\mathfrak{U}) = \tilde{O}(B^{\frac{1}{2}}\sigma^{\frac{1}{2}}).$$

The asymptotic order of SW-TS in the ASC-MAB setting upper bound reduces to the one of Theorem 2 in the case we have $B = 0$, *i.e.*, we are in an SC-MAB setting. If we apply the bound in Theorem 3 for the AC-MAB setting, in which we have $\sigma = 0$ and by fixing $\Delta = \min_i \Delta_i$ we have $|\mathcal{F}_{\Delta,N}| = 0$, thus $F = 0$, we obtain a slightly less accurate bound in terms of Δ than the one provided in Theorem 1. Nonetheless, the bound presents the same order in terms of N and τ. Finally, if we are in a stationary setting, *i.e.*, $F = B = 0$, the bound over the expected pseudo-regret reduces to the order of the one provided by the Thompson Sampling algorithm, analyzed by Kaufmann et al. (2012b) and Agrawal and Goyal (2012).



(a) (b)

Figure 1: AC-MAB: average dynamic pseudo-regret $\bar{R}_t(\mathfrak{U})$ and 95% confidence intervals. Settings with $K = 10$ arms and time horizon $N = 10^4$ (a), and with $N = 10^5$ (b).

5. Experimental Evaluation

We experimentally evaluate our algorithm w.r.t. the state-of-the-art passive algorithms with theoretical guarantees in terms of pseudo-regret performance in the AC-MAB, SC-MAB, and ASC-MAB settings. In particular, we compare SW-TS with Thompson Sampling (TS) by Thompson (1933) to evaluate the improvement obtained thanks to the employment of a sliding window τ. Furthermore, we compare SW-TS with REXP3 by Besbes et al. (2014), SW-UCB by Garivier and Moulines (2008), SW-KL-UCB by Combes and Proutiere (2014) and SER4 by Allesiardo et al. (2017) to evaluate the improvement obtained thanks to the adoption of Bayesian methods vs. frequentist ones in non-stationary settings.[9] The figures of merit we consider are the dynamic pseudo-regret $\overline{R}_N(\mathfrak{U})$, as defined in Equation (1), and the corresponding 95% confidence intervals (reported in the figures as semi-transparent areas) computed over 100 independent runs, if not specified otherwise. Finally, we perform a sensitivity analysis on the two parameters whose knowledge is required by the SW-TS algorithm, specifically α and β.

5.1 Abruptly Changing MAB Setting

Experimental Setting We use a time horizon $N \in \{10^4, 10^5, 10^6\}$ and a number of arms $K \in \{5, 10, 20, 30\}$. We split the time horizon N into four phases of equal length. The expected value $\mu_{i,\phi}$ is chosen randomly for every arm a_i during each phase. In particular, after every breakpoint, the expected value $\mu_{i,\phi}$ of each arm a_i is drawn from a uniform probability distribution over $[0, 1]$, thus assuring that there is never the same optimal arm in two different phases, *i.e.*, $a_{i^*_\phi} \neq a_{i^*_{\phi'}}$, $\forall \phi, \phi'$ with $\phi \neq \phi'$. For the sake of comparison, we choose a sliding window $\tau = 4\sqrt{N \log(N)}$ as is discussed by Garivier and Moulines (2008). We generate 10 configurations for every combination of N and K as discussed above, and

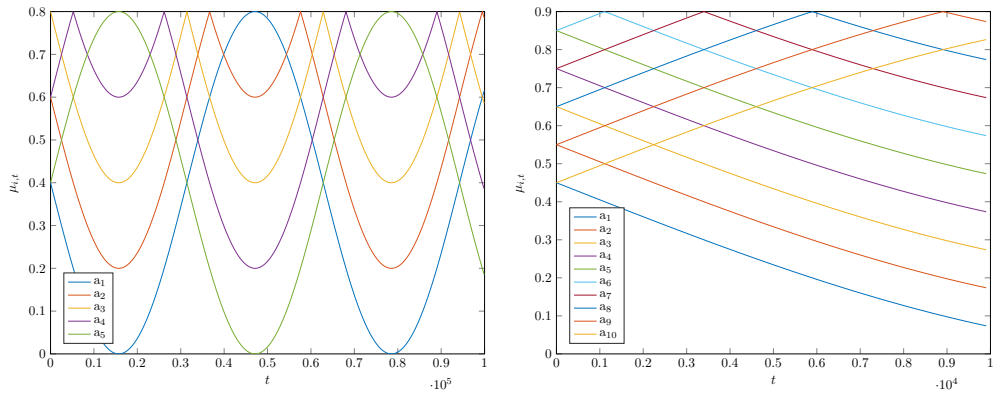
9. If not specified otherwise, the parameters of REXP3 and SER4 are set as in Corollary 3 provided by Allesiardo et al. (2017) and Theorem 2 provided by Besbes et al. (2014), respectively, since these values allow the two algorithms to have sublinear regret.

Table 2: AC-MAB: average dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ and 95% confidence intervals. Best results on average have been highlighted in boldface.

			N		
			10^4	10^5	10^6
K	5	TS	1317±52.89	12857±425.68	114476±4836.98
		SER4	2494±37.63	25601±513.99	238034±4323.34
		REXP3	1451±13.70	8448±55.21	42561±212.75
		SW-UCB	824±66.80	5687±814.94	32939±7587.28
		SW-KL-UCB	**344±7.57**	1570±31.51	6248±145.51
		SW-TS	437±13.37	**1467±30.45**	**4904±39.00**
	10	TS	1251±26.90	10927±315.30	98312±4168.24
		SER4	3151±34.63	31454±499.91	279232±6504.65
		REXP3	1913±17.85	12170±108.38	61978±345.25
		SW-UCB	1116±68.46	8143±872.37	49537±6191.14
		SW-KL-UCB	**469±7.98**	2197±45.54	8601±162.32
		SW-TS	**470±8.82**	**1632±32.85**	**5493±92.67**
	20	TS	1130±30.91	8864±139.77	69919±2447.98
		SER4	3684±26.76	33293±167.89	293844±3038.42
		REXP3	2480±17.27	16134±93.65	83042±337.96
		SW-UCB	1405±57.44	11789±503.34	68751±6651.74
		SW-KL-UCB	652±6.70	3086±48.22	11921±315.74
		SW-TS	**536±10.26**	**1858±26.82**	**6156±149.64**
	30	TS	1016±35.55	7714±170.92	61979±2001.15
		SER4	3922±19.23	33622±212.29	285382±1727.97
		REXP3	2712±22.37	18432±100.09	96851±378.67
		SW-UCB	1566±60.42	12271±804.93	82006±8424.70
		SW-KL-UCB	770±19.79	3858±84.94	15287±233.75
		SW-TS	**575±12.20**	**2067±35.65**	**7123±96.46**

we provide the results averaged over the configurations and over 100 independent trials for each configuration.

Results The numerical results in terms of $\overline{R}_N(\mathfrak{U})$ are reported in Table 2. For every combination of N and K, we highlight in bold the minimum value of $\overline{R}_N(\mathfrak{U})$ achieved. SW-TS outperforms the other algorithms in all the configurations except for the setting with $N = 10^4$ and $K = 5$ where SW-KL-UCB outperforms SW-TS. In the setting with $N = 10^4$ and $K = 10$ there is no statistical evidence to determine which algorithm is the best between SW-TS and SW-KL-UCB since the 95% confidence intervals overlap.



(a) (b)

Figure 2: SC-MAB: examples of evolution of the expected reward $\mu_{i,t}$ of the arms over time. Settings with $N = 10^4$, $K = 5$ (a), and $N = 10^5$, $K = 10$ (b).

In Figure 1, we report the results for the settings with $K = 10$ as t varies. It can be observed that, with $N = 10^4$ (Figure 1a), the performance of SW-TS and SW-KL-UCB are similar. However, the regret obtained by the algorithms is almost linear, suggesting that the algorithms are not able to learn since the problem is excessively hard. With a longer time horizon of $N = 10^5$ (Figure 1b), the sliding window τ becomes larger (we recall that we use a τ depending on N), as well as the length of the phases and, thus, SW-TS outperforms SW-KL-UCB. The SW-TS suffers from a larger regret when we enter a new phase, *e.g.*, around $t = 5 \cdot 10^4$, but once the sliding window discards the samples coming from the previous phase SW-TS can learn faster than other algorithms, which is exemplified by the lower slope of the regret between $t = 6 \cdot 10^4$ and $t = 7 \cdot 10^4$.

5.2 Smoothly Changing MAB Setting

Experimental Setting We use a time horizon $N \in \{10^4, 10^5, 10^6\}$ and a number of arms $K \in \{5, 10, 20, 30\}$. We replicate the experimental setting of Combes and Proutiere (2014), where the expected value $\mu_{i,t}$ of arm a_i changes according to the following function:

$$\mu_{i,t} = \frac{K-1}{K} - \frac{|w(t) - i|}{K},$$
$$w(t) = 1 + \frac{(K-1)(1+\sin(t\sigma))}{2}.$$

Examples of the evolution over time of the expected reward of the arms are presented in Figure 2.

We use two different sliding window lengths. In the first case, we use the values of the parameters τ and σ prescribed by our theoretical results. In particular, with the above experimental setting, Assumption 3 is satisfied for every value of $N \in \{10^4, 10^5, 10^6\}$ when $\beta = \frac{1}{2}$ and $\sigma = 0.0001$.[10] Such a value of β leads to $\tau = \sqrt{N}$. In the second case, we

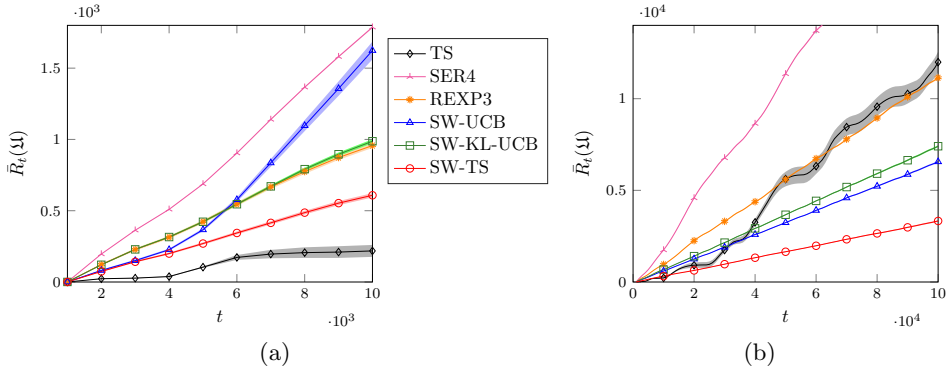
10. Details on the conditions for which Assumption 3 is satisfied are provided in Appendix E.

Table 3: SC-MAB with $\tau = \sqrt{N}$: average dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ and 95% confidence intervals. Best results on average have been highlighted in boldface.

			N		
			10^4	10^5	10^6
K	5	TS	**218±41.94**	11995±562.73	161933±3767.54
		SER4	1787±6.61	22398±61.49	212095±309.93
		REXP3	957±16.58	11141±58.95	111403±169.79
		SW-UCB	1624±62.40	6560±49.31	34615±160.23
		SW-KLUCB	987±13.79	7407±50.99	40190±165.42
		SW-TS	608±15.43	**3330±40.60**	**16403±117.46**
	10	TS	**520±42.19**	13253±579.05	169850±4434.77
		SER4	2206±14.66	26094±145.80	242464±498.22
		REXP3	1253±19.36	14391±63.09	144581±199.42
		SW-UCB	3424±105.37	36622±314.36	80256±4375.85
		SW-KLUCB	1289±12.56	11009±50.77	64518±179.97
		SW-TS	922±16.18	**5529±49.82**	**28258±149.82**
	20	TS	**549±26.74**	12843±390.73	173140±2772.27
		SER4	2361±32.85	27286±259.41	258380±1039.26
		REXP3	1470±16.91	17334±67.77	174065±219.54
		SW-UCB	4466±202.74	45089±353.93	448649±155.59
		SW-KLUCB	1330±12.17	13630±38.60	89442±157.88
		SW-TS	1180±15.48	**7971±49.70**	**44186±154.32**
	30	TS	**581±26.29**	12483±297.63	172305±2205.39
		SER4	2480±23.17	27872±354.54	279190±1680.75
		REXP3	1607±14.59	18854±59.28	189734±178.99
		SW-UCB	4348±329.65	47586±911.96	462611±443.20
		SW-KLUCB	1638±11.92	14603±37.31	102707±126.91
		SW-TS	1339±11.49	**9595±39.76**	**54298±124.20**

use the value of the parameter τ used by Combes and Proutiere (2014) to provide a direct comparison between the SW-TS and SW-KLUCB algorithms. Thus, we set $\tau = \sigma^{-\frac{4}{5}}$. Let us notice that such a choice is not optimal for SW-TS according to our theoretical results. Furthermore, we do not set σ equal to a fixed value, but we evaluate the performance of the algorithms as σ varies. In both cases, we average the results over 100 independent trials for every combination of N, K and σ.

Results First, we analyze the results for the settings with $\tau = \sqrt{N}$. The numerical results in terms of $\bar{R}_N(\mathfrak{U})$ are reported in Table 3. The SW-TS algorithm outperforms all the other



(a) (b)

Figure 3: SC-MAB: average dynamic pseudo-regret $\bar{R}_t(\mathfrak{U})$ and 95% confidence intervals. Settings with $K = 5$ arms and time horizon $N = 10^4$ (a), and $N = 10^5$ (b).

ones, except for the case with $N = 10^4$, in which SW-TS achieves the best performance w.r.t. the other algorithms using a sliding window, but it is not able to outperform TS. The reason behind this behaviour lies in the fact that, with such a small value of σ, the optimal arm remains the same until round $t = 5 \cdot 10^3$ and it is not convenient to use a sliding window approach. Conversely, if we have longer time horizons, the optimal arm changes more often. In particular, with $N = 10^5$, we have 14 changes of the optimal arm, and the performance of TS becomes the worst. In Figure 3a, we report the dynamic pseudo-regret $\bar{R}_t(\mathfrak{U})$ of the analysed algorithms as t varies in the case with $N = 10^4$. It can be observed that, when the optimal arm changes, there is a worsening in the dynamic pseudo-regret performance of TS. However, no sliding window algorithm can achieve its performance. Conversely, as it can be observed in Figure 3b, when $N = 10^5$, TS is outperformed by almost all the other algorithms, and this is because the optimal arm changes multiple times. Even if TS and SW-TS share similar behaviours in the very first rounds, the use of a sliding window allows SW-TS to provide the best performance on a longer time horizon.

Second, we analyze the results for the settings with $\tau = \sigma^{-\frac{4}{5}}$. The results in terms of dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ are reported in Table 4 for the experiments with $\sigma = 10^{-3}$. We observe that SW-TS outperforms all the other algorithms, providing in every setting the minimum $\bar{R}_N(\mathfrak{U})$ (highlighted in bold). In Figure 4, we report the dynamic pseudo-regret $\bar{R}_t(\mathfrak{U})$ as t varies in the setting with $N = 10^4$, $K = 10$ and $\sigma = 10^{-3}$. It can be observed that, in the first $4 \cdot 10^3$ rounds, TS outperforms SW-KL-UCB, but, subsequently, thanks to the use of a sliding window, SW-KL-UCB forgets the past and improves its performance. Instead, SW-TS outperforms all the other algorithms for the whole time horizon. Notably, REXP3 achieves performance similar to the one of TS, while TS outperforms SW-UCB even if this latter algorithm employs a sliding window approach. Finally, in Figure 5, we report the dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ as σ varies in the setting with $N = 10^4$. We observe that SW-TS outperforms all the other algorithms for each value of σ. As the number of arms K increases from $K = 5$ to $K = 30$, the performance of REXP3 and SW-KL-UCB gets worse and, with $\sigma = 0.01$, TS (without sliding window) outperforms both REXP3 and SW-KL-UCB. The setting with the other values of N, K and σ, are not reported since they

Table 4: SC-MAB with $\tau = \sigma^{-\frac{4}{5}}$: average dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ and 95% confidence intervals. Setting with $\sigma = 10^{-3}$. Best results on average have been highlighted in boldface.

			N		
			10^4	10^5	10^6
K	5	TS	1374±41.16	22965±186.32	211996±209.56
		REXP3	1094±6.49	11915±21.08	118999±96.21
		SW-UCB	677±8.24	7533±127.33	92532±4929.63
		SW-KL-UCB	752±7.18	8249±35.44	82469±317.70
		SW-TS	**423±8.14**	**4629±33.14**	**46291±269.55**
	10	TS	1419±36.45	25969±276.69	266539±210.95
		REXP3	1426±7.70	15505±23.81	155050±92.10
		SW-UCB	3247±27.03	41238±43.84	422468±38.42
		SW-KL-UCB	1093±7.75	11973±31.21	119631±238.20
		SW-TS	**690±8.03**	**7664±30.12**	**76583±199.37**
	20	TS	1442±29.13	26623±226.21	292824±228.08
		REXP3	1727±6.89	18863±22.90	188579±65.92
		SW-UCB	3987±41.67	45247±38.81	458502±41.66
		SW-KL-UCB	1321±7.02	14507±26.76	145215±134.47
		SW-TS	**944±6.95**	**10413±24.59**	**104074±88.18**
	30	TS	1401±26.42	26887±201.09	302460±219.72
		REXP3	1887±6.92	20529±24.08	205269±93.17
		SW-UCB	4144±71.81	46490±69.65	470566±70.75
		SW-KL-UCB	1383±8.81	15121±25.55	151357±90.92
		SW-TS	**1104±7.30**	**12107±21.06**	**120868±71.62**

provide results in line with what has been presented before, in which SW-TS outperforms state-of-the-art algorithms.

5.3 Abruptly and Smoothly Changing MAB Setting

Experimental Setting We use a time horizon $N \in \{10^4, 10^5, 10^6\}$ and a number of arms $K \in \{5, 10, 20, 30\}$. For each setting, we generate the expected reward of the arms $\mu_{i,t}$ as follows. We split the time horizon N into four phases $\mathcal{F}_\phi$ of equal length and the set the expected value $\mu_{i,t}$ of arm a_i according to the following function:

$$\mu_{i,t} = \frac{K-1}{K} - \frac{|w(t) - i|}{K},$$

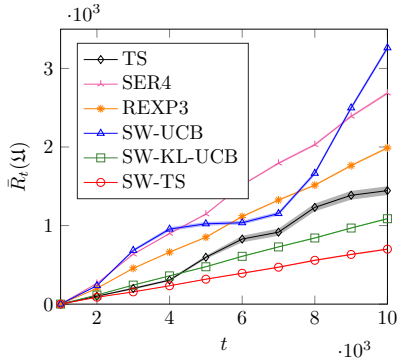


Figure 4: SC-MAB with $\tau = \sigma^{-\frac{4}{5}}$: average dynamic pseudo-regret $\bar{R}_t(\mathfrak{U})$ and 95% confidence intervals. Setting with number of arms $K = 10$, time horizon $N = 10^4$, and $\sigma = 10^{-3}$.

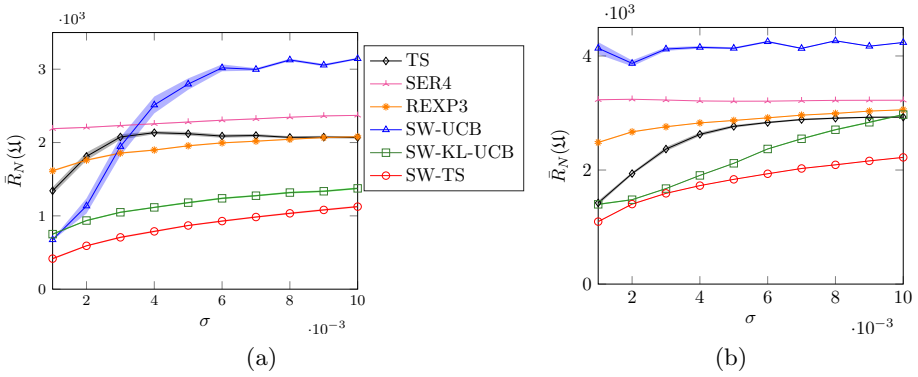


Figure 5: SC-MAB with $\tau = \sigma^{-\frac{4}{5}}$: average dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ and 95% confidence as the value of σ varies. Setting with number of arms $K = 5$ (a) and $K = 30$ (b), time horizon $N = 10^4$.

$$w(t) = 1 + \frac{(K-1)(1+\sin(t\sigma + \frac{N(\phi-1)}{4}))}{2},$$

where $\phi \in \{1, \ldots, 4\}$ represents the index of the phase. Note that the presence of the second term in the argument of the sine induces an abrupt change at the beginning of each phase by shifting the argument of the sine by an amount proportional to the time horizon N: after the first breakpoint, we shift of an number of rounds of $\frac{N}{4}$; after the second one, $\frac{N}{2}$ of N; after the third one, $\frac{3N}{4}$. As in Section 5.2, we run a set of experiments using both a sliding window $\tau = \sqrt{N}$ and $\sigma = 0.0001$, and a sliding window $\tau = \sigma^{-\frac{4}{5}}$ and $\sigma \in \{0.001, 0.002, \ldots, 0.01\}$. In both cases, we average the results over 100 independent trials for every combination of N, K and σ.

Table 5: ASC-MAB with $\tau = \sqrt{N}$: average dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ and 95% confidence intervals. Setting with $\sigma = 10^{-4}$. Best results on average have been highlighted in boldface.

			N		
			10^4	10^5	10^6
K	5	TS	**416±42.09**	11598±294.55	155795±3325.74
		SER4	2136±10.24	22894±63.04	211904±386.48
		REXP3	984±18.45	11428±55.26	111657±166.40
		SW-UCB	1424±80.93	6565±58.97	34832±162.74
		SW-KLUCB	991±16.66	7367±51.78	40395±171.51
		SW-TS	587±15.28	**3376±51.61**	**16546±143.17**
	10	TS	**513±33.02**	13322±399.70	166233±4166.45
		SER4	2677±28.61	26590±163.43	243132±652.63
		REXP3	1391±20.59	14652±65.30	144851±231.14
		SW-UCB	3807±146.66	51669±13.16	82394±5545.59
		SW-KLUCB	1434±15.94	10994±42.74	64783±166.04
		SW-TS	967±15.41	**5512±47.74**	**28539±143.63**
	20	TS	**598±30.29**	13099±243.57	172319±3124.95
		SER4	2832±56.99	28280±242.18	258039±1055.98
		REXP3	1664±18.84	17838±62.91	173867±198.21
		SW-UCB	5085±229.91	56948±35.71	450598±200.35
		SW-KLUCB	1544±12.00	13718±40.59	89463±139.87
		SW-TS	1280±12.89	**8017±52.56**	**44306±115.80**
	30	TS	**589±21.37**	12854±288.97	171534±2879.85
		SER4	2929±48.64	29651±457.64	278093±1590.27
		REXP3	1833±18.58	19585±53.51	189882±189.03
		SW-UCB	4819±418.89	59078±38.20	464238±565.19
		SW-KLUCB	1882±14.16	14718±33.40	102637±138.04
		SW-TS	1471±13.06	**9612±41.48**	**54363±134.34**

Results The results for both settings are similar to the ones presented in Section 5.2, suggesting that the abrupt changes not affect the dynamic pseudo-regret if the expected values of the arms are smoothly changing.

The numerical results in terms of dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ with $\tau = \sqrt{N}$ are reported in Table 5 for the experiments with $\sigma = 10^{-4}$. We observe that SW-TS outperforms all the other algorithms except for the case with $N = 10^4$, in which TS is the algorithm with the lowest dynamic pseudo-regret. The reason behind this behaviour lies in the fact that,

Table 6: ASC-MAB with $\tau = \sigma^{-\frac{4}{5}}$: average dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ and 95% confidence intervals. Setting with $\sigma = 10^{-3}$. Best results on average have been highlighted in boldface.

			N		
			10^4	10^5	10^6
K	5	TS	1252±39.02	23163±305.47	211871±380.91
		SER4	2253±9.44	23133±55.87	226316±294.62
		REXP3	1661±14.75	17517±42.07	175080±115.16
		SW-UCB	684±15.74	7451±74.99	90968±7174.24
		SW-KLUCB	754±9.62	8232±50.02	83436±359.15
		SW-TS	**470±11.90**	**4645±54.85**	**47197±298.63**
	10	TS	1370±39.51	26078±384.97	266289±283.49
		SER4	2745±23.31	29277±99.78	299947±456.21
		REXP3	2063±12.45	21719±49.39	217066±140.36
		SW-UCB	4736±62.98	41371±76.94	423199±99.35
		SW-KLUCB	1090±11.15	11958±48.21	120388±218.01
		SW-TS	**728±11.76**	**7630±47.36**	**77302±207.19**
	20	TS	1394±41.43	26573±341.85	292356±328.06
		SER4	3107±21.22	34016±46.53	341214±164.83
		REXP3	2389±14.57	25122±46.73	251113±143.90
		SW-UCB	5309±4.98	45305±77.67	459211±75.73
		SW-KLUCB	1345±9.73	14479±36.84	145437±133.78
		SW-TS	**980±9.92**	**10420±34.53**	**104146±107.82**
	30	TS	1401±27.51	26713±326.58	301882±317.67
		SER4	3252±13.63	35227±22.46	352542±78.03
		REXP3	2558±12.68	26911±40.75	268738±126.13
		SW-UCB	5513±4.21	46429±132.47	471276±0.00
		SW-KLUCB	1418±11.11	15161±37.74	151120±143.85
		SW-TS	**1136±10.84**	**12102±32.18**	**120625±115.00**

with such a small value for σ, the optimal arm remains the same until round $t = 5 \cdot 10^3$, and, therefore, it is not convenient to use a sliding window approach. Conversely, with longer time horizons, the optimal arm changes multiple times, and the performance of TS gets worse.

The results in terms of $\bar{R}_N(\mathfrak{U})$ with $\tau = \sigma^{\frac{4}{5}}$ are reported in Table 6 for the experiments with $\sigma = 10^{-3}$, and confirm the superior performance showed by the SW-TS algorithms in the SC-MAB setting.

6. Sensitivity Analysis

In this section, we present the results of the sensitivity analysis for parameter α in the AC-MAB when the sliding window is set as prescribed by Corollary 1, and for parameter β in the SC-MAB when the sliding window is set as prescribed by Corollary 3.

6.1 Abruptly Changing MAB Setting

Experimental Setting We compare the performance of the SW-TS algorithm using different sliding windows $\tau = N^{\frac{1-\alpha}{2}}$ with $\alpha \in \mathbb{A}$, where $\mathbb{A} = \{-1, -0.95, \dots, 0.95, 1\}$. We use the same set of arms $K \in \{5, 10, 20, 30\}$ previously used in Section 5.1, and a more fine-grained set of time horizons $N \in \{10^4, 10^5, 2 \cdot 10^5, 3 \cdot 10^5, \dots, 9 \cdot 10^5, 10^6\}$. The different configurations of expected values $\mu_{i,\phi}$ for each arm a_i are generated as described in Section 5.1. The results are averaged over these configurations and over 100 independent trials for each configuration. In addition to the sensitivity analysis w.r.t. α, we also show the change in terms of cumulative per-round pseudo-regret $\hat{R}_N := \bar{R}_N(\mathfrak{U})/N$ as the value of α varies.

Results In Figure 6, we report the values of α (denoted with α^* and using the solid red line), minimizing the dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$, as the length of the time horizon N varies. We also report the value of α (denoted with α_0 and using the solid blue line) prescribed by Corollary 1; in this case, $\alpha_0 = 0$. For a better comprehension of how the dynamic pseudo-regret increases as α gets far from the optimal α^*, we also plot the values of α corresponding to an increase of 50%, 100% and 200% of the dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ w.r.t. the one achieved with α^*. These curves are denoted with $\alpha_{50\%}$, $\alpha_{100\%}$ and $\alpha_{200\%}$, respectively, and are reported using dashed lines (notice that we have two curves for every $\alpha_{\%}$, one for α larger than α^* and another for α smaller than α^*; the curves above $\alpha = 1$ and below $\alpha = -1$ are omitted). It can be observed that, when using α_0, the increase in terms of dynamic pseudo-regret is always lower than 50% w.r.t the dynamic pseudo-regret achieved with α^*, α_0 being always between the two curves corresponding to $\alpha_{50\%}$. Moreover, α^* always corresponds to a negative value, suggesting that, on average, a sliding window τ longer than the one obtained with α_0 is preferable in these settings.

In Figure 7, for every number K of arms, we show the curves of the per-round pseudo-regret $\hat{R}_N$ for every different value of the time horizon N as the value of α varies. The minimum of each curve corresponds to α^* for the considered time horizon N. It can also be observed that the minimum per-round pseudo-regret is always achieved with almost the same value of $\alpha \approx -0.4$. Moreover, $\hat{R}_N$ grows faster moving from α^* to larger values of α rather than to smaller values of α, thus suggesting that underestimating the value of α to use in the algorithm, and therefore overestimating the length of the sliding window, is safer than overestimating it.

In Figure 8, for every value of the time horizon N, we show the curves of the per-round pseudo-regret $\hat{R}_N$ for every different number K of arms as the value of α varies. The minimum of each curve corresponds to α^* for the considered number of arms K. We observe that the value of α^* decreases as the number K of arms increases. Intuitively, this behavior is due to the fact that, when SW-TS has more arms to play, it also needs to collect

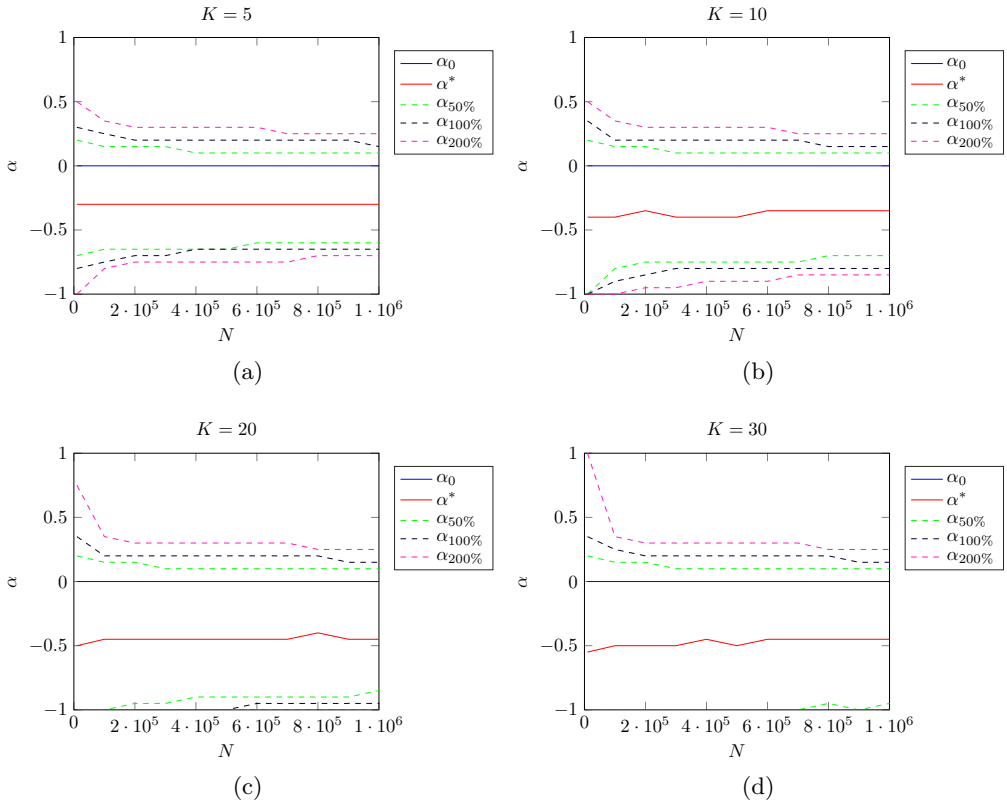


Figure 6: AC-MAB: values of α providing the best dynamic pseudo-regret (α^*), the best dynamic pseudo-regret increased by 50% ($\alpha_{50\%}$), by 100% ($\alpha_{100\%}$), and by 200% ($\alpha_{200\%}$) as the time horizon N varies; α_0 is the value prescribed by Corollary 1.

more samples for each arm in order to identify which is the optimal arm and, consequently, a larger sliding window τ is preferable.

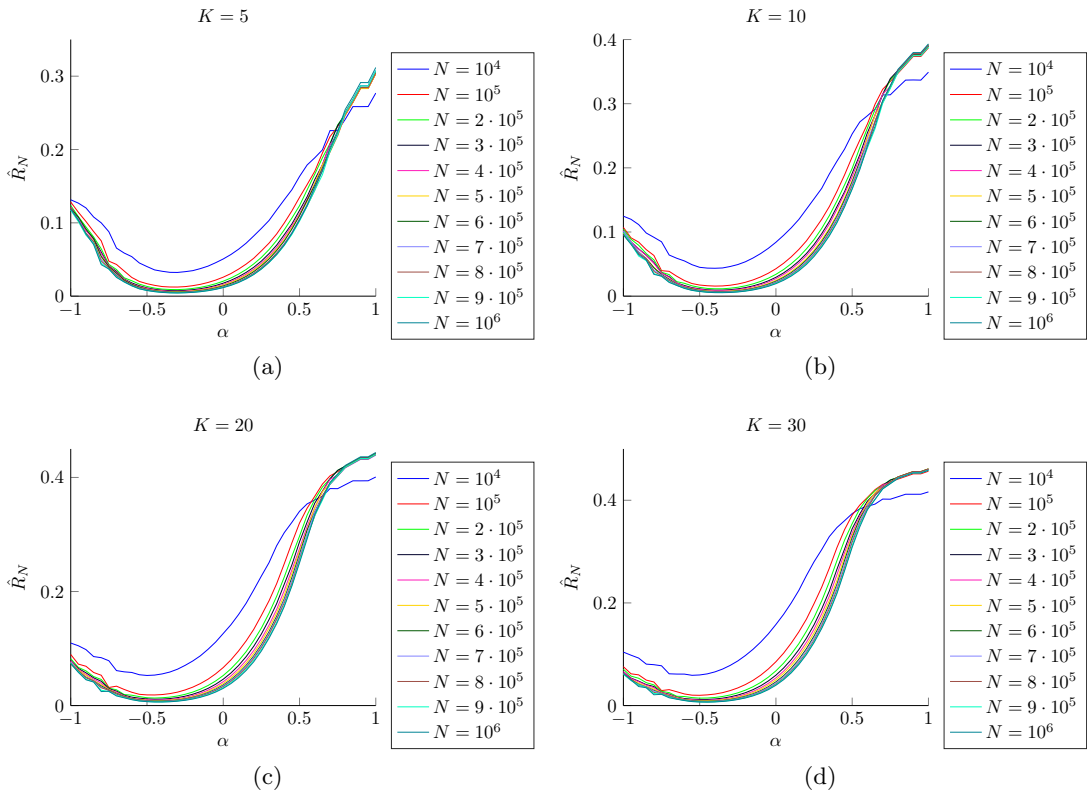


Figure 7: AC-MAB: per-round pseudo-regret $\hat{R}_N$ as the value of $\alpha \in \mathbb{A}$ varies for every value of the time horizon N and every different number of arms K.

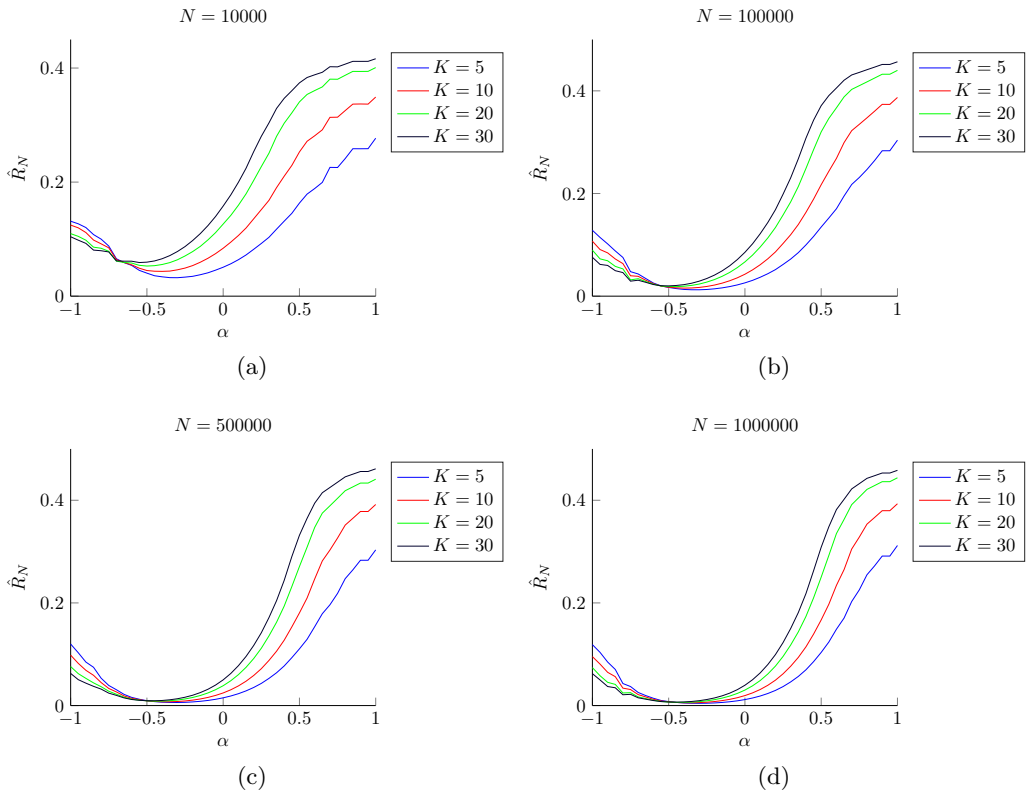


Figure 8: AC-MAB: Per-round pseudo-regret $\hat{R}_N$ as the value of $\alpha \in \mathbb{A}$ varies for every number of arms K and every different length of the time horizon N.

6.2 Smoothly Changing MAB Setting

Experimental Setting We compare the performance of the SW-TS algorithm using different sliding windows $\tau = N^{\frac{1-\beta}{2}}$ with $\beta = \mathbb{B}$, where $\mathbb{B} = \{-1, -0.95, \dots, 0.95, 1\}$. We use the same set of arms $K \in \{5, 10, 20, 30\}$ previously used in Section 5.1, and a time horizon with a length in $N \in \{10^4, 10^5, 10^6\}$. The expected value of the rewards $\mu_{i,t}$ for arm a_i changes as described in Section 5.2.

Results In Figure 9, we report the values of β (denoted with β^* and using the solid red line), minimizing the dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$, as the length of the time horizon N varies. We also report the value of β (denoted with β_0 and using the solid blue line) prescribed by Corollary 3; in this case, $\beta_0 = 0$. For a better comprehension of how the dynamic pseudo-regret increases as β gets far from the optimal β^*, we also plot the values of β corresponding to an increase of 50%, 100% and 200% of the dynamic pseudo-regret $\bar{R}_N(\mathfrak{U})$ w.r.t. the one achieved with β^*. These curves are denoted with $\beta_{50\%}$, $\beta_{100\%}$ and $\beta_{200\%}$ and are reported using dashed lines. It can be observed that, when using β_0, the increase in terms of dynamic pseudo-regret is always lower than the 50% w.r.t. the dynamic pseudo-regret achieved with β^*, β_0 being always between the two curves corresponding to $\beta_{50\%}$. As in the analysis for the parameter α, we notice that β^* always corresponds to a negative value, suggesting that a sliding window τ longer than the one obtained with β_0 is preferable.

In Figure 10, for every number K of arms, we show the curves of the per-round pseudo-regret $\hat{R}_N$ for every different value of the time horizon N as the value of β varies. The minimum of the curves corresponds to β^* for the considered time horizon N. It can also be observed that the minimum per-round pseudo-regret is always achieved with a value of $\beta \approx -0.2$. Moreover, $\hat{R}_N$ grows faster moving from β^* toward larger values of β rather than towards smaller values of β. Such behavior suggests that, in practice, using a sliding window τ slightly longer than the one of the optimal β^* is preferable w.r.t. a slightly smaller one.

In Figure 11, for every value of the time horizon N, we show the curves of the per-round pseudo-regret $\hat{R}_N$ for every different number K of arms as the value of β varies. The minimum of the curves corresponds to β^* for the considered number of arms K. We observe that the larger the number of arms K, the lower the value of β^*, which is in line with what has been observed for parameter α.

7. Conclusions and Future Work

In this paper, we study the non-stationary Multi-Armed Bandit problem, investigating settings in which two different forms of non-stationarity are present: the abruptly changing one, namely AC-MAB, in which the expected reward of the arm remains the same for sequences of rounds, and it changes at unknown rounds, and the smoothly changing one, namely SC-MAB, in which the expected reward of every arm continuously changes over rounds in a smooth way. Besides, we also study the setting in which both the non-stationarity forms coexist, namely ASC-MAB. We propose an algorithm, namely Sliding-Window Thompson Sampling (SW-TS), which chooses the next arm to play on the basis of the information collected in the last τ rounds. We derive an upper bound on the dynamic pseudo-regret

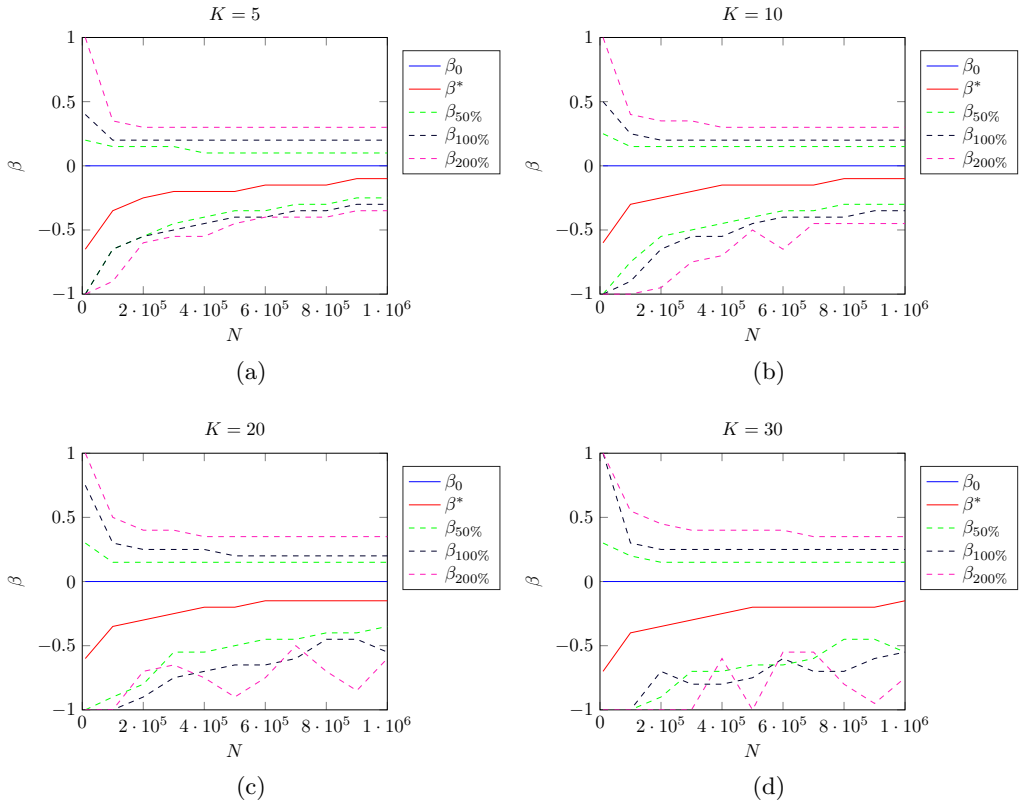


Figure 9: SC-MAB: values of β providing the best dynamic pseudo-regret (β^*), the best dynamic pseudo-regret increased by 50% ($\beta_{50\%}$), by 100% ($\beta_{100\%}$), and by 200% ($\beta_{200\%}$) as the time horizon N varies; β_0 is the value prescribed by Corollary 3.

for SW-TS when it is applied to one of the settings mentioned before. Finally, we present a thorough experimental evaluation of the performance of the SW-TS algorithm separately for the AC-MAB, SC-MAB, and ASC-MAB settings. We show that our algorithm significantly outperforms state-of-the-art approaches for NS-MAB problems in terms of dynamic pseudo-regret, achieving the lowest pseudo-regret in almost all the configurations we tested. Future development of this work may consider an analysis of our algorithm in MAB problems with structure, *e.g.*, the Unimodal MAB, in which the arm space presents a graph structure, the case of continuous decision space as studied by Nuara, Trovó, Crippa, Gatti, and Restelli (2020), and adversarial settings as studied by Marchesi, Trovó, and Gatti (2020) and Bisi, Nittis, Trovò, Restelli, and Gatti (2017). Moreover, another research line we intend to explore is the derivation of bounds depending on the characteristics of the abrupt changes, such as their magnitude and their distribution over the time horizon.

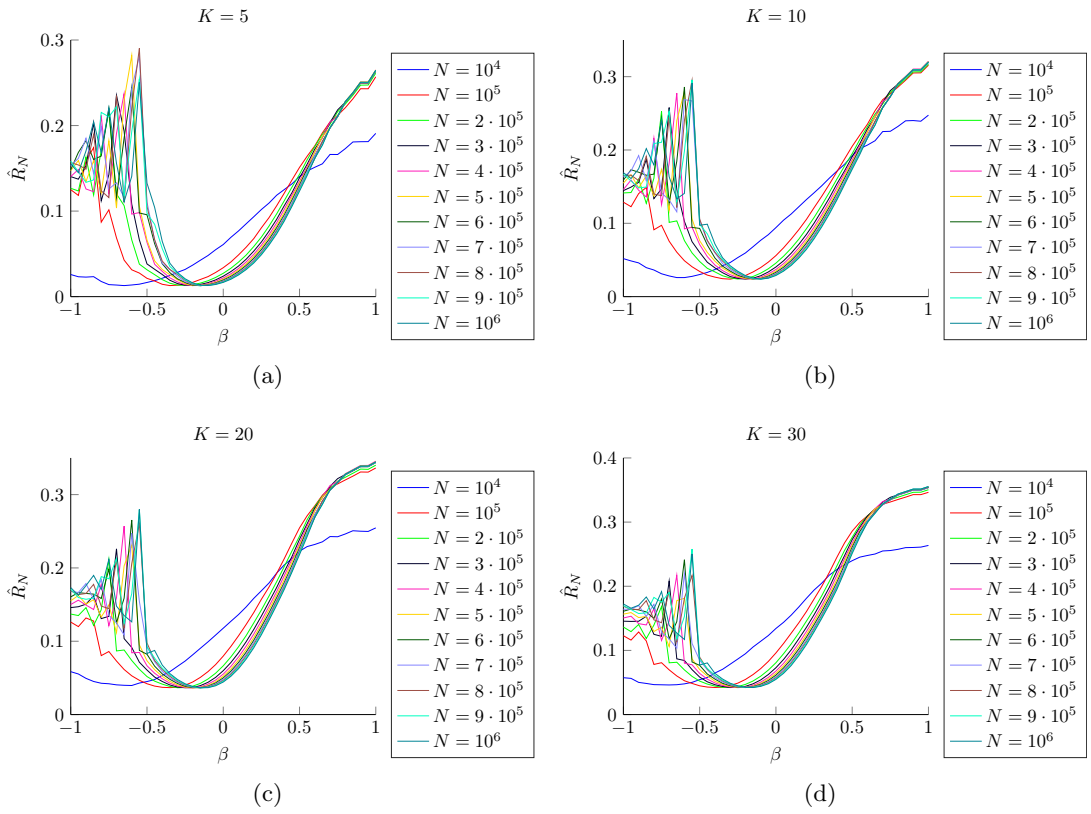


Figure 10: SC-MAB: per-round pseudo-regret $\hat{R}_N$ as the value of $\beta \in \mathbb{B}$ varies for every value of the time horizon N and every different number of arms K.

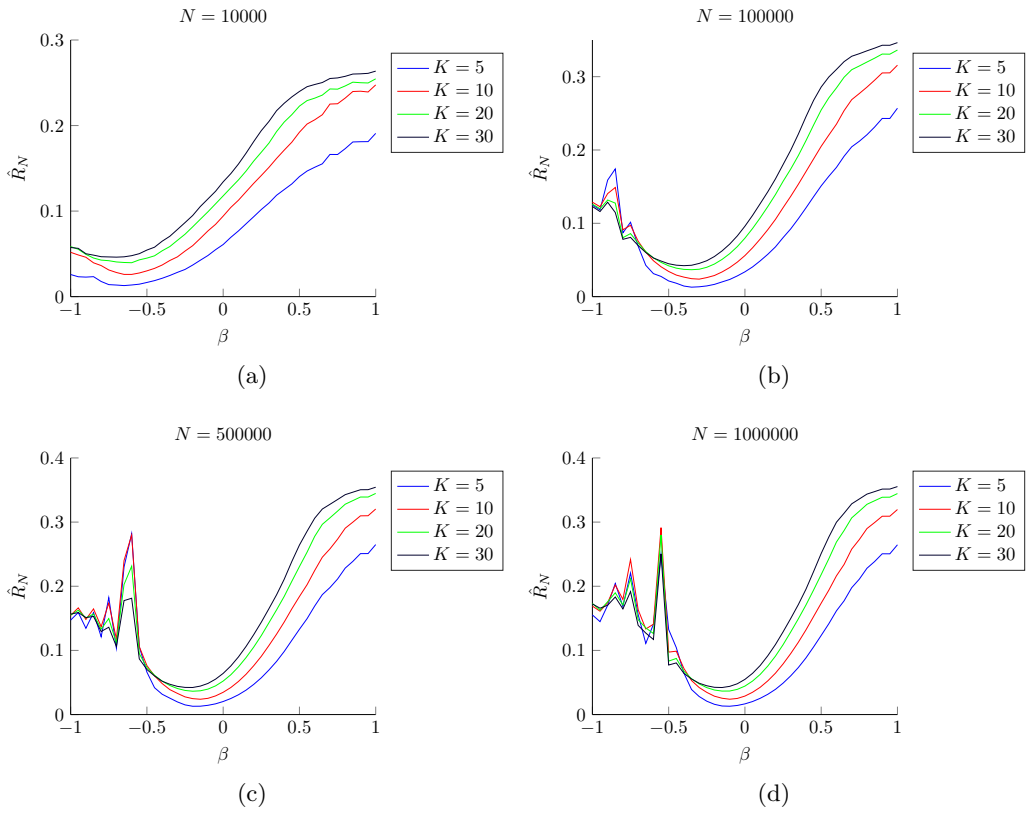


Figure 11: SC-MAB: per-round pseudo-regret $\hat{R}_N$ as the value of $\beta \in \mathbb{B}$ varies for every number of arms K and every different length of the time horizons N.

References

Agrawal, S., & Goyal, N. (2012). Analysis of Thompson Sampling for the multi-armed bandit problem. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 39.1–39.26.

Allesiardo, R., Féraud, R., & Maillard, O.-A. (2017). The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, *3*, 1–17.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2), 235–256.

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

Besbes, O., Gur, Y., & Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, pp. 199–207.

Besson, L., & Kaufmann, E. (2019). The generalized likelihood ratio test meets KLUCB: an improved algorithm for piece-wise non-stationary bandits. *arXiv preprint arXiv:1902.01575*.

Bisi, L., Nittis, G. D., Trovò, F., Restelli, M., & Gatti, N. (2017). Regret minimization algorithms for the followers behaviour identification in leadership games. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Chapelle, O., & Li, L. (2011). An empirical evaluation of Thompson Sampling. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, pp. 2249–2257.

Combes, R., & Proutiere, A. (2014). Unimodal bandits: regret lower bounds and optimal algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 521–529.

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

Farina, G., & Gatti, N. (2017). Adopting the cascade model in ad auctions: Efficiency bounds and truthful algorithmic mechanisms. *Journal of Artificial Intelligence Research*, *59*, 265–310.

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

Garivier, A., & Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, pp. 174–188.

Gasparini, M., Nuara, A., Trovò, F., Gatti, N., & Restelli, M. (2018). Targeting optimization for internet advertising by learning from logged bandit feedback. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Gatti, N., Lazaric, A., Rocco, M., & Trovò, F. (2015). Truthful learning mechanisms for multi-slot sponsored search auctions with externalities. *Artificial Intelligence*, *227*, 93–139.

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

Granmo, O.-C. (2010). Solving two-armed Bernoulli bandit problems using a Bayesian learning automaton. *International Journal of Intelligent Computing and Cybernetics*, *3*(2), 207–234.

Granmo, O.-C., & Berg, S. (2010). Solving non-stationary bandit problems by random sampling from sibling Kalman filters. In *Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE)*, pp. 199–208.

Hartland, C., Gelly, S., Baskiotis, N., Teytaud, O., & Sebag, M. (2006). Multi-armed bandit, dynamic environments and meta-bandits. Working paper.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*(301), 13–30.

Kaufmann, E., Cappé, O., & Garivier, A. (2012a). On Bayesian upper confidence bounds for bandit problems. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 592–600.

Kaufmann, E., Korda, N., & Munos, R. (2012b). Thompson Sampling: an asymptotically optimal finite-time analysis. In *Proceedings of the Algorithmic Learning Theory (ALT)*, pp. 199–213.

Kitts, B., & Leblanc, B. (2004). Optimal bidding on keyword auctions. *Electronic Markets*, *14*(3), 186–201.

Kocsis, L., & Szepesvári, C. (2006). Discounted UCB. In *PASCAL Challenges Workshop*, pp. 784–791.

Lai, L., El Gamal, H., Jiang, H., & Poor, H. V. (2011). Cognitive medium access: Exploration, exploitation, and competition. *IEEE Transactions on Mobile Computing*, *10*(2), 239–253.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*(1), 4–22.

Liu, F., Lee, J., & Shroff, N. (2018). A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Luo, H., Wei, C.-Y., Agarwal, A., & Langford, J. (2018). Efficient contextual bandits in non-stationary worlds. In *Proceedings of the Conference On Learning Theory (COLT)*, pp. 1739–1776.

Marchesi, A., Trovó, F., & Gatti, N. (2020). Learning probably approximately correct maximin strategies in simulation-based games with infinite strategy spaces. In *Proceedings*

of the International Conference on Autonomous Agents and MultiAgent Systems (AAMAS).

May, B. C., Korda, N., Lee, A., & Leslie, D. S. (2012). Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, *13*(Jun), 2069–2106.

Mellor, J., & Shapiro, J. (2013). Thompson Sampling in switching environments with Bayesian online change point detection. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 442–450.

Nuara, A., Sosio, N., Trovò, F., Zaccardi, M. C., Gatti, N., & Restelli, M. (2019). Dealing with interdependencies and uncertainty in multi-channel advertising campaigns optimization. In *Proceedings of the the World Wide Web Conference (WWW)*, pp. 1376–1386.

Nuara, A., Trovó, F., Crippa, D., Gatti, N., & Restelli, M. (2020). Driving exploration by maximum distribution in gaussian process bandits. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*.

Nuara, A., Trovò, F., Gatti, N., & Restelli, M. (2018). A combinatorial-bandit algorithm for the online joint bid/budget optimization of pay-per-click advertising campaigns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2379–2386.

Paladino, S., Trovò, F., Restelli, M., & Gatti, N. (2017). Unimodal Thompson Sampling for graph-structured arms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2457–2463.

Slivkins, A. (2011). Contextual bandits with similarity information. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 679–702.

Slivkins, A., & Upfal, E. (2008). Adapting to a changing environment: the Brownian restless bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 343–354.

St-Pierre, D. L., & Jialin, L. (2014). Differential evolution algorithm applied to non-stationary bandit problem. In *Proceeding of the IEEE Congress on Evolutionary Computation (CEC)*, pp. 2397–2403.

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

Trovò, F., Paladino, S., Restelli, M., & Gatti, N. (2018). Improving multi-armed bandit algorithms in online pricing settings. *International Journal of Approximate Reasoning*, *98*, 196–235.

Wei, C.-Y., Hong, Y.-T., & Lu, C.-J. (2016). Tracking the best expert in non-stationary stochastic environments. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, pp. 3972–3980.

Appendix A. Preliminaries

In this section, we introduce some lemmas that we use in the proofs of our main theorems. At first, we recall the link shown by Agrawal and Goyal (2012), and cited by Kaufmann et al. (2012b), between Beta and Bernoulli distributions, usually addressed in the literature as the "Beta-Binomial trick".

Lemma 1 (Agrawal & Goyal, 2012)**.** *Let us denote with $F_{a,b}^{Beta}$ the Cumulative Distribution Function (CDF) of a Beta distribution $Beta(a,b)$ with parameters a and b and with $F_{n,\mu}^{B}$ the CDF of a random variable with Binomial distribution $Bi(n,\mu)$ with parameters n and μ. It is true that:*

$$F_{a,b}^{Beta}(y) = 1 - F_{a+b-1,y}^{B}(a-1).$$

We introduce the following lemma that we use below to bound the number of times a Thompson sample $\vartheta_{i,t}$ is drawn from a high quantile of the Beta distribution.

Lemma 2. *Consider a random variable Beta with Beta distribution $Beta(S+1, T-S+1)$, where $S := \sum_{s=1}^{T} X_s$ is the sum of $T \in \mathbb{N}$ Bernoulli trials $X_s \sim Be(\mu)$ with same parameter $\mu \in [0,1]$. Consider a finite integer $\tau \in \mathbb{N}, \tau > T$, a parameter $\varepsilon > \frac{1}{2}$ and:*

$$u_T := \frac{S}{T} + \sqrt{\frac{\varepsilon \log \tau}{T}},$$
$$q_T := Q\left(1 - \frac{1}{\tau}\right),$$

where $Q(\alpha)$ is the α-quantile of the random variable Beta. We have that $q_T \leq u_T$.

Proof. Here, we use the inequalities provided by Kaufmann, Cappé, and Garivier (2012a) to bound the $(1-\frac{1}{\tau})$-quantile of a Beta distribution with the KL-divergence and elaborate over them. Consider the event that variable $B \sim \text{Beta}(S+1, T-S+1)$ is greater than the UCB-like bound u_T. We have:

$$\begin{aligned}
\mathbb{P}(Beta \geq u_T) &= 1 - F^{\text{Beta}}_{S+1,T-S+1}(u_T) && (3)\\
&= F^{\text{B}}_{T+1,u_T}(S) && (4)\\
&= 1 - F^{\text{B}}_{T+1,1-u_T}(T-S+1) = \mathbb{P}(Bi_{T+1,1-u_T} > T-S+1) && (5)\\
&\leq \exp\left\{-T \cdot KL\left(\frac{T-S+1}{T+1}, 1-u_T\right)\right\} && (6)\\
&\leq \exp\left\{-2T\left(\frac{T-S+1}{T+1} - 1 + u_T\right)^2\right\} && (7)\\
&= \exp\left\{-2T\left(\frac{T-S+1}{T+1} - 1 + \frac{S}{T} + \sqrt{\frac{\varepsilon \log \tau}{T}}\right)^2\right\} && (8)\\
&\leq \exp\left\{-2T\frac{\varepsilon \log \tau}{T}\right\} = \frac{1}{\tau^{2\varepsilon}}, && (9)
\end{aligned}$$

where $Bi_{n,\mu}$ is a random variable with Binomial distribution $\text{Bi}(n,\mu)$ with parameters n and μ, $KL(\cdot,\cdot)$ is the Kullback-Leibler divergence, the equalities in Equation (3) follow

from Lemma 1, Equation (5) follows from the properties of the Binomial distribution, Equation (6) follows from the Sanov inequality, Equation (7) follows from the Pinsker inequality. The quantile $Q\left(1-\frac{1}{\tau}\right)$ satisfies, by definition, the property:

$$\mathbb{P}\left(Beta \geq q_T\right) = \frac{1}{\tau}.$$

Since we have $\frac{1}{\tau} \geq \frac{1}{\tau^{2\varepsilon}}$ for $\varepsilon \geq \frac{1}{2}$, it follows that $q_T \leq u_T$. □

Finally, we introduce the following lemma, whose proof is provided independently in the appendices of the works by Garivier and Moulines (2011) (Lemma 1) and Combes and Proutiere (2014) (Lemma 4.1).

Lemma 3 (Garivier & Moulines, 2011; Combes & Proutiere, 2014)**.** *Let $A \subset \mathbb{N}$ and $a(t) = \sum_{t'=t-\tau}^{t-1} \mathbb{1}\{t' \in A\}$, then for every positive integer τ and every $s \in \mathbb{N}$ we have:*

$$\sum_{t=1}^{N} \mathbb{1}\{t \in A, a(t) \leq s\} \leq s\left\lceil \frac{N}{\tau} \right\rceil.$$

Appendix B. Abruptly Changing Setting: Proofs

Theorem 1. *If the SW-TS policy is run over an AC-MAB setting with $X_{i,t} \sim Be(\mu_{i,t})$, for every $\tau \in \mathbb{N}$, the dynamic pseudo-regret after N rounds is at most:*

$$\overline{R}_N(\mathfrak{U}) \leq \sum_{i=1}^{K}\left[\tau B N^{\alpha} + \sum_{\phi=1}^{B_N} \Delta_{i,\phi}\frac{N_\phi}{\tau}\left(\frac{52\log\tau}{\Delta_{i,\phi}^2} + \log\tau + 5 + \frac{19}{\log\tau}\right)\right],$$

where B and α are defined in Assumption 1 and $\Delta_{i,\phi} := \mu_{i^*,\phi} - \mu_{i,\phi}$ is the difference between the expected reward $\mu_{i^*,\phi}$ of the best arm $a_{i_\phi^*}$ and the expected reward $\mu_{i,\phi}$ of arm a_i during phase $\mathcal{F}_\phi$. By defining:

$$\Delta_i := \min_{\phi \in \{1,\dots,B_N\}} \Delta_{i,\phi}\mathbb{1}\{i \neq i_\phi^*\},$$

for all $i \in \{1,\dots,K\}$, i.e.*, the minimum over all the phases $\mathcal{F}_\phi$ of the difference of the expected rewards $\Delta_{i,\phi}$, the dynamic pseudo-regret can be written as:*

$$\overline{R}_N(\mathfrak{U}) \leq \tau K B N^{\alpha} + \frac{N}{\tau}\sum_{i=1}^{K}\left(\frac{52\log\tau}{\Delta_i} + \log\tau + 5 + \frac{19}{\log\tau}\right).$$

Proof. We adapt, to the AC-MAB setting, the proof provided by Kaufmann et al. (2012b) to bound the expected pseudo-regret of the classical Thompson Sampling algorithm. In particular, in the following, we remark the points where our proof distinguishes from the one by Kaufmann et al. (2012b).

Let us define the effective phase $\mathcal{F}'_\phi := \{t \in \mathbb{N} \text{ s.t. } b_{\phi-1} + \tau \leq t < b_\phi\}$ and denote with $T_i(\mathcal{F}'_\phi) := \sum_{t \in \mathcal{F}'_\phi} \mathbb{1}\{i_t = i, i \neq i_\phi^*\}$, *i.e.*, the number of times a suboptimal arm $a_i \neq a_{i_\phi^*}$ is played during the effective phase $\mathcal{F}'_\phi$. During every generic effective phase $\mathcal{F}'_\phi$, the MAB setting is stationary. Moreover, by the definition of effective phase $\mathcal{F}'_\phi$, we have:

$$\mathbb{E}\left[T_i(\mathcal{F}_\phi)\right] \leq \tau + \mathbb{E}\left[T_i(\mathcal{F}'_\phi)\right], \tag{10}$$

where we recall that $T_i(\mathcal{F}_\phi)$ is the number of times arm a_i is pulled during phase $\mathcal{F}_\phi$.

At first, we bound the expected number of times the algorithm chooses a suboptimal arm in a generic effective phase $\mathcal{F}'_\phi$. The rationale with which we bound $\mathbb{E}[T_i(\mathcal{F}'_\phi)]$ is to decompose $\mathbb{E}[T_i(\mathcal{F}'_\phi)]$ into two terms. The first term is conditioned to the fact that the reward of the optimal arm $a_{i^*_\phi}$ is underestimated, while the second term is conditioned to the fact that the reward of the optimal arm $a_{i^*_\phi}$ is not underestimated, but the suboptimal arm a_i is played. Hence, we have:

$$\mathbb{E}\left[T_i(\mathcal{F}'_\phi)\right] = \sum_{t\in\mathcal{F}'_\phi} \mathbb{E}\left[\mathbb{1}\{i_t = i\}\right] \tag{11}$$

$$= \sum_{t\in\mathcal{F}'_\phi} \left[\mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}, i_t = i\right) + \mathbb{P}\left(\vartheta_{i^*_\phi,t} > \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}, i_t = i\right)\right] \tag{12}$$

$$\leq \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}\right) + \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i,t} > \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}, i_t = i\right) \tag{13}$$

$$\leq \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}\right) +$$

$$+ \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i,t} > \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}, i_t = i, \vartheta_{i,t} < q_{T_{i,t,\tau}}\right) + \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i,t} \geq q_{T_{i,t,\tau}}\right) \tag{14}$$

$$\leq \underbrace{\sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}\right)}_{R_A} + \underbrace{\sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(u_{T_{i,t,\tau}} > \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}, i_t = i\right)}_{R_B} + \tag{15}$$

$$+ \underbrace{\sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i,t} \geq q_{T_{i,t,\tau}}\right)}_{R_C}, \tag{16}$$

where, in bounding Equation (12), we use the property that the Thompson sample $\vartheta_{i_t,t} = \vartheta_{i,t}$ chosen at round t is larger than the one of the optimal arm $\vartheta_{i^*_\phi,t}$ (*i.e.*, $\vartheta_{i,t} \geq \vartheta_{i^*_\phi,t}$), $q_{T_{i,t,\tau}}$ is the quantile of order $1 - \frac{1}{\tau}$ of the Beta distribution corresponding the expected value $\mu_{i,t}$ of arm a_i, and we use Lemma 2, applied to the rewards of arm a_i and with $T = T_{i,t,\tau}$ and $\varepsilon = 2$, to bound the second term in Equation (14).

Let us focus on R_A. In the analysis by Kaufmann et al. (2012b), the probability that the optimal arm is pulled in the past less than t^b times (by properly defining the constant $b \in (0,1)$) is bounded by a constant (from Proposition 1 by Kaufmann et al. (2012b)). In the setting we study, such a result does not hold as the number of samples used in the posterior distribution $\pi_{i,t}$ of the expected reward $\mu_{i,t}$ does not increase indefinitely over time due to the sliding-window approach that limits the number of samples to at most τ. Thus, we bound the probability that an arm is pulled less than $\bar{n}_A$ times by using Lemma 3

with $A = \{t | i_t = i\}$, $t \in \mathcal{F}'_\phi$ and, consequently $a(t) = T_{i,t,\tau}$. We have:

$$\sum_{t \in \mathcal{F}'_\phi} \mathbb{E}\left[\mathbb{1}\{i_t = i, T_{i,t,\tau} \leq \bar{n}_A\}\right] \leq \bar{n}_A \left\lceil \frac{N_\phi - \tau}{\tau} \right\rceil \leq \bar{n}_A \frac{N_\phi}{\tau}, \tag{17}$$

where $|\mathcal{F}'_\phi| = N_\phi - \tau \leq N_\phi$, which holds for all $i \in \{1, \dots, K\}$. Thus, by choosing $\bar{n}_A = \left\lceil \frac{19}{\log \tau} \right\rceil$, we have:

$$\begin{aligned}
R_A &= \sum_{t \in \mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5 \log \tau}{T_{i^*_\phi,t,\tau}}} \right) && (18)\\
&\leq \sum_{t \in \mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5 \log \tau}{T_{i^*_\phi,t,\tau}}}, T_{i^*_\phi,t,\tau} > \bar{n}_A \right) + \sum_{t \in \mathcal{F}'_\phi} \mathbb{P}\left(T_{i^*_\phi,t,\tau} \leq \bar{n}_A \right) && (19)\\
&\leq \sum_{t \in \mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5 \log \tau}{T_{i^*_\phi,t,\tau}}}, T_{i^*_\phi,t,\tau} > \bar{n}_A \right) + \sum_{t \in \mathcal{F}'_\phi} \mathbb{E}\left[\mathbb{1}\left\{ T_{i^*_\phi,t,\tau} \leq \bar{n}_A \right\} \right] && (20)\\
&\leq \sum_{t \in \mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5 \log \tau}{T_{i^*_\phi,t,\tau}}}, T_{i^*_\phi,t,\tau} > \bar{n}_A \right) + \bar{n}_A \frac{N_\phi}{\tau}, && (21)
\end{aligned}$$

where we use Lemma 3 to bound Equation (21).

Let us define:

- $\{U_t\}_{t \in \mathcal{F}'_\phi}$ a set of i.i.d. uniform random variables over $\Omega = [0, 1]$;
- $S_{i,t,\tau} := \sum_{h=t-\tau+1}^{t} X_{i,h} \mathbb{1}\{i_h = i\}$ the sum of the rewards received by arm a_i in the last τ rounds (with abuse of notation w.r.t. the main body of the paper);
- $\Sigma_{i,t,\tau,s} := \sum_{s=t-\tau+1}^{t-\tau+s} X_{i,h} \mathbb{1}\{i_h = i\}$ the sum of the first s rewards over the last τ rounds of arm a_i.

Recalling that $T_{i,t,\tau} := \sum_{h=\max\{t-\tau+1,1\}}^{t} \mathbb{1}\{i_h = i\}$, we have:

$$\begin{aligned}
&\mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5 \log \tau}{T_{i^*_\phi,t,\tau}}}, T_{i^*_\phi,t,\tau} > \bar{n}_A \right) && (22)\\
&= \mathbb{P}\left(U_t \leq F^{\text{Beta}}_{S_{i^*_\phi,t,\tau}+1, T_{i^*_\phi,t,\tau} - S_{i^*_\phi,t,\tau}+1}\left(\mu_{i^*_\phi,t} - \sqrt{\frac{5 \log \tau}{T_{i^*_\phi,t,\tau}}} \right), T_{i^*_\phi,t,\tau} > \bar{n}_A \right) && (23)\\
&= \mathbb{P}\left(U_t \leq 1 - F^{\text{B}}_{T_{i^*_\phi,t,\tau}+1, \mu_{i^*_\phi,t} - \sqrt{\frac{5 \log \tau}{T_{i^*_\phi,t,\tau}}}}(S_{i^*_\phi,t,\tau}), T_{i^*_\phi,t,\tau} > \bar{n}_A \right) && (24)\\
&= \mathbb{P}\left(F^{\text{B}}_{T_{i^*_\phi,t,\tau}+1, \mu_{i^*_\phi,t} - \sqrt{\frac{5 \log \tau}{T_{i^*_\phi,t,\tau}}}}(S_{i^*_\phi,t,\tau}) \leq U_t, T_{i^*_\phi,t,\tau} > \bar{n}_A \right) && (25)
\end{aligned}$$

$$\leq \mathbb{P}\left(\exists s \in \{\bar{n}_A, \ldots, \tau\} F^{\mathrm{B}}_{s+1,\mu_{i^*_\phi,t}-\sqrt{\frac{5\log\tau}{s}}}(\Sigma_{i^*_\phi,t,\tau,s}) \leq U_t\right) \tag{26}$$

$$\leq \sum_{s=\bar{n}_A}^{\tau} \mathbb{P}\left(\Sigma_{i^*_\phi,t,\tau,s} \leq (F^{\mathrm{B}})^{-1}_{s+1,\mu_{i^*_\phi,t}-\sqrt{\frac{5\log\tau}{s}}}(U_t)\right), \tag{27}$$

where we use Lemma 1 to derive Equation (24), we use the fact that $U_t \sim 1 - U_t$ to derive Equation (25), and we use the union bound to derive Equation (27).

Note that:

$$(F^{\mathrm{B}})^{-1}_{s+1,\mu_{i^*_\phi,t}-\sqrt{\frac{5\log\tau}{s}}}(U_t) \sim \mathrm{Bi}\left(s+1, \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{s}}\right). \tag{28}$$

We also remark that term $(F^{\mathrm{B}})^{-1}_{s+1,\mu_{i^*_\phi,t}-\sqrt{\frac{5\log\tau}{s}}}(U_t)$ is independent of $\Sigma_{i^*_\phi,t,\tau,s} \sim \mathrm{Bi}(s, \mu_{i^*_\phi,t})$. Similarly to what done by Kaufmann et al. (2012b), we define, for a chosen s, two i.i.d. sequences of Bernoulli random variables $\{X_{1,l}\}_{l=1}^{s}$ and $\{X_{2,l}\}_{l=1}^{s}$ of size s and $s+1$, respectively:

$$X_{1,l} \sim \mathrm{Be}\left(\mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{s}}\right), \tag{29}$$

$$X_{2,l} \sim \mathrm{Be}\left(\mu_{i^*_\phi,t}\right), \tag{30}$$

whose summations correspond to the r.h.s. and l.h.s. of the inequality present in the probability in Equation (27), respectively. Let $\{Z_l\}_{l=1}^{s}$ be another i.i.d. sequence of random variables, with $Z_l := X_{2,l} - X_{1,l}$ and $\mathbb{E}[Z_l] = \sqrt{\frac{5\log\tau}{s}}$.[11] We get:

$$\mathbb{P}\left(\Sigma_{i^*_\phi,t,\tau,s} \leq (F^{\mathrm{B}})^{-1}_{s+1,\mu_{i^*_\phi,t}-\sqrt{\frac{5\log\tau}{s}}}(U_t)\right) \tag{31}$$

$$= \mathbb{P}\left(\sum_{l=1}^{s} X_{2,l} \leq \sum_{l=1}^{s+1} X_{1,l}\right) = \mathbb{P}\left(\sum_{l=1}^{s} Z_l \leq X_{1,s+1}\right) \leq \mathbb{P}\left(\sum_{l=1}^{s} Z_l \leq 1\right) \tag{32}$$

$$= \mathbb{P}\left(\sum_{l=1}^{s}\left(Z_l - \sqrt{\frac{5\log\tau}{s}}\right) \leq -\sum_{l=1}^{s}\sqrt{\frac{5\log\tau}{s}} + 1\right) \tag{33}$$

$$= \mathbb{P}\left(\sum_{l=1}^{s}\left(Z_l - \sqrt{\frac{5\log\tau}{s}}\right) \leq -\left(\sqrt{5s\log\tau} - 1\right)\right) \tag{34}$$

$$\leq \mathbb{P}\left(\sum_{l=1}^{s}\left(Z_l - \sqrt{\frac{5\log\tau}{s}}\right) \leq -\sqrt{4s\log\tau}\right), \tag{35}$$

11. We here assume that $\mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{s}} \geq 0$, *i.e.*, that the sequence $\{X_{1,l}\}_{l=1}^{s}$ is well defined. In the case this condition does not hold, we have $R_A = 0$ since the event that the Thompson sample $\vartheta_{i^*_t,t} < 0$ has zero probability.

where we use that $s > \bar{n}_A \Rightarrow \sqrt{5s\log\tau} - 1 > \sqrt{4s\log\tau}$ to bound Equation (35). We apply the Hoeffding's inequality provided by Hoeffding (1963) to the bounded martingale difference sequence $\{Z_l\}_{l=1}^{s}$ (having support of measure 2) and we get:

$$\sum_{s=\bar{n}_A}^{\tau} \mathbb{P}\left(\Sigma_{i^*_\phi,t,\tau,s} \leq (F^{\mathrm{B}})^{-1}_{s+1,\mu_{i^*_\phi,t}-\sqrt{\frac{5\log\tau}{s}}}(U_t)\right) \leq \sum_{s=\bar{n}_A}^{\tau} \exp\left(-\frac{2s(\sqrt{4s\log\tau})^2}{\sum_{h=1}^{s} 2^2}\right) \tag{36}$$

$$= \sum_{s=\bar{n}_A}^{\tau} \exp\left(-\frac{(\sqrt{4s\log\tau})^2}{2s}\right) = \sum_{s=\bar{n}_A}^{\tau} e^{-2\log\tau} \leq \sum_{s=1}^{\tau} \frac{1}{\tau^2} = \frac{1}{\tau}. \tag{37}$$

Finally, we get:

$$R_A = \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\vartheta_{i^*_\phi,t} \leq \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}\right) \leq \bar{n}_A \frac{N_\phi}{\tau} + \sum_{t\in\mathcal{F}'_\phi} \frac{1}{\tau^{\frac{3}{2}}} \tag{38}$$

$$\leq \frac{N_\phi}{\tau}\left(\frac{19}{\log\tau} + 1\right) + \frac{N_\phi}{\tau} = \frac{19N_\phi}{\tau\log\tau} + \frac{2N_\phi}{\tau}. \tag{39}$$

Let us focus on R_B. Differently from what done by Kaufmann et al. (2012b), we use the Hoeffding's inequality to bound each probability term used in R_B. Let us define $\hat{\mu}_{i,t,\tau} := \frac{\sum_{s=t-\tau+1}^{t} X_{i,s}\mathbb{1}\{i_s=i\}}{T_{i,t,\tau}}$, *i.e.*, the estimator of the expected value $\mu_{i,\phi}$ of the rewards of arm a_i computed over the last τ rounds and choose $\bar{n}_{B*} = \left\lceil \frac{20\log\tau}{\Delta^2_{i,\phi}} \right\rceil$ and $\bar{n}_B = \left\lceil \frac{32\log\tau}{\Delta^2_{i,\phi}} \right\rceil$, where we recall that $\Delta_{i,\phi} := \mu_{i^*_\phi,t} - \mu_{i,t}$ with $t \in \mathcal{F}'_\phi$. If the two properties $T_{i^*_\phi,t,\tau} > \bar{n}_{B*}$ and $T_{i,t,\tau} > \bar{n}_B$ hold, we have the following:

$$-\left(2\sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} + \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}\right) > -\Delta_{i,\phi} \tag{40}$$

and thus:

$$R_B \leq \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(u_{T_{i,t,\tau}} > \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}, i_t = i\right) \tag{41}$$

$$= \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} + \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}, i_t = i\right) \tag{42}$$

$$\leq \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} + \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}, T_{i^*_\phi,t,\tau} > \bar{n}_{B*}, T_{i,t,\tau} > \bar{n}_B\right)$$

$$+ \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(T_{i^*_\phi,t,\tau} \leq \bar{n}_{B*}\right) + \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(T_{i,t,\tau} \leq \bar{n}_B\right) \tag{43}$$

$$\leq \sum_{t\in\mathcal{F}'_\phi} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} + \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i^*_\phi,t} - \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}, T_{i^*_\phi,t,\tau} > \bar{n}_{B*}, T_{i,t,\tau} > \bar{n}_B\right) \tag{44}$$

$$+ \bar{n}_{B*}\frac{N_\phi}{\tau} + \bar{n}_B\frac{N_\phi}{\tau} \tag{45}$$

$$\leq \sum_{t\in\mathcal{F}'_\phi}\mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t} + \underbrace{\mu_{i^*_\phi,t} - \mu_{i,t}}_{=\Delta_{i,\phi}} \underbrace{- \left(2\sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} + \sqrt{\frac{5\log\tau}{T_{i^*_\phi,t,\tau}}}\right)}_{>-\Delta_{i,\phi}}\right) \tag{46}$$

$$+ \frac{N_\phi}{\tau}\left(\frac{52\log\tau}{\Delta^2_{i,\phi}} + 2\right) \tag{47}$$

$$\leq \sum_{t\in\mathcal{F}'_\phi}\mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t}\right) + \frac{N_\phi}{\tau}\frac{52\log\tau}{\Delta^2_{i,\phi}} + \frac{2N_\phi}{\tau}, \tag{48}$$

where Equation (42) is derived from the definition of $u_{T_{i,t,\tau}}$.

From Corollary 21 by Garivier and Moulines (2008), we have for all $\eta > 0$:

$$\sum_{t\in\mathcal{F}'_\phi}\mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t}\right) \leq \sum_{t\in\mathcal{F}'_\phi}\frac{\log\tau}{\log(1+\eta)}\exp\left(-12\log\tau\left(1 - \frac{\eta^2}{16}\right)\right) \tag{49}$$

and, since $\eta = 4\sqrt{1-\frac{1}{12}}$, the bound can be written as:

$$\sum_{t\in\mathcal{F}'_\phi}\mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t}\right) \leq \sum_{t\in\mathcal{F}'_\phi}\frac{\log\tau}{\tau} \leq \frac{N_\phi\log\tau}{\tau}. \tag{50}$$

Hence, we can write:

$$R_B \leq \frac{N_\phi}{\tau}\frac{52\log\tau}{\Delta^2_{i,\phi}} + \frac{2N_\phi}{\tau} + \frac{N_\phi\log\tau}{\tau}. \tag{51}$$

Let us focus on R_C. The R_C term is upper bounded by:

$$R_C = \sum_{t\in\mathcal{F}'_\phi}\mathbb{P}\left(\vartheta_{i,t} \geq q_{T_{i,t,\tau}}\right) = \sum_{t\in\mathcal{F}'_\phi}\frac{1}{\tau} \leq \frac{N_\phi}{\tau}, \tag{52}$$

using the definition of quantiles $q_{T_{i,t,\tau}}$.

Pseudo-regret. Since $\sum_{\phi=1}^{B_N} N_\phi = N$ and recalling that if $t \in \mathcal{F}_\phi \supset \mathcal{F}'_\phi$ we have $\mu_{i,t} = \mu_{i,\phi}$, the dynamic pseudo-regret over all the phases can be written as:

$$\overline{R}_N(\mathfrak{U}) = \mathbb{E}\left[\sum_{t=1}^{N}\left(\mu_{i^*,t} - \mu_{i_t,t}\right)\right] = \sum_{\phi=1}^{B_N}\mu_{i^*,\phi}N_\phi - \mathbb{E}\left[\sum_{t=1}^{N}\mu_{i_t,t}\right] \tag{53}$$

$$\overline{R}_N(\mathfrak{U}) = \mathbb{E}\left[\sum_{t=1}^{N}\left(\mu_{i^*,t} - \mu_{i_t,t}\right)\right] = \sum_{\phi=1}^{B_N}\mu_{i^*,\phi}N_\phi - \mathbb{E}\left[\sum_{t=1}^{N}\mu_{i_t,t}\right] \tag{54}$$

$$= \sum_{\phi=1}^{B_N} \left(\mu_{i^*,\phi} N_\phi - \mathbb{E}\left[\sum_{t \in \mathcal{F}_\phi} \mu_{i_t,t} \right] \right) = \sum_{\phi=1}^{B_N} \left(\mu_{i^*,\phi} N_\phi - \sum_{i=1}^{K} \mu_{i,\phi} \mathbb{E}[T_i(\mathcal{F}_\phi)] \right) \tag{55}$$

$$= \sum_{\phi=1}^{B_N} \left(\sum_{i=1}^{K} (\mu_{i^*,\phi} - \mu_{i,\phi}) \mathbb{E}[T_i(\mathcal{F}_\phi)] \right) = \sum_{i=1}^{K} \left(\sum_{\phi=1}^{B_N} (\mu_{i^*,\phi} - \mu_{i,\phi}) \mathbb{E}[T_i(\mathcal{F}_\phi)] \right) \tag{56}$$

$$= \sum_{i=1}^{K} \left(\sum_{\phi=1}^{B_N} \Delta_{i,\phi} \mathbb{E}[T_i(\mathcal{F}_\phi)] \right) \leq \sum_{i=1}^{K} \left[\sum_{\phi=1}^{B_N} \Delta_{i,\phi} \left(\tau + \mathbb{E}[T_i(\mathcal{F}'_\phi)] \right) \right] \tag{57}$$

$$\leq \sum_{i=1}^{K} \left[\tau B_N + \sum_{\phi=1}^{B_N} \Delta_{i,\phi} \left(R_A + R_B + R_C \right) \right] \tag{58}$$

$$\leq \sum_{i=1}^{K} \left[\tau B N^\alpha + \sum_{\phi=1}^{B_N} \Delta_{i,\phi} \left(\frac{19 N_\phi}{\tau \log \tau} + \frac{2 N_\phi}{\tau} + \frac{N_\phi}{\tau} \frac{52 \log \tau}{\Delta_{i,\phi}^2} + \frac{2 N_\phi}{\tau} + \frac{N_\phi \log \tau}{\tau} + \frac{N_\phi}{\tau} \right) \right] \tag{59}$$

$$\leq \sum_{i=1}^{K} \left[\tau B N^\alpha + \sum_{\phi=1}^{B_N} \Delta_{i,\phi} \left(\frac{19 N_\phi}{\tau \log \tau} + \frac{N_\phi}{\tau} \frac{52 \log \tau}{\Delta_{i,\phi}^2} + \frac{N_\phi \log \tau}{\tau} + \frac{5 N_\phi}{\tau} \right) \right] \tag{60}$$

$$\leq \sum_{i=1}^{K} \left[\tau B N^\alpha + \sum_{\phi=1}^{B_N} \Delta_{i,\phi} \frac{N_\phi}{\tau} \left(\frac{52 \log \tau}{\Delta_{i,\phi}^2} + \log \tau + 5 + \frac{19}{\log \tau} + \frac{1}{\tau^{\frac{1}{2}}} \right) \right], \tag{61}$$

where B_N is the number of breakpoints before N.

By defining:

$$\Delta_i := \min_{\phi \in \{1,\dots,B_N\}} \Delta_{i,\phi} \mathbb{1}\{i \neq i^*_\phi\} \quad \forall i \in \{1,\dots,K\}, \tag{62}$$

i.e., the minimum, over all the phases $\mathcal{F}_\phi$ in which the arm a_i is not optimal, of the difference between the expected reward $\mu_{i^*_\phi,\phi}$ of the best arm $a_{i^*_\phi}$ and the expected reward $\mu_{i,\phi}$ of arm a_i, the dynamic pseudo-regret can be written as:

$$\overline{R}_N(\mathfrak{U}) \leq \tau K B N^\alpha + \frac{N}{\tau} \sum_{i=1}^{K} \left(\frac{52 \log \tau}{\Delta_i} + \log \tau + 5 + \frac{19}{\log \tau} \right), \tag{63}$$

which concludes the proof. □

Appendix C. Smoothly Changing Setting: Proofs

Theorem 2. *If the SW-TS policy is run over a SC-MAB setting with $X_{i,t} \sim Be(\mu_{i,t})$, Lipschitz constant $\sigma > 0$ and there exists $\Delta_0 \in (0,1)$ as in Assumption 3, for any $\tau \in \mathbb{N}$ s.t. $2\sigma\tau < \Delta \leq \Delta_0$, the dynamic pseudo-regret after N rounds is at most:*

$$\overline{R}_N(\mathfrak{U}) \leq F \Delta N^\beta + \frac{NK}{\tau} \left[\frac{52 \log \tau}{(\Delta - 2\sigma\tau)^2} + \log \tau + 5 + \frac{19}{\log \tau} \right] + \\ + K \left[\frac{52 \log \tau}{(\Delta - 2\sigma\tau)^2} + 3 + \frac{19}{\log \tau} \right].$$

Proof. Let us define:

- $\mathcal{F}_{\Delta,N} := \{t \in \{1,\dots,N\} \text{ s.t. } \exists i \neq j, |\mu_{i,t} - \mu_{j,t}| < \Delta\}$, *i.e.*, the set of the rounds in which there are two arms whose expected values differing less than Δ;
- $\mathcal{F}_{\Delta^C,N} := \{1,\dots,N\} \setminus \mathcal{F}_{\Delta,N}$, *i.e.*, the set of the rounds in which the expected rewards of the arms are well separated ($|\mu_{i,t} - \mu_{j,t}| > \Delta, \forall i \neq j\}$);
- $T_i(\mathcal{F}_{\Delta,N}) := \sum_{t \in \mathcal{F}_{\Delta,N}} \mathbb{1}\{i_t = i, i \neq i_t^*\}$, *i.e.*, the number of rounds arm a_i is played when it is not optimal during rounds $t \in \mathcal{F}_{\Delta,N}$;
- $T_i(\mathcal{F}_{\Delta^C,N}) := \sum_{t \in \mathcal{F}_{\Delta^C,N}} \mathbb{1}\{i_t = i, i \neq i_t^*\}$, *i.e.*, the number of rounds arm a_i is played when it is not optimal during rounds $t \in \mathcal{F}_{\Delta^C,N}$.

For every Δ s.t. $2\sigma\tau \leq \Delta \leq \Delta_0$, we have:

$$\overline{R}_N(\mathfrak{U}) = \mathbb{E}\left[\sum_{t=1}^{N} \left(\mu_{i_t^*,t} - \mu_{i_t,t}\right)\right] \leq \sum_{t=1}^{N} \mathbb{E}\left[\mathbb{1}\{i_t = i, i \neq i_t^*\}\right] \tag{64}$$

$$= \sum_{i=1}^{K} \sum_{t=1}^{N} \mathbb{E}\left[\mathbb{1}\{i_t = i, i \neq i_t^*\}\right] \tag{65}$$

$$\leq \sum_{i=1}^{K} \mathbb{E}[T_i(\mathcal{F}_{\Delta,N})] + \sum_{i=1}^{K} \mathbb{E}[T_i(\mathcal{F}_{\Delta^C,N})]. \tag{66}$$

The first term in Equation (66), *i.e.*, $\mathbb{E}[T_i(\mathcal{F}_{\Delta,N})]$, can be directly bounded by using Assumption 3. Instead, bounding the second term, *i.e.*, $\mathbb{E}[T_i(\mathcal{F}_{\Delta^C,N})]$, requires a more complex procedure. In our proof, we adapt the line provided by Kaufmann et al. (2012b) to our setting. Indeed, we remark the result by Kaufmann et al. (2012b) cannot be directly applied as the reward distributions vary at every round.[12] We define two events: in the first one the optimal arm $a_{i_t^*}$ is underestimated; in the second one the optimal arm $a_{i_t^*}$ is not underestimated, but the suboptimal arm a_i is played. Hence, we have:

$$\begin{aligned}
\mathbb{E}\left[T_i(\mathcal{F}_{\Delta^C,N})\right] &\leq \sum_{t \in \mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\vartheta_{i_t^*,t} \leq \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right) \\
&+ \sum_{t \in \mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\vartheta_{i,t} > \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}, i_t = i\right) \\
&\leq \sum_{t \in \mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\vartheta_{i_t^*,t} \leq \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right) \\
&+ \sum_{t \in \mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\vartheta_{i,t} > \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}, i_t = i, \vartheta_{i,t} \leq q_{T_{i,t,\tau}}\right)
\end{aligned} \tag{67}$$

12. For the sake of concision, we will omit the derivations which are the same of those discussed in the proof of Theorem 1.

$$+ \sum_{t \in \mathcal{F}_{\Delta C,N}} \mathbb{P}\left(\vartheta_{i,t} \geq q_{T_{i,t,\tau}}\right) \tag{68}$$

$$\leq \underbrace{\sum_{t \in \mathcal{F}_{\Delta C,N}} \mathbb{P}\left(\vartheta_{i_t^*,t} \leq \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right)}_{R_A}$$

$$+ \underbrace{\sum_{t \in \mathcal{F}_{\Delta C,N}} \mathbb{P}\left(u_{T_{i,t,\tau}} > \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right)}_{R_B}$$

$$+ \underbrace{\sum_{t \in \mathcal{F}_{\Delta C,N}} \mathbb{P}\left(\vartheta_{i,t} \geq q_{T_{i,t,\tau}}\right)}_{R_C}, \tag{69}$$

where we use Lemma 2 applied to the rewards of the arm a_i with $T = T_{i,t,\tau}$ and $\varepsilon = 2$, to bound the expression in Equation (69).

Let us focus on R_A. By applying Lemma 3 and defining $\bar{n}_A = \left\lceil \frac{19}{\log\tau} \right\rceil$, we have:

$$R_A = \sum_{t \in \mathcal{F}_{\Delta C,N}} \mathbb{P}\left(\vartheta_{i_t^*,t} \leq \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right) \tag{70}$$

$$\leq \sum_{t \in \mathcal{F}_{\Delta C,N}} \mathbb{P}\left(\vartheta_{i_t^*,t} \leq \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}, T_{i_t^*,t,\tau} > \bar{n}_A\right)$$

$$+ \sum_{t \in \mathcal{F}_{\Delta C,N}} \mathbb{P}\left(T_{i_t^*,t,\tau} \leq \bar{n}_A\right) \tag{71}$$

$$\leq \sum_{t \in \mathcal{F}_{\Delta C,N}} \mathbb{P}\left(\vartheta_{i_t^*,t} \leq \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}, T_{i_t^*,t,\tau} > \bar{n}_A\right) + \bar{n}_A \left\lceil \frac{N}{\tau} \right\rceil \tag{72}$$

$$\leq \sum_{t \in \mathcal{F}_{\Delta C,N}} \mathbb{P}\left(\vartheta_{i_t^*,t} \leq \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}, T_{i_t^*,t,\tau} > \bar{n}_A\right) + \bar{n}_A \left(\frac{N}{\tau} + 1\right). \tag{73}$$

While in the proof of Theorem 1, as well as in the proof by Kaufmann et al. (2012b), the expected values of the rewards are constant over the last τ rounds, in the case we study here such property does not hold. In what follows, we define a set of auxiliary random variables $\underline{S}_{i,t,\tau}$ whose expected values are constant over the last τ rounds and smaller than the one of the optimal arm. Then, using the random variables $\underline{S}_{i,t,\tau}$, we can apply Lemma 1 to transform the Beta distribution estimating the expected reward of an arm into a Binomial random variable, which, finally, allows us to bound the quantity R_A.

Let us define:

- $\{U_t\}_{t \in \mathcal{F}_{\Delta C,N}}$ as a sequence of i.i.d. uniform random variables over $\Omega = [0,1]$;

- $S_{i,t,\tau} := \sum_{s=t-\tau+1}^{t} \mathbb{1}\{i_s = i\} X_{i,s}$, *i.e.*, the number of successes of arm a_i at round t in the previous τ rounds (with abuse of notation w.r.t. the main body of the paper);
- $\tilde{X}_{i,s} := X_{i,s} + \mu_{i,t} - \mu_{i,s} - \sigma\tau, \forall s \in \{t-\tau+1, t\}$, *i.e.*, a set of auxiliary variables having $\tilde{X}_{i,s} \leq X_{i,s}$ (since $|\mu_{i,t} - \mu_{i,s}| \leq \sigma\tau$) and $\underline{\mu}_{i,t} := \mathbb{E}[\tilde{X}_{i,s}] = \mu_{i,t} - \sigma\tau$;
- $\underline{S}_{i,t,\tau} := \sum_{s=t-\tau+1}^{t} \mathbb{1}\{i_s = i\} \tilde{X}_{i,s}$, the number of successes of an arm $\underline{a}_i$ having rewards $\tilde{X}_{i,s}$ at round s in the rounds $\{t-\tau+1, \dots, t\}$;
- $\Sigma_{i,t,\tau,s} := \sum_{h=t-\tau+1}^{t-\tau+s} \mathbb{1}\{i_h = i\} \tilde{X}_{i,h}$, *i.e.*, the sum of the random variables $\tilde{X}_{i,t-\tau+1}, \dots, \tilde{X}_{i,t-\tau+s}$.

Note that the arm $\underline{a}_{i_t^*}$, whose expected reward is $\underline{\mu}_{i_t^*,t}$, is optimal since we are focusing on rounds $t \in \mathcal{F}_{\Delta^C,N}$. Hence, we have:

$$\mathbb{P}\left(\vartheta_{i_t^*,t} \leq \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}, T_{i_t^*,t,\tau} > \bar{n}_A\right) \tag{74}$$

$$= \mathbb{P}\left(U_t \leq F^{\text{Beta}}_{S_{i_t^*,t,\tau}+1, T_{i_t^*,t,\tau}-S_{i_t^*,t,\tau}+1}\left(\mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right), T_{i_t^*,t,\tau} > \bar{n}_A\right) \tag{75}$$

$$= \mathbb{P}\left(U_t \leq 1 - F^{\text{B}}_{T_{i_t^*,t,\tau}+1, \mu_{i_t^*,t}-\sigma\tau-\sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}}(S_{i_t^*,t,\tau}), T_{i_t^*,t,\tau} > \bar{n}_A\right) \tag{76}$$

$$= \mathbb{P}\left(F^{\text{B}}_{T_{i_t^*,t,\tau}+1, \mu_{i_t^*,t}-\sigma\tau-\sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}}(S_{i_t^*,t,\tau}) \leq U_t, T_{i_t^*,t,\tau} > \bar{n}_A\right) \tag{77}$$

$$\leq \mathbb{P}\left(F^{\text{B}}_{T_{i_t^*,t,\tau}+1, \underline{\mu}_{i_t^*,t}-\sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}}(\underline{S}_{i_t^*,t,\tau}) \leq U_t, T_{i_t^*,t,\tau} > \bar{n}_A\right) \tag{78}$$

$$\leq \mathbb{P}\left(\exists s \in \{\bar{n}_A, \dots, \tau\} \text{ s.t. } F^{\text{B}}_{s+1, \underline{\mu}_{i_t^*,t}-\sqrt{\frac{5\log\tau}{s}}}(\Sigma_{i_t^*,t,\tau,s}) \leq U_t\right) \tag{79}$$

$$= \sum_{s=\bar{n}_A}^{\tau} \mathbb{P}\left(\Sigma_{i_t^*,t,\tau,s} \leq (F^{\text{B}})^{-1}_{s+1, \underline{\mu}_{i_t^*,t}-\sqrt{\frac{5\log\tau}{s}}}(U_t)\right), \tag{80}$$

where we use Lemma 1 to derive Equation (76), Equation (77) follows from $U_t \sim 1 - U_t$, and we bound Equation (78) exploiting that $S_{i,t,\tau} \geq \underline{S}_{i,t,\tau}, \forall i$, which follows from the definition of $\underline{S}_{i,t,\tau}$.

Note that:

$$(F^{\text{B}})^{-1}_{s+1, \underline{\mu}_{i_t^*,t}-\sqrt{\frac{5\log\tau}{s}}}(U_t) \sim \text{Bi}\left(s+1, \underline{\mu}_{i_t^*,t} - \sqrt{\frac{5\log\tau}{s}}\right) \tag{81}$$

and $(F^{\text{B}})^{-1}_{s+1, \underline{\mu}_{i_t^*,t}-\sqrt{\frac{5\log\tau}{s}}}(U_t)$ is independent of $\Sigma_{i_t^*,t,\tau,s} \sim \text{Bi}(s, \underline{\mu}_{i_t^*,t})$. Consider, for a chosen s, two i.i.d. sequences of random variables $\{X_{1,l}\}_{l=1}^{s}$ and $\{X_{2,l}\}_{l=1}^{s}$ of size s and $s+1$,

respectively:

$$X_{1,l} \sim \text{Be}\left(\underline{\mu}_{i_t^*,t} - \sqrt{\frac{5\log\tau}{s}}\right), \tag{82}$$

$$X_{2,l} \sim \text{D}\left(\underline{\mu}_{i_t^*,t}\right), \tag{83}$$

whose summations correspond to the r.h.s. and l.h.s. of the inequality that is the argument of the probability operator in Equation (80), respectively. In Equation (82), we denote with $\text{Be}(\mu)$ a Bernoulli distribution with mean μ, and, in Equation (83), we denote with D a discrete distribution defined over $\Omega = \{1 + \mu_{i_t^*,t} - \mu_{i_t^*,s} - \sigma\tau, \mu_{i_t^*,t} - \mu_{i_t^*,s} - \sigma\tau\}$ and expected value equal to $\underline{\mu}_{i_t^*,t}$. Let $\{Z_l\}_{l=1}^s$ be another i.i.d. sequence of random variables, with $Z_l := X_{2,l} - X_{1,l}$, having support of measure 2 and $\mathbb{E}[Z_l] = \sqrt{\frac{5\log\tau}{s}}$.[13] We get:

$$\mathbb{P}\left(\Sigma_{i_t^*,t,\tau,s} \leq (F^{\text{B}})^{-1}_{s+1,\underline{\mu}_{i_t^*,t} - \sqrt{\frac{5\log\tau}{s}}}(U_t)\right) = \mathbb{P}\left(\sum_{l=1}^{s} X_{2,l} \leq \sum_{l=1}^{s+1} X_{1,l}\right) \tag{84}$$

$$= \mathbb{P}\left(\sum_{l=1}^{s} Z_l \leq X_{1,s+1}\right) \leq \mathbb{P}\left(\sum_{l=1}^{s} Z_l \leq 1\right) \tag{85}$$

$$= \mathbb{P}\left(\sum_{l=1}^{s}\left(Z_l - \sqrt{\frac{5\log\tau}{s}}\right) \leq -\sum_{l=1}^{s}\sqrt{\frac{5\log\tau}{s}} + 1\right) \tag{86}$$

$$= \mathbb{P}\left(\sum_{l=1}^{s}\left(Z_l - \sqrt{\frac{5\log\tau}{s}}\right) \leq -\left(\sqrt{5s\log\tau} - 1\right)\right) \tag{87}$$

$$\leq \mathbb{P}\left(\sum_{l=1}^{s}\left(Z_l - \sqrt{\frac{5\log\tau}{s}}\right) \leq -\sqrt{4s\log\tau}\right), \tag{88}$$

where we use the property $s > \bar{n}_A \Rightarrow \sqrt{5s\log\tau} - 1 > \sqrt{4s\log\tau}$. We apply the Hoeffding's inequality to the bounded martingale difference sequence $\{Z_l\}_{l=1}^s$ and we get:

$$\sum_{s=\bar{n}_A}^{\tau} \mathbb{P}\left(\Sigma_{i_t^*,t,\tau,s} \leq (F^{\text{B}})^{-1}_{s+1,\underline{\mu}_{i_t^*,t} - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}}(U_t)\right) \tag{89}$$

$$\leq \sum_{s=\bar{n}_A}^{\tau} \exp\left(-2\frac{(\sqrt{4s\log\tau})^2}{4s}\right) = \sum_{s=\bar{n}_A}^{\tau} e^{-2\log\tau} \leq \sum_{s=1}^{\tau}\frac{1}{\tau^2} = \frac{1}{\tau}. \tag{90}$$

Finally, we get:

$$R_A = \sum_{t\in\mathcal{F}_{\Delta C,N}} \mathbb{P}\left(\vartheta_{i_t^*,t} \leq \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right) \tag{91}$$

13. Similarly to what we do in the proof of Theorem 1, here, we focus only on the case in which the sequence $\{X_{1,l}\}_{l=1}^s$ is well defined.

$$\leq \bar{n}_A \left(\frac{N}{\tau}+1\right) + \sum_{t\in\mathcal{F}_{\Delta^C,N}} \frac{1}{\tau} \leq \frac{19N}{\tau\log\tau} + \frac{2N}{\tau} + \frac{19}{\log\tau} + 1. \tag{92}$$

Let us focus on R_B. Let us define $\hat{\mu}_{i,t,\tau} := \frac{\sum_{s=t-\tau+1}^{t} X_{i,s}\mathbb{1}\{i_s=i\}}{T_{i,t,\tau}}$, *i.e.*, the estimator of the expected value of the rewards of arm a_i computed over the last τ rounds and its expected value $\mu_{i,t,\tau} := \frac{\sum_{s=t-\tau+1}^{t} \mu_{i,s}\mathbb{1}\{i_s=i\}}{T_{i,t,\tau}}$. Note that $-\mu_{i,t,\tau} \geq -\mu_{i,t} - \sigma\tau$ due to Assumption 2.

We can rewrite term R_B and apply Lemma 3 with $\bar{n}_{B*} = \left\lceil \frac{20\log\tau}{(\Delta-2\sigma\tau)^2} \right\rceil$ and $\bar{n}_B = \left\lceil \frac{32\log\tau}{(\Delta-2\sigma\tau)^2} \right\rceil$ as follows:

$$R_B = \sum_{t\in\mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(u_{T_{i,t,\tau}} > \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_\phi^*,t,\tau}}}, i_t = i\right) \tag{93}$$

$$= \sum_{t\in\mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} + \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}, i_t = i\right) \tag{94}$$

$$\leq \sum_{t\in\mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} + \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i_t^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_\phi^*,t,\tau}}}, T_{i_t^*,t,\tau} > \bar{n}_{B*}, T_{i,t,\tau} > \bar{n}_B\right)$$
$$+ \sum_{t\in\mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(T_{i_t^*,t,\tau} \leq \bar{n}_{B*}\right) + \sum_{t\in\mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(T_{i,t,\tau} \leq \bar{n}_B\right) \tag{95}$$

$$\leq \sum_{t\in\mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} + \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i_\phi^*,t} - \sigma\tau - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}, T_{i_t^*,t,\tau} > \bar{n}_{B*}, T_{i,t,\tau} > \bar{n}_B\right)$$
$$+ \bar{n}_{B*}\left\lceil\frac{N}{\tau}\right\rceil + \bar{n}_B\left\lceil\frac{N}{\tau}\right\rceil \tag{96}$$

$$= \sum_{t\in\mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t,\tau} + \mu_{i_t^*,t} - \mu_{i,t,\tau} - \sigma\tau - 2\sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right)$$
$$+ \left(\frac{N}{\tau}+1\right)\left[\frac{52\log\tau}{(\Delta-2\sigma\tau)^2}+2\right] \tag{97}$$

$$\leq \sum_{t\in\mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t,\tau} + \mu_{i_t^*,t} - \mu_{i,t} - \sigma\tau - \sigma\tau - 2\sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} - \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right)$$
$$+ \frac{N}{\tau}\frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + \frac{2N}{\tau} + \frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + 2 \tag{98}$$

$$\leq \sum_{t\in\mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t,\tau} + \Delta_{i,t} - 2\sigma\tau \underbrace{- \left(2\sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} + \sqrt{\frac{5\log\tau}{T_{i_t^*,t,\tau}}}\right)}_{\geq -(\Delta-2\sigma\tau)}\right)$$
$$+ \frac{N}{\tau}\frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + \frac{2N}{\tau} + \frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + 2 \tag{99}$$

$$\leq \sum_{t \in \mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t,\tau}\right) + \frac{N}{\tau}\frac{52\log\tau}{(\Delta - 2\sigma\tau)^2} + \frac{2N}{\tau} + \frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + 2, \quad (100)$$

where we use the property $\Delta_{i,t} > \Delta \forall i$, $\forall t \in \mathcal{F}_{\Delta^C,N}$ to bound Equation (99).

By using Corollary 21 by Garivier and Moulines (2008), we have the following for every $\eta > 0$:

$$\sum_{t \in \mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t,\tau}\right) \quad (101)$$

$$\leq \sum_{t \in \mathcal{F}_{\Delta^C,N}} \frac{\log\tau}{\log(1+\eta)} \exp\left(-12\log\tau\left(1 - \frac{\eta^2}{16}\right)\right) \quad (102)$$

thus, by using $\eta = 4\sqrt{1-\frac{1}{12}}$, we have:

$$\sum_{t \in \mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\hat{\mu}_{i,t,\tau} - \sqrt{\frac{2\log\tau}{T_{i,t,\tau}}} > \mu_{i,t,\tau}\right) \leq \sum_{t \in \mathcal{F}_{\Delta^C,N}} \frac{\log\tau}{\tau} \leq \frac{N\log\tau}{\tau}.$$

Hence, we can write:

$$R_B \leq \frac{N}{\tau}\frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + \frac{2N}{\tau} + \frac{N\log\tau}{\tau} + \frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + 2. \quad (103)$$

Let us focus on R_C. The R_C term is upper bounded by:

$$R_C = \sum_{t \in \mathcal{F}_{\Delta^C,N}} \mathbb{P}\left(\vartheta_{i,t} \geq q_{T_{i,t,\tau}}\right) = \sum_{t \in \mathcal{F}_{\Delta^C,N}} \frac{1}{\tau} \leq \frac{N}{\tau}.$$

Pseudo-regret. Summing all the derived bounds, the dynamic pseudo-regret becomes:

$$\bar{R}_N(\mathfrak{U}) = \mathbb{E}\left[\sum_{t=1}^{N}\left(\mu_{i_t^*,t} - \mu_{i_t,t}\right)\right] \quad (104)$$

$$\leq \sum_{i=1}^{K}\left(\mathbb{E}[T_i(\mathcal{F}_{\Delta,N})] + \mathbb{E}[T_i(\mathcal{F}_{\Delta^C,N})]\right) \quad (105)$$

$$\leq |\mathcal{F}_{\Delta,N}| + K\left(R_A + R_B + R_C\right) \quad (106)$$

$$= F\Delta N^\beta + K\left(\frac{19N}{\tau\log\tau} + \frac{2N}{\tau} + \frac{19}{\log\tau} + 1\right. \quad (107)$$

$$\left. + \frac{N}{\tau}\frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + \frac{2N}{\tau} + \frac{N\log\tau}{\tau} + \frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + 2 + \frac{N}{\tau}\right) \quad (108)$$

$$\leq F\Delta N^\beta + \frac{NK}{\tau}\left[\frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + \log\tau + 5 + \frac{19}{\log\tau}\right] \quad (109)$$

$$+ K\left[\frac{52\log\tau}{(\Delta-2\sigma\tau)^2} + 3 + \frac{19}{\log\tau}\right], \quad (110)$$

where we use the property $\sum_{i=1}^{K} \mathbb{E}[T_i(\mathcal{F}_{\Delta,N})] \leq |\mathcal{F}_{\Delta,N}| \leq F\Delta N^{\beta}$ that holds as Assumption 3 is satisfied. □

Corollary 3. *If the SW-TS policy is run over an SC-MAB setting with no switches between expected rewards of the arms ($P = 0$), in which Assumption 3 holds with $\beta \in [1 - \log_N\left(\frac{\Delta}{2\sigma}\right), 1]$, and using a sliding window $\tau := N^{1-\beta}$, for each $\Delta \leq \Delta_0$ the dynamic pseudo-regret is at most:*

$$\overline{R}_N(\mathfrak{U}) = \tilde{O}(N^{\beta}).$$

Proof. It is easy to derive that, if we choose a sliding window $\tau := N^{1-\beta}$, we minimize the asymptotic dynamic pseudo-regret w.r.t. N. By substituting τ in the expression of the dynamic pseudo-regret bound used in Theorem 2, we obtain the result stated in Corollary 3. However, the result stated in Theorem 2 holds only if the condition $2\sigma\tau \leq \Delta$ is satisfied. Since we set $\tau = N^{1-\beta}$, we have that:

$$\begin{aligned} 2\sigma N^{1-\beta} &\leq \Delta \\ N^{1-\beta} &\leq \frac{\Delta}{2\sigma} \\ 1 - \beta &\leq \log_N\left(\frac{\Delta}{2\sigma}\right) \\ \beta &\geq 1 - \log_N\left(\frac{\Delta}{2\sigma}\right), \end{aligned}$$

which concludes the proof. □

Corollary 4. *If the SW-TS policy is run over an SC-MAB setting with $P \in \mathbb{N}$ switches between expected rewards of the arms, in which Assumption 3 holds with:*

$$\beta \in \left[\max\left\{1 - \log_N\left(\frac{\Delta}{2\sigma}\right), \frac{1}{2} - \log_N\sqrt{\frac{F\Delta}{P}}\right\}, 1\right],$$

where $\max\{a, b\}$ denotes the maximum between a and b, and using a sliding window $\tau := N^{1-\beta}$, F is defined in Assumption 3, for each $\Delta \leq \Delta_0$ the dynamic pseudo-regret is at most:

$$\overline{R}_N(\mathfrak{U}) = \tilde{O}(N^{\beta}).$$

Proof. If two arms switch their average rewards over time, *i.e.*, it exists t s.t. $(\mu_{i,t} - \mu_{j,t})(\mu_{i,t+1} - \mu_{j,t+1}) \leq 0$, then the set $\mathcal{F}_{\Delta,N}$ is nonempty. In particular, for every switch, we have at least $\frac{2\Delta}{\sigma}$ rounds during which the difference of the average rewards is smaller than Δ. This is because a variation of the expected reward by Δ occurs in at least $\frac{\Delta}{\sigma}$ rounds before the switch and $\frac{\Delta}{\sigma}$ rounds after it.[14] Specifically, given Δ and having P switches over the time horizon N, we have that the number of rounds belonging to $\mathcal{F}_{\Delta,N}$ is at least:

$$\frac{P\Delta}{2\sigma} \leq |\mathcal{F}_{\Delta,N}| \leq F\Delta N^{\beta}, \tag{111}$$

14. Assume, for sake of simplicity, that Δ is a multiple of σ.

where the second inequality comes from Assumption 3, thus leading to:

$$\sigma \geq \frac{P}{2FN^{\beta}}. \tag{112}$$

From the assumption that $2\sigma\tau \leq \Delta$ and given that the length of the sliding window is $\tau = N^{1-\beta}$, we have:

$$\sigma \leq \frac{\Delta}{2N^{1-\beta}}. \tag{113}$$

To make both Equations (112) and (113) hold at the same time, we need to have an SC-MAB problem s.t.:

$$\frac{\Delta}{2N^{1-\beta}} \geq \frac{P}{2FN^{\beta}} \tag{114}$$

$$F\Delta N^{\beta} \geq PN^{1-\beta} \tag{115}$$

$$N^{2\beta-1} \geq \frac{P}{F\Delta} \tag{116}$$

$$2\beta - 1 \geq \log_N\left(\frac{P}{F\Delta}\right) \tag{117}$$

$$\beta \geq \frac{1}{2} - \log_N\sqrt{\frac{F\Delta}{P}}. \tag{118}$$

Given that Inequality (118) holds and condition $\beta \in [1 - \log_N\left(\frac{\Delta}{2\sigma}\right), 1]$, derived in Corollary 3, is satisfied, we can provide following range for β:

$$\beta \in \left[\max\left\{1 - \log_N\left(\frac{\Delta}{2\sigma}\right), \frac{1}{2} - \log_N\sqrt{\frac{F\Delta}{P}}\right\}, 1\right]. \tag{119}$$

□

Corollary 5. *If the SW-TS policy is run over an SC-MAB setting in which Assumption 3 holds with $\beta = 1$ and using a sliding window $\tau \propto \sigma^{-\frac{3}{4}}$, the average pseudo-regret is:*

$$\overline{AR}_N(\mathfrak{U}) = \tilde{O}(\sigma^{\frac{1}{2}}).$$

Proof. Defining $\varepsilon := \Delta - 2\sigma\tau$ ($\varepsilon \in (0, 1]$), independent of σ and τ, the dynamic pseudo-regret of the SW-TS algorithm is bounded by:

$$\begin{aligned}\frac{\overline{R}_N(\mathfrak{U})}{N} &\leq F\Delta + \frac{K}{\tau}\left[\frac{52\log\tau}{(\Delta - 2\sigma\tau)^2} + \log\tau + 5 + \frac{19}{\log\tau}\right] \\ &\quad + \frac{K}{N}\left[\frac{52\log\tau}{(\Delta - 2\sigma\tau)^2} + 3 + \frac{19}{\log\tau}\right] &(120)\\ &= F(\Delta - 2\sigma\tau) + 2F\sigma\tau + \frac{K}{\tau}\left[\frac{52\log\tau}{\varepsilon^2} + \log\tau + 5 + \frac{19}{\log\tau}\right] + \frac{K}{N}o(\log\tau) &(121)\\ &= 2F\sigma\tau + \frac{K}{\tau}\left[\frac{52\log\tau}{\varepsilon^2} + \log\tau + 5 + \frac{19}{\log\tau}\right] + F\varepsilon + \frac{K}{N}o(\log\tau) &(122)\end{aligned}$$

By substituting $\tau = C\sigma^{-\frac{1}{2}}$, we have:

$$\frac{\overline{R}_N(\mathfrak{U})}{N} \leq 2FC\sigma^{\frac{1}{2}} + \frac{58K\sigma^{\frac{1}{2}}\log\left(C\sigma^{-\frac{1}{2}}\right)}{C\varepsilon^2} + o(\sigma^{\frac{1}{2}}) + \frac{K}{N}o(\sigma^{\frac{1}{2}}), \tag{123}$$

which, as $N \to +\infty$, provides the inequality given in the corollary statement. □

Appendix D. Abrupt and Smoothly Changing Setting: Proofs

Theorem 3. *If the SW-TS policy is run over an ASC-MAB setting with $X_{i,t} \sim Be(\mu_{i,t})$, Lipschitz constant $\sigma > 0$ as in Assumption 4 and there exists $\Delta_0 \in (0,1)$ as in Assumption 3, for any $\tau \in \mathbb{N}$ s.t. $2\sigma\tau < \Delta \leq \Delta_0$, the dynamic pseudo-regret after N rounds is at most:*

$$\overline{R}_N(\mathfrak{U}) \leq F\Delta N^{\beta} + \tau B N^{\alpha} + \frac{NK}{\tau}\left[\frac{52\log\tau}{(\Delta - 2\sigma\tau)^2} + \log\tau + 5 + \frac{19}{\log\tau}\right],$$

where B and α are defined in Assumption 1 and F and β are defined in Assumption 3.

Proof. Let us define:

- $\mathcal{F}_{\Delta,N} := \{t \in \{1,\dots,N\} \text{ s.t. } \exists i \neq j, |\mu_{i,t} - \mu_{j,t}| < \Delta\}$, *i.e.*, the set of the rounds in which there are two arms whose expected values differ by less than Δ;
- $\mathcal{F}_{\Delta^C,N} := \{\tau,\dots,N\} \setminus \mathcal{F}_{\Delta,N}$, *i.e.*, the set of the rounds $t \geq \tau$, in which the expected rewards of the arms are well separated ($|\mu_{i,t} - \mu_{j,t}| > \Delta, \forall i \neq j\}$);
- $\mathcal{F}_{\Delta^C,\phi} := \{b_{\phi-1},\dots,b_\phi\} \setminus \mathcal{F}_{\Delta,N}$, *i.e.*, the set of the rounds of phase $\mathcal{F}_\phi$, in which the expected rewards of the arms are well separated;
- $\mathcal{F}'_{\Delta^C,\phi} := \{b_{\phi-1}+\tau,\dots,b_\phi\} \setminus \mathcal{F}_{\Delta,N}$, *i.e.*, the set of the rounds of phase $\mathcal{F}_\phi$ except the first τ rounds, in which the expected rewards of the arms are well separated;
- $T_i(\mathcal{F}) := \sum_{t\in\mathcal{F}} \mathbb{1}\{i_t = i, i \neq i_t^*\}$, *i.e.*, the number of rounds the arm a_i is played when it is not optimal during rounds $t \in \mathcal{F}$.

For every Δ s.t. $2\sigma\tau \leq \Delta$, we have that:

$$\overline{R}_N(\mathfrak{U}) = \mathbb{E}\left[\sum_{t=1}^{N}\left(\mu_{i_t^*,t} - \mu_{i_t,t}\right)\right] \leq \sum_{t=1}^{N}\mathbb{E}\left[\mathbb{1}\{i_t = i, i \neq i_t^*\}\right] = \sum_{i=1}^{K}\sum_{t=1}^{N}\mathbb{E}\left[\mathbb{1}\{i_t = i, i \neq i_t^*\}\right] \tag{124}$$

$$= \sum_{i=1}^{K}\mathbb{E}[T_i(\mathcal{F}_{\Delta,N})] + \sum_{i=1}^{K}\mathbb{E}[T_i(\mathcal{F}_{\Delta^C,N})] = \sum_{i=1}^{K}\mathbb{E}[T_i(\mathcal{F}_{\Delta,N})] + \sum_{\phi=1}^{B_N}\sum_{i=1}^{K}\mathbb{E}[T_i(\mathcal{F}_{\Delta^C,\phi})] \tag{125}$$

$$\sum_{i=1}^{K}\mathbb{E}[T_i(\mathcal{F}_{\Delta,N})] + \sum_{\phi=1}^{B_N}\left(\tau + \sum_{i=1}^{K}\mathbb{E}[T_i(\mathcal{F}'_{\Delta^C,\phi})]\right) \tag{126}$$

$$= \sum_{i=1}^{K} \mathbb{E}[T_i(\mathcal{F}_{\Delta,N})] + \tau B_N + \sum_{\phi=1}^{B_N} \sum_{i=1}^{K} \mathbb{E}[T_i(\mathcal{F}'_{\Delta^C,\phi})]. \tag{127}$$

The first term in Equation (127), *i.e.*, $\mathbb{E}[T_i(\mathcal{F}_{\Delta,N})]$, can be directly bounded by using Assumption 3. Each element $\mathbb{E}[T_i(\mathcal{F}'_{\Delta^C,\phi})]$ of the summation in the third term of Equation (127) can be bounded as we do for the second term in Equation (66) in the proof of Theorem 2 when using an opportune time horizon of length $N_\phi - \tau$ for phase $\mathcal{F}_\phi$. This follows since, when we focus on rounds belonging to a single phase, the setting is an SC-MAB. Formally, we have:

$$\mathbb{E}[T_i(\mathcal{F}'_{\Delta^C,\phi})] \leq \frac{N_\phi - \tau}{\tau} \left[\frac{52 \log \tau}{(\Delta - 2\sigma\tau)^2} + \log \tau + 5 + \frac{19}{\log \tau} \right] \tag{128}$$

$$\leq \frac{N_\phi}{\tau} \left[\frac{52 \log \tau}{(\Delta - 2\sigma\tau)^2} + \log \tau + 5 + \frac{19}{\log \tau} \right]. \tag{129}$$

Here, we omit the details about the derivation of the last equation. They are the same of the proof of Theorem 2, except for the fact that the time horizon we use here is $N_\phi - \tau$ instead of N and that the quantities $\left\lceil \frac{N_\phi - \tau}{\tau} \right\rceil$ can be upper bounded with $\frac{N_\phi}{\tau}$. This leads to Equation (129), which, differently from Equation (110), lacks of some terms that are not depending on N.

Finally, we have:

$$\overline{R}_N(\mathfrak{U}) \leq \sum_{i=1}^{K} \mathbb{E}[T_i(\mathcal{F}_{\Delta,N})] + \tau B_N + \sum_{\phi=1}^{B_N} \sum_{i=1}^{K} \mathbb{E}[T_i(\mathcal{F}'_{\Delta^C,\phi})] \tag{130}$$

$$\leq F\Delta N^\beta + \tau B N^\alpha + \sum_{\phi=1}^{B_N} \sum_{i=1}^{K} \frac{N_\phi}{\tau} \left[\frac{52 \log \tau}{(\Delta - 2\sigma\tau)^2} + \log \tau + 5 + \frac{19}{\log \tau} \right] \tag{131}$$

$$\leq F\Delta N^\beta + \tau B N^\alpha + \frac{NK}{\tau} \left[\frac{52 \log \tau}{(\Delta - 2\sigma\tau)^2} + \log \tau + 5 + \frac{19}{\log \tau} \right], \tag{132}$$

where we use Assumption 1 and Assumption 3 in Equation (131), and we use the property $\sum_{\phi=1}^{B_N} N_\phi = N$ in Equation (132). This concludes the proof. □

Corollary 6. *If the SW-TS policy is run over an ASC-MAB setting with no switches between expected rewards of the arms ($P = 0$) and Assumption 1 and Assumption 3 hold with $\alpha \in (1 - 2\log_N\left(\frac{\Delta}{2\sigma}\right), 1)$ and $\beta \in (0, 1)$, respectively, for each $\Delta \leq \Delta_0$, using a sliding window of $\tau := N^{\frac{1-\alpha}{2}}$, the dynamic pseudo-regret is at most:*

$$\overline{R}_N(\mathfrak{U}) = \begin{cases} \tilde{O}\left(N^{\frac{1+\alpha}{2}}\right) & \text{if } \beta \leq \frac{1+\alpha}{2} \\ \tilde{O}\left(N^{\beta}\right) & \text{if } \beta > \frac{1+\alpha}{2} \end{cases}.$$

Proof. The problem of choosing the proper order of the sliding window reduces here to minimizing the order in N of each of the first three terms of the dynamic pseudo-regret bound in Theorem 3. More specifically, the dynamic pseudo-regret can be bounded by:

$$\overline{R}_N(\mathfrak{U}) \leq C_1 N^\beta + C_2 \tau N^\alpha + C_3 \frac{N}{\tau} \tag{133}$$

where C_1, C_2 and C_3 are proper constants. Using a sliding window $\tau := N^\gamma$, with $\gamma \in (0,1)$, the minimization of the dynamic pseudo-regret order can be obtained by choosing the value of γ that minimizes $\max\{\beta, \alpha+\gamma, 1-\gamma\}$.

If $\beta \leq \frac{1+\alpha}{2}$, the minimum is obtained with an order $\gamma = \frac{1-\alpha}{2}$ and, thus, the overall dynamic pseudo-regret is of order $\tilde{O}(N^{\frac{1+\alpha}{2}})$. If $\beta > \frac{1+\alpha}{2}$, the minimization problem does not admit a single solution. However, a solution is $\gamma = \frac{1-\alpha}{2}$, thus providing an overall dynamic pseudo-regret of order $\tilde{O}(N^\beta)$.

In this case, the condition on the sliding window $2\sigma\tau \leq \Delta$ is:

$$\begin{aligned}
&2\sigma N^{\frac{1-\alpha}{2}} \leq \Delta \\
&N^{\frac{1-\alpha}{2}} \leq \frac{\Delta}{2\sigma} \\
&\frac{1-\alpha}{2} \leq \log_N\left(\frac{\Delta}{2\sigma}\right) \\
&\alpha \geq 1 - 2\log_N\left(\frac{\Delta}{2\sigma}\right),
\end{aligned}$$

which concludes the proof. □

Corollary 7. *If the SW-TS policy is run over an ASC-MAB setting with $P \in \mathbb{N}$ switches between expected rewards of the arms, and Assumption 1 and Assumption 3 hold with $\alpha \in (1 - 2\log_N\left(\frac{\Delta}{2\sigma}\right), 1)$ and $\beta \in (0,1)$, respectively, for each $\Delta \leq \Delta_0$, using a sliding window of $\tau := N^{\frac{1-\alpha}{2}}$, if $\beta + \frac{\alpha}{2} \geq \frac{1}{2} - \log_N\left(\frac{F\Delta}{P}\right)$ holds, the dynamic pseudo-regret is at most:*

$$\overline{R}_N(\mathfrak{U}) = \begin{cases} \tilde{O}\left(N^{\frac{1+\alpha}{2}}\right) & \text{if } \beta \leq \frac{1+\alpha}{2} \\ \tilde{O}\left(N^\beta\right) & \text{if } \beta > \frac{1+\alpha}{2} \end{cases}.$$

Proof. The analysis is similar to the one provided for Corollary 6. In addition to the properties we require in Corollary 6, we need to enforce another condition on the parameters α and β due to the number of switches between the expected rewards of the arms. More specifically, as shown in Equation (111) we have:

$$\sigma \geq \frac{P}{2FN^\beta},$$

and, since we use a sliding window of $\tau = N^{\frac{1-\alpha}{2}}$, we require that:

$$\sigma \leq \frac{\Delta}{2N^{\frac{1-\alpha}{2}}}. \tag{134}$$

Using the previous inequality together, we have:

$$\frac{\Delta}{2N^{\frac{1-\alpha}{2}}} \geq \frac{P}{2FN^\beta} \tag{135}$$

$$F\Delta N^\beta \geq PN^{\frac{1-\alpha}{2}} \tag{136}$$

$$N^{\frac{2\beta+\alpha-1}{2}} \geq \frac{P}{F\Delta} \tag{137}$$

$$\frac{2\beta+\alpha-1}{2} \geq \log_N\left(\frac{P}{F\Delta}\right) \tag{138}$$

$$\beta+\frac{\alpha}{2} \geq \frac{1}{2}-\log_N\left(\frac{F\Delta}{P}\right), \tag{139}$$

which concludes the proof. □

Corollary 8. *If the SW-TS policy is run over an SC-MAB setting in which Assumption 1 holds with $\alpha=1$, Assumption 3 holds with $\beta=1$, and using a sliding window $\tau \propto B^{-\frac{1}{4}}\sigma^{-\frac{3}{4}}$, the average pseudo-regret is:*

$$\overline{AR}_N(\mathfrak{U}) = \tilde{O}(B^{\frac{1}{2}}\sigma^{\frac{1}{2}}).$$

Proof. As in Corollary 5, let us define $\varepsilon := \Delta - 2\sigma\tau$ and, using the dynamic pseudo-regret bound in Theorem 3, we have:

$$\frac{\overline{R}_N(\mathfrak{U})}{N} \leq F(\Delta-2\sigma\tau)+2F\sigma\tau+\tau B+\frac{K}{\tau}\left[\frac{52\log\tau}{(\Delta-2\sigma\tau)^2}+\log\tau+5+\frac{19}{\log\tau}\right]. \tag{140}$$

Using a sliding window $\tau = CB^{-\frac{1}{2}}\sigma^{-\frac{1}{2}}$ we have:

$$\frac{\overline{R}_N(\mathfrak{U})}{N} \leq F\varepsilon+2FB^{\frac{1}{2}}\sigma^{\frac{1}{2}}+\frac{52KB^{\frac{1}{2}}\sigma^{\frac{1}{2}}\log\left(CB^{-\frac{1}{2}}\sigma^{-\frac{1}{2}}\right)}{\varepsilon^2}+o(B^{\frac{1}{2}}\sigma^{\frac{1}{2}}), \tag{141}$$

which concludes the proof. □

Appendix E. Smoothly Changing Setting: Satisfaction of Assumption 3 in the Experimental Setting of Section 5.2

We want to show that the number of the rounds, in which at least one pair $i,j \in \{1,\dots,K\}$ such that $i \neq j$ the inequality $|\mu_{i,t}-\mu_{j,t}| < \Delta$ holds, is upper bounded by $F\Delta$, where F is defined in Assumption 3. Let us set $\Delta_0 = \frac{1}{3}$, which leads to $\Delta \leq \frac{1}{3}$.

The evolution of the expected values of the arms over time in the SC-MAB we evaluate in Section 5 is the following:

$$\mu_{i,t} = \frac{K-1}{K} - \frac{\left|1+\frac{1}{2}(K-1)(1+\sin(t\sigma))-i\right|}{K}.$$

If we are in phase $\mathcal{F}_{\Delta,N}$, there exists a couple of index i and j, $i \neq j$ such that:

$$\begin{aligned}
&|\mu_{i,t}-\mu_{j,t}| \\
&= \left|\frac{K-1}{K} - \frac{\left|1+\frac{1}{2}(K-1)(1+\sin(t\sigma))-i\right|}{K} - \frac{K-1}{K} + \frac{\left|1+\frac{1}{2}(K-1)(1+\sin(t\sigma))-j\right|}{K}\right| \\
&= \left|-\frac{\left|1+\frac{1}{2}(K-1)(1+\sin(t\sigma))-i\right|}{K} + \frac{\left|1+\frac{1}{2}(K-1)(1+\sin(t\sigma))-j\right|}{K}\right| \\
&= \frac{1}{K}\Big|\left|1+\tfrac{1}{2}(K-1)(1+\sin(t\sigma))-j\right| - \left|1+\tfrac{1}{2}(K-1)(1+\sin(t\sigma))-i\right|\Big| \leq \Delta,
\end{aligned}$$

thus, we have:

$$-K\Delta < \left|1 + \tfrac{1}{2}(K-1)(1+\sin(t\sigma)) - j\right| - \left|1 + \tfrac{1}{2}(K-1)(1+\sin(t\sigma)) - i\right| < K\Delta.$$

In the following, we distinguish the analysis in two cases: the one in which the arguments of the absolute values have the same sign, formally t is such that $(1 + \frac{1}{2}(K-1)(1+\sin(t\sigma)) - j)(1 + \frac{1}{2}(K-1)(1+\sin(t\sigma)) - i) > 0$, from the one in which they have opposite signs, formally t is such that $(1 + \frac{1}{2}(K-1)(1+\sin(t\sigma)) - j)(1 + \frac{1}{2}(K-1)(1+\sin(t\sigma)) - i) < 0$.

Case 1. Let us consider the case in which both $1 + \frac{1}{2}(K-1)(1+\sin(t\sigma) > 0$ and $1 + \frac{1}{2}(K-1)(1+\sin(t\sigma)) - i > 0$. The same holds in the case both terms are negative and by inverting the roles of i and j. In the former case, the inequality becomes:

$$\begin{aligned}
&-K\Delta < 1 + \tfrac{1}{2}(K-1)(1+\sin(t\sigma)) - j - 1 - \tfrac{1}{2}(K-1)(1+\sin(t\sigma)) + i < K\Delta \\
&-K\Delta < i - j < K\Delta.
\end{aligned}$$

If $i > j$, the inequality $-K\Delta < i - j$ is always satisfied, while we need to examine whether $i - j < K\Delta$ or not. The worst case is when the two arms are $i = K$ and $j = 1$:

$$\begin{aligned}
&i - j < K\Delta \\
&K - 1 < K\Delta \\
&\Delta > \frac{K-1}{K},
\end{aligned}$$

which is always false if $K > 2$ since $\Delta_0 = \frac{1}{3}$. Thus, the set of the rounds in this case is empty.

If $i < j$, the inequality $i - j < K\Delta$ is always satisfied, while we need to verify $-K\Delta < i - j$. The worst case is when $i = 1$ and $j = K$, thus:

$$\begin{aligned}
&-K\Delta < i - j \\
&-K\Delta < 1 - K \\
&\Delta > \frac{K-1}{K},
\end{aligned}$$

thus, the same reasoning made for the previous case holds and we have an empty set of rounds.

Case 2. Even in this case we analyse only the case in which $1 + \frac{1}{2}(K-1)(1+\sin(t\sigma)) - j > 0$ and $1 + \frac{1}{2}(K-1)(1+\sin(t\sigma)) - i < 0$, the opposite one being analogous. We have that the initial condition becomes:

$$\begin{aligned}
&-K\Delta < 1 + \tfrac{1}{2}(K-1)(1+\sin(t\sigma)) - j + 1 + \tfrac{1}{2}(K-1)(1+\sin(t\sigma)) - i < K\Delta \\
&-K\Delta < 2 + (K-1)(1+\sin(t\sigma)) - j - i < K\Delta \\
&-K\Delta - 2 + j + i < (K-1)(1+\sin(t\sigma)) < K\Delta - 2 + j + i \\
&\frac{-K\Delta - 2 + j + i}{K-1} - 1 < \sin(t\sigma) < \frac{K\Delta - 2 + j + i}{K-1} - 1 \\
&\frac{1}{\sigma}\arcsin\left(\frac{-K\Delta - 2 + j + i}{K-1} - 1\right) < t < \frac{1}{\sigma}\arcsin\left(\frac{K\Delta - 2 + j + i}{K-1} - 1\right)
\end{aligned}$$

We are interested in the number of rounds for which the inequalities hold, *i.e.*,:

$$\begin{aligned}\overline{T} &= \left|\left\{t : \frac{1}{\sigma}\arcsin\left(\frac{-K\Delta-2+j+i}{K-1}-1\right) < t < \frac{1}{\sigma}\arcsin\left(\frac{K\Delta-2+j+i}{K-1}-1\right)\right\}\right| \\ &= \frac{1}{\sigma}\arcsin\underbrace{\left(\frac{K\Delta-2+j+i}{K-1}-1\right)}_{\Delta_a} - \frac{1}{\sigma}\arcsin\underbrace{\left(\frac{-K\Delta-2+j+i}{K-1}-1\right)}_{\Delta_b},\end{aligned}$$

where $|\cdot|$ is the cardinality operator.

By relying on the following inequalities:

$$\begin{aligned}\arcsin(x) &\leq 2x \qquad x \leq 0, \\ \arcsin(x) &\geq 2x \qquad x \geq 0,\end{aligned}$$

we have that if $\Delta_a \leq 0$ and $\Delta_b \geq 0$, we can write:

$$\overline{T} \leq \frac{2}{\sigma}\left(\frac{K\Delta-2+j+i}{K-1}-1\right) - \frac{2}{\sigma}\left(\frac{-K\Delta-2+j+i}{K-1}-1\right) = \frac{4K\Delta}{\sigma(K-1)},$$

thus Assumption 3 is satisfied with $F := \frac{4K}{\sigma(K-1)}$.

Finally, we have to show that $\Delta_a \leq 0$ and $\Delta_b \geq 0$. Let us start with $\Delta_a \leq 0$. The value minimizing Δ_a for the indexes are $i = 1$ and $j = 2$, consequently, we have:

$$\begin{aligned}&\frac{K\Delta-2+j+i}{K-1}-1 = \frac{K\Delta-2+2+1-K+1}{K-1} = \frac{K\Delta-K+2}{K-1} = \frac{(\Delta-1)K+2}{K-1} \leq 0 \\ &\Delta \leq \frac{K-2}{K},\end{aligned}$$

which is satisfied since $\Delta_0 \leq \frac{1}{3}$ for $K \geq 3$.

Let us consider $\Delta_b \geq 0$. Even in this case the choice of $i = 1$ and $j = 2$ is the one providing the most restrictive conditions. We have:

$$\frac{-K\Delta-2+j+i}{K-1}-1 = \frac{-K\Delta-2+2+1-K+1}{K-1} = \frac{-K\Delta-K+2}{K-1} = \frac{-(\Delta+1)K+2}{K-1} \geq 0,$$

which is the same condition as in the $\Delta_a \geq 0$ derivations.