Improving Multi-Armed Bandit Algorithms in Online Pricing Settings

Francesco Trovò, Stefano Paladino, Marcello Restelli, Nicola Gatti

[a] *Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, piazza Leonardo da Vinci* 32*, Milano, Italy* 20133

Abstract

The design of effective bandit algorithms to learn the optimal price is a task of extraordinary importance in all the settings in which the demand curve is not *a priori* known and the estimation process takes a long time, as customary, e.g., in e-commerce scenarios. In particular, the adoption of effective pricing algorithms may allow companies to increase their profits dramatically. In this paper, we exploit the structure of the pricing problem in online scenarios to improve the performance of state-of-the-art general-purpose bandit algorithms. More specifically, we make use of the *monotonicity* of the customer demand curve, which suggests the same behaviour of the conversion rates, and we exploit the fact that, in many scenarios, companies have *a priori information* about the order of magnitude of the conversion rate. We design techniques—applicable in principle to any bandit algorithm—capable of exploiting these two properties, and we apply them to Upper Confidence Bound policies both in stationary and nonstationary environments. We show that algorithms exploiting these two properties may significantly outperform state-of-the-art bandit policies in most of the configurations and we also show that the improvement increases as the number of arms increases. In particular, simulations based on real-world data show that our algorithms may increase the profit by 300% or more when compared to the performance achieved by state-of-the-art bandit algorithms. Furthermore, we formally prove that the empirical improvement provided by our algorithms can be achieved without incurring any cost in terms of theoretical guarantees. Indeed, our algorithms present the same asymptotic worst-case regret bounds of the bandit algorithms previously known in the state of the art.

1. Introduction

We focus on the problem of learning the best price to apply to a good (a.k.a. *pricing problem*) when a seller has an unlimited amount of non-perishable

Email addresses: `francesco1.trovo@polimi.it` (Francesco Trovò), `stefano.paladino@polimi.it` (Stefano Paladino), `marcello.restelli@polimi.it` (Marcello Restelli), `nicola.gatti@polimi.it` (Nicola Gatti)

goods in online scenarios [1]. Although such a setting is basic, it perfectly fits with (or is a sufficiently accurate approximation of) many real-world applications, such as the sale of streaming services (e.g., movies and music) and digital products (e.g., software). Basically, the pricing problem is characterized by a *price*, defined as the sum of the *cost* of the good for the seller and the *gross margin* chosen by the seller, and a *conversion rate*, measuring the probability that the good will be sold at a given price, which may be unknown to the seller. Furthermore, in a pricing problem, the behavior of the customers may be *nonstationary*, thus making the average conversion rate to change over time. Usually, there is a sequence of *phases* such that during each one of them the behavior is stationary (in this case, the change between two consecutive phases may be due to, e.g., a new product entering the market). Extremely low conversion rates (as customary in e-commerce) make the estimation process excessively long, usually longer than the time between two consecutive phases. As a result, the estimation process rarely converges to stable solutions, and it is in a transient for most of the time.[1] Therefore, the effectiveness of a pricing algorithm mainly depends on its performance during the transient, and this makes the problem of finding the best price an *online learning problem*.

In an online learning problem, a *learner* chooses at each round an option, customarily called *arm* (in our case, corresponding to a gross margin), and observes the stochastic reward associated with the arm. The goal of the learner is to identify the best arm in terms of expected reward while minimizing the loss incurred from pulling sub-optimal arms. The two main opposite settings are the *bandit* setting—a.k.a. Multi-Armed Bandit (MAB) [2]—, in which at each round the learner observes the reward associated with *only* the pulled arm, and the *expert* setting [3], where at each round the learner observes the reward associated with *all* the arms. The bandit and expert settings are well assessed in the scientific literature, while the study of all the situations lying between these two extremes represents a widely unexplored research area, whose goal is the exploitation of the problem structure to improve the performance of the traditional bandit algorithms.

The main work dealing with bandits for pricing is provided in [1], where the authors study the value of knowing the demand curve—assumed continuous and smooth—in stationary settings both in the stochastic and in the adversarial cases and provide a tight regret-bound analysis. Furthermore, the authors provide an algorithm to select a finite set of arms to which each state-of-the-art MAB algorithm can be applied. In this paper, we focus on a different problem. Specifically, we exploit the properties of the conversion rates in the pricing problem to design algorithms capable of improving the empirical performance of the state-of-the-art MAB algorithms both in stationary and nonstationary stochastic settings. Thus, in principle, our algorithms can be paired with the work provided in [1]. Empirically, our algorithms significantly outperform the

[1] As customary, with "transient" we mean the early stages of the learning process that are far from the convergence.

techniques available in the state of the art. This is achieved without incurring any cost in terms of theoretical guarantees. Indeed, our algorithms present the same asymptotic regret bounds of the algorithms for pricing previously known. The specific properties we exploit in our paper, unexplored in the literature so far, are as follows.

The first property is the *weak decreasing monotonicity* of the customer's demand curve in the price. That is, the larger the price, the smaller the number of customers that buy the good, or, equivalently, the larger the price, the smaller the conversion probability. Monotonicity assumption is very common in economics [4] and applies to a wide spectrum of settings. For instance, the demand curves are monotonic in markets in which the goods/services are very similar, and thus every competitor has a similar market share. Many Internet services are characterized by a monotonic demand curve, e.g., for all those that provide basic accounts for free and require to pay for premium features (a.k.a. *freemium* services) [5]. The free basic account is crucial for attracting users to increase the popularity of the service and, thus, to increase the value for a user in adopting the service. In this scenario a vast majority of the users are not willing to pay the premium account at any price, while the remaining fraction of users will consider to pay a fee to be able to access the extra services. Restricting the analysis on these users, we have that the demand curve for the purchase of the premium account is monotonically decreasing in the price. Notice that monotonicity is common also in many other application domains different from the pricing. For instance, in multi-slot *online advertising*, where it is necessary to estimate the Click-Through Rates (CTRs) of ads [6] and the expected value of the CTR of an ad monotonically decreases from the slot in the top to the one in the bottom; and in *bandwidth allocation*, where it is necessary to estimate the best packet size for the link between some servers [7] and, if a packet has been successfully transmitted, also a smaller one would have been received too.

When monotonicity in the price holds, each time a buyer makes a purchase at a given price, we can infer that the sale would have also been made at any lower price, and, *vice versa*, each time the buyer refuses to buy at a certain price, we can infer that all the higher prices would not have been accepted too. Providing such information to the algorithms allows them to speed up the learning process. We are also interested in the weak decreasing monotonicity of the demand curve in the gross margin. When this property holds, our algorithms can be used to maximize the profit. Instead, when the demand curve is monotonic in the price but not in the gross margin, our algorithms can be used to maximize the revenue.

The second property concerns the fact that e-commerce sellers have *a priori* information about the customer behavior, coming from past transactions. Usually, this information is not sufficient for producing sufficiently accurate estimates to avoid a cold start of the learning algorithms, e.g., in the case the seller pulled in the past a limited number of arms or the market is nonstationary and, therefore, too old information is not meaningful. However, such information is sufficient to estimate a lower bound to the percentage of the buyers that are only interested in checking the price without buying the item, which leads to a low probability of purchasing a good [8] (e.g., it is common that human users check the price for some days before buying an item as well as it is common that companies use bots to frequently check the prices of the competitors). As a result, for every specific pricing setting (e.g., associated with a product) we can set an upper bound over the curve of the conversion rate as a function of the price (or the gross margin). This may allow exploiting tighter concentration inequalities, thus reducing the experience needed to get accurate estimates of the expected conversion rate, and, consequently, reducing the loss due to the algorithm exploration.

1.1. Related Works

Several previous works exploit the structure of specific classes of sequential games to improve the performance provided by the general-purpose algorithms. For instance, in [9, 10] the authors present policies for MAB problems where the expected reward is *unimodal* over partially ordered arms. However, the assumption of unimodal reward is strong and rarely met in practice in microeconomics problems. Interestingly, the assumptions of monotonicity and unimodality are orthogonal, none of them being a special case of the other one and, therefore, the results known for unimodal bandits cannot be directly adopted in monotonic settings. A different approach is presented in [11, 12], where the authors study a graph model for the arm feedback in an adversarial setting under the assumption that the realizations are correlated and that this correlation is known. The treatment of this last assumption is different from the treatment of the monotonicity assumption, where, conversely, the correlation is over the expected value of the arms and not over the realizations. In [13, 14], the authors propose a more general setting named *partial monitoring* games, for which several studies on asymptotic regret bounds have been produced in the last decade both in stochastic [15, 16] and adversarial [7, 14, 17] settings. To the best of our knowledge, no work takes advantage of the monotonicity property as defined above or exploits *a priori* information about the magnitude order of low conversion probabilities.

In the economics literature and, more precisely, in the subarea of *learning and earning*, several works study the pricing problem [18, 19, 20, 21]. Most of these works assume that *a priori* information on the structure of the problem is available (e.g., on the product supply availability or the user behavior). More specifically, [18] considers a limited initial inventory of a single product and designs a parametric and a non-parametric algorithm to estimate the demand function. In [19], the authors design the LLVD algorithm, based on the Bayesian framework, which assumes that the demand curve linearly decreases with the price. Several works propose techniques to learn the optimal price under the assumption that the expected revenue curve has a unique global optimal solution [20, 21, 22]. Finally, in [23] the authors consider the case of an adversarial model for the user in an online posted-price auction and directly applies the Exp3 algorithm [24] to minimize the regret. Remarkably, most of the works in the *learning and earning* field do not provide any theoretical guarantee on the regret bounds. Even if heuristic algorithms might perform better than the algorithms with theoretical guarantees, the lack of worst-case guarantees discourages their employment in practice.

A problem related to pricing is the design of *nearly-optimal auctions* in the case the bidders' valuations are drawn from an unknown distribution [25, 26]. The proposed solution relies on statistical learning theory techniques to compute the number of samples required to bound the distance of the approximated solution from the real expected revenue.

The related works discussed above focus on stationary settings. The MAB literature also provides several works addressing nonstationary settings. In [27, 28], the authors study an abruptly changing environment and propose the SW-UCB algorithm, that exploits a sliding window approach. In economic domains, an abrupt change can be due to the invasion of the market by a new product. Instead, in [9, 10], the authors present policies working in a unimodal and smoothly changing environment. Finally, [29] presents an evolutionary algorithm for a nonstationary setting. The proposed algorithm outperforms the state-of-the-art ones, but no theoretical guarantees on the regret are provided.

1.2. Original Contributions

In the present paper, we study the stochastic MAB setting on a finite number of arms, and we propose techniques to exploit the monotonicity property of conversion rates as well as the *a priori* information on the maximum conversion rate. Our techniques can be paired, in principle, with any frequentist MAB algorithm, while the extension to Bayesian MAB policies (e.g., Thompson Sampling [30]) is left open. In this paper, we tailor our techniques for two main Upper Confidence Bound (UCB) like algorithms working in stationary settings: UCB1 [31], being the most popular and basic MAB algorithm, and UCBV [32], being one of the UCB-like algorithms with the best empiric performance. We prove that the asymptotic regret bounds of our algorithms are of the same order as UCB1 and UCBV. Furthermore, we provide an analysis for nonstationary settings, tailoring our techniques for SW-UCB and providing a regret bound analysis for the modified algorithm. We present a thorough experimental evaluation of our algorithms in several different configurations based on real-world data. We compare our algorithms with the main general-purpose frequentist stochastic MAB policies both in stationary and nonstationary settings, showing that exploiting the two aforementioned properties allows one to significantly improve the profit—up to 300%. Overall, the empirical analysis shows that our algorithms provide significant advantages with respect to general-purpose MAB algorithms in the early stages of the learning process. This is crucial in real pricing scenarios, where very low conversion rates (that require a long exploration phase to have accurate estimations) and nonstationary buyers' demands make the algorithms to work in a never-ending transient.

1.3. Paper Organization

The remaining part of the paper is structured as follows. Section 2 provides the formulation for the MAB setting we study. Section 3 describes the proposed techniques in stationary settings, while Section 4 describes the proposed techniques in nonstationary settings. Section 5 provides experimental results in stationary settings, while Section 6 provides experimental results in nonstationary settings. Finally, in Section 7 the conclusions of this work are drawn. In the appendices, we provide supplemental material. More precisely, the proofs of the theorems are reported in Appendix A, the pseudocode of some algorithms can be found in Appendix B, and additional experimental results are provided in Appendix C.

2. Problem Formulation

We study a scenario where an unlimited non-perishable amount of goods is available to a monopolistic seller, who proposes the product she is selling to some unknown buyers at a chosen price. For the sake of simplicity, we assume the costs of the seller to be a constant equal to zero, and therefore the gross margin and the price are equal.[2] We model our problem as a MAB problem [31], where at each round $t \in \{1, \dots, N\}$ over a finite horizon N the seller selects an arm, corresponding to a gross margin, among a strictly ordered finite set of K different arms $A = \{a_1, \dots, a_K\}$ with $a_i \in (0, +\infty)$. As customary in microeconomics, each buyer is modeled as a deterministic agent who buys the item only if the proposed gross margin is lower than or equal to a threshold $s \in \mathbb{R}^+$. Thus, all the gross margins that are at most s lead to a sale, while all the ones that are higher than s lead to a non-sale. Since buyers generally have different thresholds s, we model s as realizations of a random variable S with a probability density function (pdf) $\mathcal{S}$ over the finite support $\Omega \subset \mathbb{R}^+$. In stationary settings, the pdf $\mathcal{S}$ is unique for all the rounds, whereas in nonstationary settings each round t presents a potentially different pdf $\mathcal{S}_t$. We assume that the pdfs are unknown to the seller and therefore that the seller needs to estimate them. The gross margin a_i also represents the reward received by the seller once she sold the product. Thus, the seller aims at maximizing of the total expected profit over the time horizon N. A MAB *policy* is an algorithm $\mathfrak{U}(h_t)$ that chooses the next arm a_{i_t} to play at round t given history h_t, defined as the sequence of past plays and obtained rewards. At each round t the algorithm observes a single realization of the reward V_{i_t} obtained from the arm $a_{i_t} = \mathfrak{U}(h_t)$.

2.1. Stationary Pricing Model

In the case of stationary settings, the reward gained by pulling an arm a_i is a bounded random variable $V_i = a_i X_i$, where $X_i \sim Be(\mu_i)$ is a Bernoulli variable that represents the *outcome* (buy/not buy) of the transaction, where $\mu_i := \mathbb{E}[X_i]$ is the expected value of the outcome corresponding to arm a_i, i.e.,

[2] Our algorithms can be used to maximize the profit whenever the costs are known and the demand curve is weakly monotonically decreasing in the gross margin. In this case, the problem of finding the best price can be formulated as the problem of finding the best gross margin. If the demand curve is not monotonic in the gross margin or the costs are not known, our algorithms can be used to maximize the revenue. In this case, our algorithms control directly the price and not the gross margin.

the conversion rate. We denote with $V_{i,n}$ and $X_{i,n}$ the random variable of the reward and the outcome of the n-th pull of the i-th arm, respectively, and with $v_{i,n}$ and $x_{i,n}$ their realizations. We denote with $T_i(t) = \sum_{m=1}^{t} \mathbb{1}\{\mathfrak{U}(h_m) = a_i\}$ the number of times the arm a_i was pulled in the first t rounds, where $\mathbb{1}\{B\}$ is the indicator function of the event B. The objective of a policy is the maximization of the expected cumulative reward or, equivalently, the minimization of the loss with respect to the optimal decision (in terms of reward). This loss is usually addressed as *(cumulative) pseudo-regret*, whose definition over the time horizon N is:

$$\overline{R}_N = a_{i^*}\mu_{i^*}N - \sum_{i=1}^{K} a_i \mu_i \mathbb{E}[T_i(N)],$$

where $i^* = \arg\max_{i \in \{1,\dots,K\}} a_i \mu_i$ is the optimal arm and $\mathbb{E}[\cdot]$ is the expectation with respect to the stochastic components of the policy.

2.2. Nonstationary Pricing Model

In the case of nonstationary settings, we analyse an *abruptly changing environment*, similarly to what has been studied in [27], where the pdf $\mathcal{S}_j$ describing the buyer behavior is constant during sequences of rounds called *phases* and changes at unknown rounds called *breakpoints*. Thus, differently from the stationary scenario, the expected value of the outcome $\mu_{i,t}$ of an arm a_i at round t changes over the phases and therefore the best arm $a_{i^*,t}$ might change after each breakpoint.

A breakpoint $b \in \{1, \dots, N\}$ is a round such that $\exists i \mid \mu_{i,b-1} \neq \mu_{i,b}$, i.e., a round b where the expected reward of at least one arm changed with respect to the one at round $b-1$. In a nonstationary environment $\mathcal{S}^{(B)}$ with time horizon N we have a set of breakpoints $B = \{b_1, \dots, b_{\Upsilon_N}\}$ of cardinality Υ_N (for sake of notation we define $b_1 = 1$), which determines a set of phases $\{\Phi_\phi\}_{\phi=1}^{\Upsilon_N}$, where $\Phi_\phi = \{t | b_{\phi-1} \leq t < b_\phi\}$, i.e., the set of rounds between two consecutive breakpoints. During phase Φ_ϕ, we denote (with abuse of notation) with $\mu_{i,\phi}$ the expected value of the outcome of the i-th arm a_i and with $\mu_{i^*,\phi}$ the expected conversion probability corresponding to the best arm $a_{i^*,\phi}$. By defining $N_\phi = |\Phi_\phi|$, the cumulative pseudo-regret of a generic policy over a nonstationary environment is:

$$\begin{aligned}
\overline{R}_N &= \mathbb{E}\left[\sum_{t=1}^{N} (a_{i^*,t}\mu_{i^*,t} - a_{i_t}\mu_{i_t,t})\right] = \sum_{\phi=1}^{\Upsilon_N} a_{i^*,\phi}\mu_{i^*,\phi}N_\phi - \mathbb{E}\left[\sum_{t=1}^{N} a_{i_t}\mu_{i_t,t}\right] \\
&= \sum_{\phi=1}^{\Upsilon_N} \left(a_{i^*,\phi}\mu_{i^*,\phi}N_\phi - \mathbb{E}\left[\sum_{t\in\Phi_\phi} a_{i_t}\mu_{i_t,t}\right]\right) \\
&= \sum_{\phi=1}^{\Upsilon_N} \left(a_{i^*,\phi}\mu_{i^*,\phi}N_\phi - \sum_{i=1}^{K} a_i\mu_{i,\phi}\mathbb{E}[T_i(\Phi_\phi)]\right),
\end{aligned}$$

where $\sum_{\phi=1}^{\Upsilon_N} N_\phi = N$, $T_i(\Phi_\phi) = \sum_{m\in\Phi_\phi} \mathbb{1}\{\mathfrak{U}(h_m) = a_i\}$ is the number of times the i-th arm a_i has been pulled during phase Φ_ϕ and $\mathbb{E}[\cdot]$ is the expectation with respect to the stochastic components of the policy.

2.3. Properties of the Pricing Problem

We exploit two properties of the probability distributions of the random variables $X_{i,n}$ representing the outcomes of the transactions. The first property is the *dependency* between arms. While in the classic MAB setting the rewards produced by different arms are assumed to be drawn from independent probability distributions, in our setting this does not hold anymore, since the realizations at time t (i.e., $x_{1,T_1(t)}, \ldots, x_{K,T_K(t)}$) of the outcome variables $X_{1,T_1(t)}, \ldots, X_{K,T_K(t)}$ are correlated by the threshold of the buyer that plays at round t. The expected *conversion probability* $\mu_{i,\phi}$ corresponding to gross margin a_i at phase Φ_ϕ is defined as the probability that a user purchases the product or formally:

$$\mu_{i,\phi} := \mathbb{P}_{S_\phi}(s \geq a_i) = 1 - \int_0^{a_i} \mathcal{S}_\phi(x)dx.$$

Notice that, in stationary settings, we have a single probability distribution, thus $\mathcal{S}_\phi = \mathcal{S}$ and $\mu_{i,\phi} = \mu_i$. From the non-negativity of the probability distribution function $\mathcal{S}_\phi$ and from the properties of the integral, it clearly follows that $a_i < a_j \Rightarrow \mu_{i,\phi} \geq \mu_{j,\phi}$, i.e., the expected conversion probability is *monotonically* (weakly) decreasing with respect to the gross margin.

The second property concerns the *low conversion rates*, which are common in many e-commerce applications. In this case, the seller knows that only a certain percentage of the buyers $\mu_{\max} \in [0, 1]$ (typically $\mu_{\max} \ll 1$) really considers the possibility of purchasing the good, while the remaining part $1 - \mu_{\max}$ would not buy at any price. Such behavior can be introduced in the user model by considering $\mathcal{S}_\phi$ with pdf equal to $\mathcal{S}_\phi(x) = (1 - \mu_{\max}) \cdot \delta(0) + \mu_{\max} \cdot \mathcal{C}_\phi(x), x \in \Omega$, where $\delta(0)$ is a Dirac delta probability distribution centered in 0 and $\mathcal{C}_\phi(\cdot)$ is a pdf defined over Ω.

3. Exploiting Pricing Property in Stationary Settings

In this section, we describe techniques exploiting the pricing problem structure in stationary settings. We use the monotonicity structure of the expected value of the outcome $\{\mu_i\}_{i=1}^K$ of the arms to tighten the UCBs used in the frequentist approach. The proposed techniques are then applied to UCB1 [31] and UCBV [32], as interesting case studies. Furthermore, to exploit the prior knowledge about low conversion rates, we propose the use of a form of the Chernoff's bound [33] which, in this case, is tighter than the Hoeffding's one. Finally, we provide an algorithm that combines both techniques.

3.1. Exploiting the Monotonicity Property

Given an arm a_i, the realizations of all the outcomes X_j with $j < i$ provide information that can be exploited for the computation of the UCB on the expected value μ_i. Indeed, since $\mu_i \leq \mu_j$, we can use the realizations drawn so far from X_j as optimistic samples to estimate μ_i. In what follows, we will derive a set of bounds which exploit the samples coming from arms with lower values and consider the tightest among them to design an algorithm for the pricing scenario. Let $\overline{X}_{i,t}$ be the empirical mean, at round t, of the outcomes obtained by pulling arm a_i for $T_i(t-1)$ rounds (i.e., an estimator of the expected conversion rate μ_i of arm a_i) and $\overline{x}_{i,t}$ its realization, or formally:

$$\overline{X}_{i,t} := \frac{1}{T_i(t-1)} \sum_{n=1}^{T_i(t-1)} X_{i,n}, \quad \overline{x}_{i,t} := \frac{1}{T_i(t-1)} \sum_{n=1}^{T_i(t-1)} x_{i,n}.$$

Similarly, given $1 \leq j \leq i$, let $\overline{X}_{ji,t}$ be the following convex combination of the sample means $\overline{X}_{j,t}, \ldots, \overline{X}_{i,t}$ and let $\overline{x}_{ji,t}$ be its realization:

$$\overline{X}_{ji,t} := \frac{\sum_{k=j}^{i} T_k(t-1)\overline{X}_{k,T_k(t-1)}}{T_{ji}(t-1)}, \qquad \overline{x}_{ji,t} := \frac{\sum_{k=j}^{i} T_k(t-1)\overline{x}_{k,t}}{T_{ji}(t-1)},$$

where $T_{ji}(t-1) = \sum_{k=j}^{i} T_k(t-1)$, corresponding to the cumulative number of rounds all the arms from j to i have been pulled. Since, given the monotonicity property, it holds:

$$\mu_{ji,t} = \mathbb{E}\left[\overline{X}_{ji,t}\right] \geq \mu_i,$$

any upper bound on $\mu_{ji,t}$ is also an upper bound on μ_i. This allows us to bound the expected value μ_i of the outcome X_i of arm a_i by using samples drawn from the set of outcomes $\{X_1, \ldots, X_i\}$. In other words, we can compute an upper bound on the expected conversion rate associated with arm a_i by taking into account also the experience collected when lower arms were selected. By considering concentration bounds over the aggregated variables X_{ji} with $j \in \{1, \ldots, i\}$, we may find a tighter bound, which also holds for the expected value of the outcome X_i. In what follows, we apply this idea to the concentration bounds used in UBC1 and UCBV policies.

3.1.1. UCB1 with Monotonic Arms (UCB1-M)

Applying the Hoeffding's inequality [34] to the random variables $\overline{X}_{ji,t}$, with probability at least $1 - \frac{p}{i}$ where $p \in [0, 1]$, we have the following UCBs (from now on denoted as UCB1-M):

$$u_{ji,t}^{(\text{UCB1-M})} = \overline{x}_{ij,t} + \sqrt{\frac{\log(i) - \log(p)}{2T_{ji}(t-1)}} > \mu_{ji,t} \geq \mu_i \qquad \forall j \in \{1, \ldots, i\}. \tag{1}$$

Since, for each $j \in \{1, \ldots, i\}$, $u_{ji,t}^{(\text{UCB1-M})}$ is a valid upper bound on μ_i holding with at least probability $1 - \frac{p}{i}$, by setting $u_{i,t}^{(\text{UCB1-M})} = \min_{j \in \{1,\ldots,i\}} u_{ji,t}^{(\text{UCB1-M})}$ and resorting to a union bound, we have the tightest bound among those provided by Equation (1), holding with at least probability $1 - p$.

The use of the UCB1-M bound constitutes a potential improvement over the traditional one used by the UCB1 algorithm and obtained by considering realizations coming from a single arm. Indeed, this novel UCB exploits $T_{ji}(t-1) \geq T_i(t-1)$ samples and may be tighter than the UCB1 one. If the observed empirical means are consistent with the monotonicity property (i.e., $\overline{x}_{i,t} < \overline{x}_{j,t},\ \forall i > j$) the use of a larger number of samples coming from other arms may allow one (specially in the early stages) to tighten the bound. The proposed method is even more advantageous when empirical means are not consistent with the monotonicity property (i.e., $\exists\ i > j$ such that $\overline{x}_{i,t} > \overline{x}_{j,t}$). In this case, the bound $u_{i,t}^{(\text{UCB1-M})}$ is significantly improved over the original UCB1 bound. Such a situation is exemplified in Figure 1, where we have that, in contrast with the monotonicity over $A = \{a_1, a_2\}$, the empirical mean of the outcome corresponding to arm $a_1 = 1$, i.e., $\overline{x}_{1,t}$, is lower than the one of arm $a_2 = 2$, i.e., $\overline{x}_{2,t}$. This happens because arm a_2 has been selected much less often than arm a_1 and so its empirical mean is more uncertain. The samples drawn from arm a_1 allow to tighten the UCB for arm a_2 from the value denoted by the blue circle to the value denoted by the red square in Figure 1 (top). The use of the proposed UCB for arm a_2 does not imply a reduction in the confidence level since the two values have been obtained from different bounds. Indeed, they share the same confidence level $1 - p$, as shown in Figure 1 (bottom).

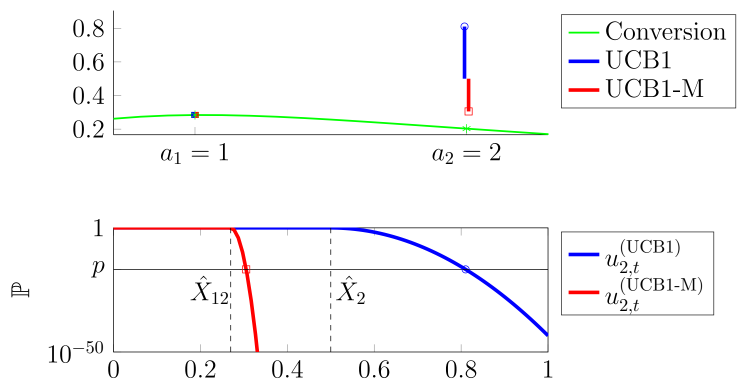

Figure 1: Example of empirical means not consistent with the monotonicity property and UCBs corresponding to UCB1 and UCB1-M. The top figure presents the real conversion rate function (green line) and two bars going from the estimated expected reward and the two bounds (blue and red lines). The bottom figure represents the dependence of the two bounds over arm a_2 (blue and red lines) with respect to the confidence level one wants to keep $[1\ 10^{-50}]$; p is the confidence level used to draw the top figure and the dashed lines are the empirical means of X_2 and X_{12}.

ALGORITHM 1: UCB1-M

Initialization

for $t \in \{1, \dots, K\}$ **do**

Play arm a_t and observe $x_{t,1}$

Loop

for $t \in \{K+1, \dots, N\}$ **do**

for $i \in \{1, \dots, K\}$ **do**

Compute:

$$u_{i,t}^{(\text{UCB1-M})} = \min_{j \in \{1,\dots,i\}} \left\{ \overline{x}_{ji,t} + \sqrt{\frac{4\log(t) + \log(i)}{2T_{ji}(t-1)}} \right\}$$

Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\dots,K\}} a_i u_{i,t}^{(\text{UCB1-M})}$ and observe $x_{i_t, T_{i_t}(t)}$

The algorithm corresponding to the previously derived UCB, namely *UCB1 with Monotonic arms* (UCB1-M), is presented in Algorithm 1. At first, the algorithm selects each arm once, to have at least one outcome realization coming from each arm. Subsequently, for each round t, it assigns for each arm a_i:

$$u_{i,t}^{(\text{UCB1-M})} = \min_{j \in \{1,\dots,i\}} \left\{ \overline{x}_{ji,t} + \sqrt{\frac{4\log(t) + \log(i)}{2T_{ji}(t-1)}} \right\},$$

where, we considered $p = t^{-4}$ and we selected the $j \in \{1, \dots, i\}$ minimizing $u_{ji,t}^{(\text{UCB1-M})}$. Finally, the algorithm selects for the next round t the arm a_{i_t} providing the maximum upper bound $a_{i_t} u_{i_t,t}^{(\text{UCB1-M})}$ over the expected reward $a_i \mu_i$.

By using the UCB1-M algorithm we are able to show that:

Theorem 1. *If policy UCB1-M is run over a stationary MAB setting with a monotonic set A, the expected regret after N rounds is at most:*

$$\overline{R}_N \leq \sum_{i | a_i \neq a_{i^*}} \frac{8a_i^2 \log(N)}{\Delta_i} + \sum_{i | a_i \neq a_{i^*}} \frac{2a_i^2 \log(K)}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \sum_{i=1}^{K} \Delta_i,$$

where $\Delta_i := a_{i^*}\mu_{i^*} - a_i \mu_i, \forall i \in \{1, \dots, K\}$.

The previous theorem guarantees that the proposed algorithm has, in the worst case, $O(\log(N))$ regret, as the UCB1 policy. Nevertheless, we show that, empirically, UCB1-M dramatically outperforms UCB1.

3.1.2. UCBV with Monotonic Arms (UCBV-M)

Similarly, by resorting to the bound presented Theorem 1 in [32], it is possible to derive an UCB that also considers the empirical variance $\overline{V}_{ji,t}$ of the variable

$X_{ji,t}$ by using its realization $\overline{v}_{ji,t}$, formally defined as:

$$\overline{V}_{ji,t} = \frac{\sum_{k=j}^{i} \sum_{n=1}^{T_k(t-1)} \left(X_{k,n} - \overline{X}_{ji,t}\right)^2}{T_{ji}(t-1)},$$
$$\overline{v}_{ji,t} = \frac{\sum_{k=j}^{i} \sum_{n=1}^{T_k(t-1)} \left(x_{k,n} - \overline{x}_{ji,t}\right)^2}{T_{ji}(t-1)},$$

respectively. The bound, from now on denoted as UCBV-M, holding with probability at least $1 - 3\left(\frac{p}{i}\right)^{\xi}$, with $p \in [0,1]$ is:

$$u_{ji,t}^{(\text{UCBV-M})} = \overline{x}_{ji,t} + \sqrt{\frac{2\overline{v}_{ji,t}\xi[\log(i) - \log(p)]}{T_{ji}(t-1)}} + \frac{3c\xi[\log(i) - \log(p)]}{T_{ji}(t-1)} > \mu_{ji,t},$$

where $\xi, c \in \mathbb{R}, \xi > 1, c \geq 1$; see [32] for details. Note that, if we choose $\xi > 1 - \frac{\log(3)}{\log(p)}$, the previous bound holds with probability at least $1 - \frac{p}{i}$, i.e., with the same confidence the UCB1-M holds.

The algorithm, based on the bound derived above and called *UCBV with Monotonic arms* (UCBV-M), is described in Algorithm 2. Similarly to UCB1-M, it chooses each arm once in the initial phase and, after that, it selects the next arm to play on the basis of the upper confidence bounds $u_{i,t}^{(\text{UCBV-M})} = u_{\bar{j}i,t}^{(\text{UCBV-M})}$, where $\bar{j}$ is chosen to minimize $u_{i,t}^{(\text{UCBV-M})}$ and $p = t^{-1}$. It is possible to show that:

Theorem 2. *If policy UCBV-M is run with $\xi = 1.2$ and $c = 1$ over a setting with a monotonic set A, the expected regret after N rounds is at most:*

$$\overline{R}_N \leq \frac{12}{5} \sum_{i|a_i \neq a_{i^*}} a_i^2 \left(\frac{\sigma_i^2}{\Delta_i} + \frac{32}{15}\right) \log(N) + \sum_{i|a_i \neq a_{i^*}} \Delta_i \left[1 + a_i^2 \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2}{\Delta_i}\right) \log(K)\right],$$

where $\sigma_i^2 := Var(X_{i,n}),\ \forall i \in \{1, \ldots, K\}, \forall\ n \in \{1, \ldots, T_i(N)\}$.

Even in this theorem the asymptotic behaviour is of order of $O(\log(N))$ as the one presented in [32] for the UCBV algorithm.

ALGORITHM 2: UCBV-M

Initialization

Input: ξ, c

for $t \in \{1, \dots, K\}$ **do**

Play arm a_t and observe $x_{t,1}$

Loop

for $t \in \{K+1, \dots, N\}$ **do**

for $i \in \{1, \dots, K\}$ **do**

Compute:

$$u_{i,t}^{(\text{UCBV-M})} = \min_{j \in \{1,\dots,i\}} \left\{ \overline{x}_{ji,t} + \sqrt{\frac{2\overline{v}_{ji,t}[\zeta \log(t) + \log(i)]}{T_{ji}(t-1)}} + \frac{3c[\zeta \log(t) + \log(i)]}{T_{ji}(t-1)} \right\}$$

Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\dots,K\}} a_i u_{ji,t}^{(\text{UCBV-M})}$ and observe $x_{i_t, T_{i_t}(t)}$

3.2. Exploiting the Low Conversion Rates Property

When it is *a priori* known that the conversion rates of all the arms are upper bounded by a value $\mu_{\max} \leq \frac{1}{2}$, i.e., $\mu_i \leq \mu_{\max}$ for every $i \in \{1, \dots, K\}$, it is possible to exploit probabilistic bounds that achieve better results than the one based on the Hoeffding's inequality [34]. More specifically, one of the approximations used in the derivation of the Hoeffding's inequality for the generic outcome X_i is:

$$\mathbb{P}(\overline{X}_{i,t} + \varepsilon \leq \mu_i) \leq e^{-T_i(t-1)D(\mu_i+\varepsilon||\mu_i)} \leq e^{-2T_i(t-1)\varepsilon^2}, \tag{2}$$

where $D(\mu_i+\varepsilon||\mu_i)$ is the Kullback-Leibler (KL) divergence between two Bernoulli variables with mean $\mu_i + \varepsilon$ and μ_i, respectively.

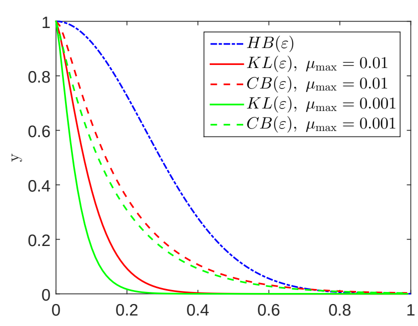

Figure 2: Example of bounds $y = e^{-x(\varepsilon)}$ obtained with different $x(\varepsilon)$: Hoeffding's Bound ($HB(\varepsilon)$), Kullback-Leiber divergence ($KL(\varepsilon)$) and Chernoff's Bound ($CB(\varepsilon)$).

As shown in Figure 2, the bound based on the KL divergence (solid lines) and the one on Hoeffding's inequality (dash-dotted line) diverge as $\mu_{\max}$ decreases. To reduce the gap, we consider the following result that is one of the formulations of the Chernoff's bound [33]:

Theorem 3 (Theorem 4 in [35], Lower tail). *Given a set of $T_i(t-1)$ independent and identically distributed random variables $\{X_{i,1}, \ldots, X_{i,T_i(t-1)}\}$ such that $X_{i,s} \sim Be(\mu_i)$, for any $\varepsilon > 0$ we have:*

$$\mathbb{P}(\overline{X}_{i,t} + \varepsilon \leq \mu_i) \leq e^{-\frac{T_i(t-1)\varepsilon^2}{2\mu_i}}.$$

Since μ_i is unknown, the above concentration inequality cannot be used in practice. On the other hand, under the assumption that $\mu_i \leq \mu_{\max}$, we can replace μ_i with $\mu_{\max}$, thus getting an upper confidence bound that is tighter than the Hoeffding's one and gets close to the one obtained by knowing the KL divergence (see dashed lines in Figure 2). To obtain an upper confidence bound over μ_i with confidence $1-p$, with $p \in [0, 1]$, we resort to Theorem 3 and get:

$$\mathbb{P}(\overline{X}_{i,t} + \varepsilon \leq \mu_i) \leq e^{-\frac{T_i(t-1)\varepsilon^2}{2\mu_i}} \leq e^{-\frac{T_i(t-1)\varepsilon^2}{2\mu_{\max}}} = p, \tag{3}$$

where the last inequality derives from the trivial fact that $\mu_{\max} \geq \mu_i$ for every $i \in \{1, \ldots, K\}$. Thus, with probability at least $1-p$ we have the following UCBs (from now on denoted as UCB-L):

$$u_{i,t}^{(\text{UCB-L})} := \overline{x}_{i,t} + \sqrt{-\frac{2\mu_{\max}\log(p)}{T_i(t-1)}} \geq \mu_i, \tag{4}$$

where the square root term is computed by considering the positive root of the second order equality in Equation (3). By comparing the two bounds provided by Hoeffding's and Chernoff's inequalities, it is possible to compute a sufficient condition that identifies when the former is tighter than the latter: when $\mu_{\max} > \frac{1}{2}$ the bound in Equation (3) is larger than the one in the right hand side of Equation (2). As a consequence, if we cannot guarantee low conversion probabilities, it is better to resort to the traditional Hoeffding's bound.

ALGORITHM 3: UCB-L

Initialization
Input: $\mu_{\max}$
for $t \in \{1, \ldots, K\}$ **do**
  Play arm a_t and observe $x_{t,1}$
Loop
for $t \in \{K+1, \ldots, N\}$ **do**
  for $i \in \{1, \ldots, K\}$ **do**
    Compute:

$$u_{i,t}^{(\text{UCB-L})} = \overline{x}_{i,t} + \sqrt{\frac{8\mu_{\max}\log(t)}{T_i(t-1)}}$$

  Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\ldots,K\}} a_i u_{i,t}^{(\text{UCB-L})}$ and observe $x_{i_t,T_{i_t}(t)}$

The proposed algorithm, namely *Upper Confidence Bound with Low conversion rates* (UCB-L) is presented in Algorithm 3, where we set $p = t^{-4}$ and we choose the next arm to be pulled by selecting the one having the maximum expected revenue. The execution is analogous to the one already described for UCB1-M and UCBV-M, where we have an initial round robin over all the arms and, after that, the choice of the arm to be played in the next round is based on the upper bound of the regret $a_i u_{i,t}^{(\text{UCB-L})}$.

In this case it is possible to show that:

Theorem 4. *If policy UCB-L is run over a stationary MAB setting with a set of arms A in which each arm $a_i \in A$ has outcome $X_{i,t}$ such that $\mathbb{E}[X_{i,t}] = \mu_i \leq \mu_{\max} \leq \frac{1}{2}$ for each $t \in \{1, \dots, N\}$, the expected regret after N rounds is at most:*

$$\overline{R}_N \leq \sum_{i|a_i \neq a_{i^*}} \frac{32\mu_{\max} a_i^2 \log(N)}{\Delta_i} + \left[1 + \frac{\pi^2}{6} + \zeta\left(\frac{10}{7}\right)\right] \sum_{i=1}^{K} \Delta_i,$$

where $\zeta(\cdot)$ is the Riemann zeta function.

Even by resorting by this newly designed bound the asymptotic order is $O(\log(N))$, thus we are assured to lose only a logarithmic amount of reward in the learning process.

3.3. Exploiting both Properties

Here, we show how to combine both the monotonic and the low conversion rates properties into a single algorithm. The resulting algorithm, named UCB-LM, consists of computing for each arm a_i the minimum upper confidence bound among the ones built using $\overline{X}_{ji,t}$, with $j \in \{1, \dots, i\}$, but, differently from UCB1-M, the UCBs are built exploiting the Chernoff's inequality and the assumption over the maximum conversion rate as it happens in UCB-L.

ALGORITHM 4: UCB-LM

Initialization
Input: $\mu_{\max}$
for $t \in \{1, \dots, K\}$ **do**
  Play arm a_t and observe $x_{t,1}$
Loop
for $t \in \{K+1, \dots, N\}$ **do**
  for $i \in \{1, \dots, K\}$ **do**
    Compute:

$$u_{i,t}^{(\text{UCB-LM})} = \min_{j \in \{1,\dots,i\}} \left\{ \overline{x}_{ji,t} + \sqrt{\frac{2\mu_{\max}[4\log(t) + \log(i)]}{T_{ji}(t-1)}} \right\}$$

  Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\dots,K\}} a_i u_{i,t}^{(\text{UCB-LM})}$ and observe $x_{i_t, T_{i_t}(t)}$

The resulting algorithm (in which we choose $p = t^{-4}$) is summarized in Algorithm 4. Also in this case, we can state the following result:

Theorem 5. *If policy UCB-LM is run over a stationary MAB setting with a monotonic set A in which each arm $a_i \in A$ has outcome $X_{i,t}$ such that $\mathbb{E}[X_{i,t}] = \mu_i \leq \mu_{\max} \leq \frac{1}{2}$ for each t, the expected regret after N rounds is at most:*

$$\overline{R}_N \leq \sum_{i | a_i \neq a_{i^*}} \frac{32 \mu_{\max} a_i^2 \log(N)}{\Delta_i} + \sum_{i | a_i \neq a_{i^*}} \frac{8 \mu_{\max} a_i^2 \log(K)}{\Delta_i} + \left[1 + \frac{\pi^2}{6} + \zeta\left(\frac{10}{7}\right)\right] \sum_{i=1}^{K} \Delta_i,$$

where $\zeta(\cdot)$ is the Riemann zeta function.

This bound presents the same characteristics of the one derived for UCB-L, e.g., $O(\log(N))$ regret and constant dependent from $\mu_{\max}$. The experimental results, presented in Section 5, provide empirical evidence that the introduction of the monotonicity assumption is improving the performance of UCB-LM even when we use the Chernoff bound to design MAB policies.

4. Exploiting the Monotonicity Property in Nonstationary Environment

Since in a nonstationary environment $\mathcal{S}^{(B)}$ the outcome expected values $\mu_{i,\phi}$ might change as a new phase starts, we employ, similarly to [27], a Sliding Window (SW) approach for UCB-like algorithms. This approach takes decisions relying on what happened during the last τ rounds and, therefore, is capable of forgetting information coming from previous phases. At the same time, we integrate the information coming from the monotonicity property to speed up the learning process. The choice of the SW length τ for such a setting is out of the scope of this paper (more information can be found in [27]).

In what follows, we use the estimator for the outcome average value μ_i over the last $\min\{\tau, t\}$ rounds $\overline{X}_{i,t,\tau}$ and its realization $\overline{x}_{i,t,\tau}$, which are defined as:

$$\overline{X}_{i,t,\tau} := \frac{1}{T_i(t-1,\tau)} \sum_{s=T_i(\max\{t-\tau,1\})}^{T_i(t-1)} X_{i,s},$$

$$\overline{x}_{i,T_i(t-1,\tau),\tau} := \frac{1}{T_i(t-1,\tau)} \sum_{s=T_i(\max\{t-\tau,1\})}^{T_i(t-1)} x_{i,s},$$

where $T_i(t,\tau) = T_i(t) - T_i(\max\{t-\tau+1, 1\})$ is the number of rounds the arm a_i has been selected in the last $\min\{\tau, t\}$ ones. Similarly to what has been considered for the UCB1-M algorithm, for each $1 \leq j \leq i$, let $\overline{X}_{ji,t,\tau}$ be the following linear combination of the random variables $\overline{X}_j, \dots, \overline{X}_i$ and $\overline{x}_{ji,t,\tau}$ its realization, defined as:

$$\overline{X}_{ji,t,\tau} := \frac{\sum_{k=j}^{i} T_k(t-1,\tau)\overline{X}_{k,t,\tau}}{T_{ji}(t-1,\tau)},$$
$$\overline{x}_{ji,t,\tau} := \frac{\sum_{k=j}^{i} T_k(t-1,\tau)\overline{x}_{k,t,\tau}}{T_{ji}(t-1,\tau)},$$

where $T_{ji}(t,\tau) = \sum_{k=j}^{i} T_k(t,\tau)$ is the number of rounds one of the arms in $\{a_j, \dots, a_i\}$ has been selected in the last $\min\{\tau, t\}$ ones. Given the monotonicity property and assuming to have samples to compute $\overline{x}_{ji,t,\tau}$ coming from the same phase Φ_ϕ we have:

$$\mu_{ji,\phi} = \mathbb{E}\left[\overline{X}_{ji,t,\tau}\right] \geq \mu_{i,\phi}.$$

Consider the following:

Theorem 6 (Corollary 21 in [27]). *Given a sequence $\{X_1, \dots, X_t\}$ of $t \in \mathbb{N}$ random variables with support $\Omega \subseteq [0,1]$ with expectation $\mu_h := \mathbb{E}[X_h]$ and a sequence $\{\epsilon_1, \dots, \epsilon_t\}$ a previsible sequence of Bernoulli random variables. For all $\tau \in \mathbb{N}$ and $\eta > 0$ it holds:*

$$\mathbb{P}\left(\frac{\sum_{h=\min\{t-\tau+1,1\}}^{t}(X_h - \mu_h)\epsilon_h}{\sum_{h=\min\{t-\tau+1,1\}}^{t}\epsilon_h}\right) \leq \left\lceil \frac{\log(\min\{t,\tau\})}{\log(1+\eta)} \right\rceil \exp\left\{-2\delta^2\left(1 - \frac{\eta^2}{16}\right)\right\}.$$

If we apply the previous result to the random variable $\overline{X}_{ji,t,\tau}$ and $\eta = 4\sqrt{1-\frac{2}{\xi}}$, with probability at least $1 - \frac{p}{i}$, with $p \in [0,1]$, we have the following UCBs (from now on denoted as SW-UCB-M):

$$u_{ji,t}^{(\text{SW-UCB-M})} = \overline{x}_{ji,t,\tau} + \sqrt{\frac{\xi[\log(i) - \log(p)]}{T_{ji}(t-1,\tau)}} > \mu_{ji,\phi} \geq \mu_{i,\phi}, \tag{5}$$

where $\xi \in \mathbb{R}^+$ is a parameter used in the bound in [27].[3] Even in this case, we select $u_{i,t}^{(\text{SW-UCB-M})}$ as the tightest bound for $1 \leq j \leq i$, which holds with at least probability $1 - p$, to decide which arm to pull next.

[3]Here we assume that all the variables used to obtain $\overline{X}_{ji,t,\tau}$ are coming from a single phase Φ_ϕ.

ALGORITHM 5: SW-UCB-M

Initialization

for $t \in \{1, \dots, K\}$ **do**

Play arm a_i and observe $x_{t,1}$

Loop

for $t \in \{K+1, \dots, N\}$ **do**

for $i \in \{1, \dots, K\}$ **do**

Compute:

$$u_{i,t}^{(\text{SW-UCB-M})} = \min_{j \in \{1,\dots,i\}} \left\{ \overline{x}_{ji,t,\tau} + \sqrt{\frac{\xi \left(4 \log(\min\{t,\tau\}) + \log(i)\right)}{T_{ji}(t-1,\tau)}} \right\}$$

Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\dots,K\}} a_i u_{i,t}^{(\text{SW-UCB-M})}$ and observe $x_{i_t, T_{i_t}(t)}$

The pseudocode of the algorithm employing the aforementioned a bound with $p = (\min\{t,\tau\})^{-4}$ is presented in Algorithm 5 and presents characteristics similar to the bounds we propose in the previous section. Focusing on the SW-UCB-M algorithm, we can show that:

Theorem 7. *If policy SW-UCB-M is run over a nonstationary MAB setting $\mathcal{S}^{(B)}$, for any $\tau \in \mathbb{N}$ and $\xi > \frac{1}{2}$, the expected regret after N rounds is at most:*

$$\overline{R}_N \leq \sum_{i=1}^{K} \left[\frac{N}{\tau} \frac{4 a_i^2 \xi [\log(i) + \log(\tau)]}{\Delta_i} + a_i \Upsilon_N \tau + \frac{2N}{\tau} \left[\frac{\log(\tau)}{\log\left(1 + 4\sqrt{1 - \frac{1}{2\xi}}\right)} \right] \right],$$

where Υ_N is the number of breakpoints before N and

$$\Delta_i := \min_{\phi \in \{1,\dots,\Upsilon_N\}} \left(a_{i^*_\phi} \mu_{i^*_\phi,\phi} - a_i \mu_{i,\phi} \right) \mathbb{1}\{i \neq i^*_\phi\} \quad \forall i \in \{1,\dots,K\},$$

denotes the minimum, over all the phases Φ_ϕ in which the arm a_i is not optimal, of the difference of the expected reward $a_{i^*_\phi} \mu_{i^*_\phi,\phi}$ of the best arm $a_{i^*_\phi}$ and the expected reward $a_i \mu_{i,\phi}$ of the arm a_i.

5. Experimental Analysis in Stationary Environments

We provide a thorough experimental evaluation of our algorithms in stationary environments, comparing them with the corresponding algorithms that do not exploit the two properties of the pricing problem we study.

5.1. Experimental Setting and Performance Indices

We evaluate our algorithms on a wide spectrum of configurations of pricing settings characterized by a different number of arms in A, by different pdfs $\mathcal{S}$, and by a different $\mu_{\max}$. In particular, we use a number of arms

$K \in \{5, 9, 17, 33\}$ evenly spaced over the interval $[1, 17]$, whose values can be interpreted as euros. We use a minimum of 5 arms, since a smaller number would provide an excessively coarse discretization of the demand curve, leading to an important loss in terms of profit. Furthermore, we use 9, 17, 33 arms such that we iteratively halve the distance between each couple of consecutive arms, thus making the discretization more accurate.

In order to use a realistic experimental setting, we estimate the demand curves by gathering historical data coming from past transactions of an European Online Travel Agency. These curves are monotonically decreasing with respect to the gross margin. More precisely, we estimate the conversion probabilities corresponding to the applied gross margins and we fit the $(1-)$CDF of a Gaussian distribution (minimizing the mean squared error). The Gaussian distribution provides a probability distribution over the acceptance threshold $\mathcal{S}$ of the customers. In doing that, we only use the data related to the purchases of tickets when the availability of the supply was sufficiently large and, thus, it did not affect the customers' behavior. We focus on two classes of probability distributions, called $\mathcal{S}_L$ and $\mathcal{S}_H$. In $\mathcal{S}_H$, the arm maximizing the profit is among the arms with largest gross margins, and, in $\mathcal{S}_L$, the arm maximizing the profit is among the arms with smallest gross margins. Configurations $\mathcal{S}_L$ and $\mathcal{S}_H$ represent the two extreme and most significant cases for the class of algorithms that make the assumption of optimism against uncertainty. More precisely, $\mathcal{S}_H$ is an easy configuration independently from the number of the arms we choose for discretization since any algorithm based on the assumption of optimism against uncertainty can discard most of the arm with a few pulls and, therefore, identify the best arm with a little exploration cost. Instead, $\mathcal{S}_L$ is a challenging configuration, since the identification of the best arm requires a large exploration cost.

For each class $\mathcal{S}_L$ and $\mathcal{S}_H$, we generate 5 probability distributions distinguishing for $\mu_{\max}$. We use the values of $\mu_{\max}$ in $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$, corresponding, in the case of the reselling of flight tickets, to different routes and markets. Let us observe that such a range includes the values of $\mu_{\max}$ of many scenarios different from the one we study, allowing us to provide an experimental evaluation of our algorithms also in other scenarios. More precisely, according to [36], $\mu_{\max} = 10^{-1}$ corresponds to Bing, Google, Yahoo!; $\mu_{\max} = 10^{-2}$ corresponds to Facebook, Pinterest, Twitter; $\mu_{\max} = 10^{-3}$ corresponds to LinkedIn; $\mu_{\max} = 10^{-4}$ corresponds to StumbleUpon.

Summarily, the threshold pdfs $\mathcal{S}$ are as follows:[4]

- $\mathcal{S}_H \sim \mathcal{N}(20, 6)$, representing a situation where $a_{i^*} \geq 15$, i.e., the optimal gross margin is among the highest values in $[1, 17]$ and for every i we have $\mu_i \in [0.68\mu_{\max}, \mu_{\max}]$, and
- $\mathcal{S}_L \sim \mathcal{N}(3, 5)$, representing a situation where $a_{i^*} \leq 5$, i.e., the optimal

[4] Here, we denote with $\mathcal{N}(\mu, \sigma)$ the normal distribution with mean μ and standard deviation σ.

gross margin is among the lowest values in $[1, 17]$ and for every i we have $\mu_i \in [0.0025\mu_{\max}, 0.66\mu_{\max}]$.

For each combination of $(K, \mathcal{S}, \mu_{\max})$, we average over 100 independent trials of length $N = 10^7$ rounds and in each round the threshold s is independently drawn from $\mathcal{S}$.

We compare our algorithms UCB1-M, UCB-L, and UCB-LM with the corresponding frequentist algorithms that do not exploit the two properties of the pricing problem we study: UCB1, UCBV, and UCBV-M (for the UCBV and UCBV-M algorithms, the parameters we use are $c = \xi = 1$). In our evaluation, we use the following performance indices, for each $t \leq N$:

$$\begin{aligned} R_{\%}(t) &= \frac{\bar{R}_t(\mathfrak{U})}{\bar{R}_t(\text{UCB1})} \\ \Delta P(t) &= \sum_{t'=1}^{t} \mathbb{E}\left[V_{i_{(\mathfrak{U},t')}}\right] - \sum_{t'=1}^{t} \mathbb{E}\left[V_{i_{(\text{UCB1},t')}}\right] \\ \Delta P_{\%}(t) &= \frac{\Delta P(t)}{\sum_{t'=1}^{t} \mathbb{E}[V_{i_{(\text{UCB1},t')}}]} \end{aligned}$$

where $\mathfrak{U}$ is a generic policy, $i_{(\mathfrak{U},t)}$ is the index chosen by policy $\mathfrak{U}$ at time t. $R_{\%}(t)$ is defined as the ratio between the total regret of policy $\mathfrak{U}$ after t rounds and the regret of UCB1 that we use here as the baseline—a value of $R_{\%}$ lower than 1 means that $\mathfrak{U}$ outperforms UCB1 and the lower the value the greater the improvement—; $\Delta P(t)$ is the difference between the cumulative expected reward of policy $\mathfrak{U}$ and the one obtained with UCB1; $\Delta P_{\%}$ is defined as the ratio between $\Delta P(t)$ and the cumulative expected reward obtained with UCB1. A value of ΔP (and $\Delta P_{\%}$) greater than 0 means that $\mathfrak{U}$ improves the profit with respect to UCB1 and the higher the value the greater the improvement.

5.2. Regret Analysis

The average $R_{\%}(N)$ and the 95% confidence intervals are reported in Table 1 (the results of UCB-L and UCB-LM are omitted for $\mu_{\max} = 1$, their bound being theoretically worse than the one of UCB1). We omit the evaluation of $R_{\%}(t)$ for $t < N$, since we provide in the next section a detailed discussion about how the profit provided by the algorithms changes as t changes, and we believe this latter evaluation is more significant in practice than the evaluation of the dependency of the regret on time.

We initially focus on the results obtained with $\mathcal{S}_L$. Here, UCBV-M outperforms all the other algorithms, with $R_{\%}(N)$ decreasing from 0.55 to 0.02 of the UCB1 regret. Furthermore, we observe that all the algorithms in the table outperform UCB1. While UCB1-M performs better than UCB-L only in some specific settings, UCB-LM outperforms both UCB1-M and UCB-L in all the configurations. Furthermore, UCB-LM performs usually worse than UCBV, except for very low values of $\mu_{\max}$ and many arms. These results strengthen the evidence that the use of the Chernoff's bound is effective when $\mu_{\max} \ll 1$. Instead, UCBV-M always outperforms UCBV reducing the regret of UCBV by a ratio up to 2/3. We observe that the (relative) performance of the algorithms exploiting the monotonicity increases as the number of arms increases. This is because these algorithms better exploit the correlation among the arms. Finally, we observe that the best improvement (in terms of reduction of the regret) of our algorithms with respect to the performance of UCB1 is for $\mu_{\max} = 10^{-1}$. This is because when $\mu_{\max} = 1$ all the algorithms converge to the best arm before $N = 10^7$ rounds, minimizing the differences in terms of regret among them; when $\mu_{\max} \in \{10^{-1}, 10^{-2}\}$ our algorithms converge to the best arm before 10^7 rounds, while UCB1 does not, thus maximizing the differences in terms of regret among the algorithms; when $\mu_{\max} \in \{10^{-3}, 10^{-4}\}$ no algorithm converges to the best arm by 10^7 rounds, but some algorithms select the best arm more frequently than others.

Table 1: Results concerning $R_{\%}(N)$ (averaged values over 100 runs, $\pm$ 95% confidence intervals). The best results for each configuration are in boldface.

		S_L				
$\mu_{\max}$	$\lvert A\rvert$	UCB1-M	UCBL	UCB-LM	UCBV	UCBV-M
1	5	0.81 ± 0.01	——	——	0.22 ± 0.00	$\mathbf{0.20 \pm 0.00}$
	9	0.72 ± 0.01	——	——	0.24 ± 0.00	$\mathbf{0.19 \pm 0.00}$
	17	0.67 ± 0.01	——	——	0.26 ± 0.00	$\mathbf{0.20 \pm 0.00}$
	33	0.61 ± 0.01	——	——	0.31 ± 0.01	$\mathbf{0.23 \pm 0.01}$
10^{-1}	5	0.80 ± 0.00	0.42 ± 0.00	0.34 ± 0.00	0.03 ± 0.00	$\mathbf{0.02 \pm 0.00}$
	9	0.66 ± 0.00	0.45 ± 0.00	0.30 ± 0.00	$\mathbf{0.03 \pm 0.00}$	$\mathbf{0.03 \pm 0.00}$
	17	0.50 ± 0.00	0.50 ± 0.00	0.27 ± 0.00	0.05 ± 0.00	$\mathbf{0.04 \pm 0.00}$
	33	0.32 ± 0.00	0.54 ± 0.00	0.20 ± 0.00	0.06 ± 0.00	$\mathbf{0.04 \pm 0.00}$
10^{-2}	5	0.87 ± 0.00	0.30 ± 0.00	0.24 ± 0.00	$\mathbf{0.02 \pm 0.00}$	$\mathbf{0.02 \pm 0.00}$
	9	0.78 ± 0.00	0.49 ± 0.00	0.31 ± 0.00	0.05 ± 0.00	$\mathbf{0.04 \pm 0.00}$
	17	0.73 ± 0.00	0.65 ± 0.00	0.30 ± 0.00	0.11 ± 0.00	$\mathbf{0.07 \pm 0.00}$
	33	0.70 ± 0.00	0.77 ± 0.00	0.28 ± 0.00	0.17 ± 0.00	$\mathbf{0.08 \pm 0.00}$
10^{-3}	5	0.91 ± 0.00	0.83 ± 0.00	0.71 ± 0.00	0.17 ± 0.00	$\mathbf{0.15 \pm 0.00}$
	9	0.88 ± 0.00	0.88 ± 0.00	0.64 ± 0.00	0.33 ± 0.00	$\mathbf{0.22 \pm 0.00}$
	17	0.86 ± 0.00	0.92 ± 0.00	0.59 ± 0.00	0.47 ± 0.00	$\mathbf{0.22 \pm 0.00}$
	33	0.85 ± 0.00	0.94 ± 0.00	0.58 ± 0.00	0.60 ± 0.00	$\mathbf{0.22 \pm 0.00}$
10^{-4}	5	0.92 ± 0.00	0.96 ± 0.00	0.86 ± 0.00	0.67 ± 0.01	$\mathbf{0.55 \pm 0.01}$
	9	0.89 ± 0.00	0.97 ± 0.00	0.81 ± 0.00	0.73 ± 0.00	$\mathbf{0.50 \pm 0.01}$
	17	0.87 ± 0.00	0.98 ± 0.00	0.78 ± 0.00	0.77 ± 0.00	$\mathbf{0.48 \pm 0.01}$
	33	0.86 ± 0.00	0.98 ± 0.00	0.77 ± 0.00	0.80 ± 0.00	$\mathbf{0.48 \pm 0.01}$

		S_H				
$\mu_{\max}$	$\lvert A\rvert$	UCB1-M	UCBL	UCB-LM	UCBV	UCBV-M
1	5	1.01 ± 0.02	——	——	$\mathbf{0.20 \pm 0.01}$	$\mathbf{0.21 \pm 0.01}$
	9	1.01 ± 0.03	——	——	$\mathbf{0.28 \pm 0.01}$	0.31 ± 0.01
	17	1.02 ± 0.02	——	——	$\mathbf{0.45 \pm 0.02}$	0.50 ± 0.02
	33	1.02 ± 0.01	——	——	$\mathbf{0.37 \pm 0.01}$	0.42 ± 0.01
10^{-1}	5	1.03 ± 0.02	0.60 ± 0.02	0.60 ± 0.02	$\mathbf{0.23 \pm 0.01}$	$\mathbf{0.24 \pm 0.01}$
	9	0.98 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	$\mathbf{0.22 \pm 0.01}$	$\mathbf{0.23 \pm 0.01}$
	17	0.86 ± 0.01	0.65 ± 0.01	0.59 ± 0.01	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.29 \pm 0.01}$
	33	0.67 ± 0.01	0.69 ± 0.01	0.54 ± 0.01	0.42 ± 0.01	$\mathbf{0.36 \pm 0.01}$
10^{-2}	5	0.93 ± 0.00	0.30 ± 0.01	0.29 ± 0.01	$\mathbf{0.21 \pm 0.01}$	$\mathbf{0.22 \pm 0.01}$
	9	0.85 ± 0.00	0.38 ± 0.01	0.35 ± 0.01	$\mathbf{0.25 \pm 0.01}$	$\mathbf{0.25 \pm 0.01}$
	17	0.75 ± 0.00	0.37 ± 0.00	0.28 ± 0.00	0.29 ± 0.01	$\mathbf{0.22 \pm 0.01}$
	33	0.67 ± 0.00	0.42 ± 0.00	0.25 ± 0.00	0.37 ± 0.00	$\mathbf{0.21 \pm 0.00}$
10^{-3}	5	1.26 ± 0.00	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.30 \pm 0.01}$	0.33 ± 0.01	$\mathbf{0.32 \pm 0.01}$
	9	1.28 ± 0.00	0.44 ± 0.01	$\mathbf{0.36 \pm 0.01}$	0.46 ± 0.01	$\mathbf{0.37 \pm 0.01}$
	17	1.30 ± 0.00	0.49 ± 0.01	$\mathbf{0.34 \pm 0.01}$	0.58 ± 0.01	$\mathbf{0.34 \pm 0.01}$
	33	1.30 ± 0.00	0.57 ± 0.00	$\mathbf{0.35 \pm 0.01}$	0.74 ± 0.01	$\mathbf{0.35 \pm 0.01}$
10^{-3}	5	1.46 ± 0.00	$\mathbf{0.55 \pm 0.01}$	$\mathbf{0.54 \pm 0.01}$	0.89 ± 0.02	0.78 ± 0.02
	9	1.51 ± 0.00	0.68 ± 0.01	$\mathbf{0.63 \pm 0.01}$	1.03 ± 0.01	0.83 ± 0.02
	17	1.57 ± 0.00	0.74 ± 0.01	$\mathbf{0.64 \pm 0.01}$	1.20 ± 0.01	0.86 ± 0.02
	33	1.59 ± 0.00	0.79 ± 0.01	$\mathbf{0.66 \pm 0.01}$	1.33 ± 0.01	0.86 ± 0.02

Now, we focus on the results obtained with $\mathcal{S}_H$. Here, there is no algorithm that always outperforms the others. We observe that, for large values of $\mu_{\max}$, UCBV is the best algorithm, for intermediate values, UCBV-M outperforms the others, and for small values, UCB-LM is the best. The best algorithm presents $R_{\%}(N)$ in the range between 0.20 and 0.66. In details, UCB1 performs better than UCB1-M for some cases and, surprisingly, better than UCBV when $\mu_{\max}$ is very small. We observe that UCB-L, UCB-LM, and UCBV-M always perform better than UCB1. In some configurations UCB-LM improves over UCBV, halving the UCBV regret, e.g., in the configuration with $K = 33$ arms and $\mu_{\max} = 10^{-4}$, providing a significant improvement over UCBV performance. Differently from the case with $\mathcal{S}_L$, with $\mathcal{S}_H$ the relative improvement of algorithms exploiting the monotonicity does not increase as the number of arms increases. The same holds for the UCBV algorithm, which does not exploit any assumption, suggesting that the performance of the UCB1 algorithm improves as the number of arms increases in the $\mathcal{S}_H$ setting. This is probably due to the fact that UCB1 excludes many arms easily in the case the optimal values of the expected reward are realized on high arms, e.g., if $\mu_i a_i > a_j$ it will not play arms lower or equal to a_j. Thus, UCB1 is effectively working on a smaller set of arms than A, and this leads to low regret even for this policy.

To summarize, our algorithms, specifically UCBV-M and UCB-LM, provide a significant improvement in terms of regret with respect to the algorithms available in the state of the art.

5.3. Profit Analysis

The average $\Delta P_{\%}(t)$ and $\Delta P(t)$ for $t \in \{1, \ldots, 10^7\}$ obtained with UCB-LM and UCBV-M (with respect to the results obtained with UCB1 and 5 arms) are reported in Figure 3 and Figure 4, respectively.[5] More detailed results about all the algorithms can be found in Appendix C.

Initially, we focus on the results obtained with $\mu_{\max} \in \{1, 10^{-1}\}$ and $\mathcal{S}_L$. The value of $\Delta P_{\%}(t)$ dramatically changes during the time horizon. It reaches a maximum around $t = 10^4$ for $\mu_{\max} = 1$ and $t = 10^5$ for $\mu_{\max} = 10^{-1}$ and then it decreases approaching the value of zero at $t = 10^7$. The improvement is significant, the maximum of $\Delta P_{\%}(t)$ being about 2.2 for UCBV-M (i.e., the profit of UCB1 is more than tripled) and about 1.3 for UCB-LM (i.e., the profit is more than doubled). In the case of $\mathcal{S}_H$, the value of $\Delta P_{\%}(t)$ initially reaches a minimum and subsequently a maximum, and finally approaches zero as t goes to 10^7. In this case, the improvement is less significant than the one we have in the case of $\mathcal{S}_L$ and $\Delta P_{\%}(t)$ is about 0.003 when $\mu_{\max} = 1$ and 0.033 when $\mu_{\max} = 10^{-1}$, meaning an improvement of 0.3% and 3.3% over the UCB1 profit, respectively.

[5] Results for UCB-LM in the case $\mu_{\max} = 1$ are not reported since this algorithm requires $\mu_{\max} < \frac{1}{2}$ to be effective. Moreover, the results for $\Delta P(t)$ for $\mathcal{S}_H$ are not reported since they are less significant.

Now we focus on the results obtained with $\mu_{\max} = 10^{-2}$ and $\mathcal{S}_L$. The maximum of $\Delta P_{\%}(t)$ is reached in the range between $t = 10^6$ and $t = 10^7$ (that is, very close to the termination of the time horizon). The improvement is very significant, $\Delta P_{\%}(t)$ achieving values larger than 3.3. The behavior for $\mathcal{S}_H$ is analogous with respect to the one with larger $\mu_{\max}$. Here, we can observe that the minimum is achieved for a larger t than in the setting with $\mu_{\max} \in \{1, 10^{-1}\}$. The maximum of $\Delta P_{\%}(t)$ is about 0.04. Finally, we focus on the results obtained with $\mu_{\max} \in \{10^{-3}, 10^{-4}\}$ and $\mathcal{S}_L$. The $\Delta P_{\%}(t)$ trend suggests that its maximum might be beyond 10^7 rounds. Nevertheless, the improvement is very significant: $\Delta P_{\%}(t)$ is larger than 3 for $\mu_{\max} = 10^{-3}$ and almost 2 for $\mu_{\max} = 10^{-4}$. As for smaller values of $\mu_{\max}$, the improvements with $\mathcal{S}_H$ are less significant. Nevertheless, UCB-LM presents a maximum of $\Delta P_{\%}(t)$ that is almost 0.08 even with $\mu_{\max} = 10^{-4}$.

Furthermore, we observe how the performance of the algorithms varies as the number of arms varies in the two different pdfs. With $\mathcal{S}_L$ the best improvement is achieved when the number of arms is 33 with $\mu_{\max} \leq 10^{-1}$ and 5 otherwise. Instead, with $\mathcal{S}_H$, the best improvement is achieved, in the most cases, when using 33 arms.

To summarize, our algorithms, specifically UCBV-M and UCB-LM, provide a significant improvement in terms of relative profit especially in the early stages of the learning process.

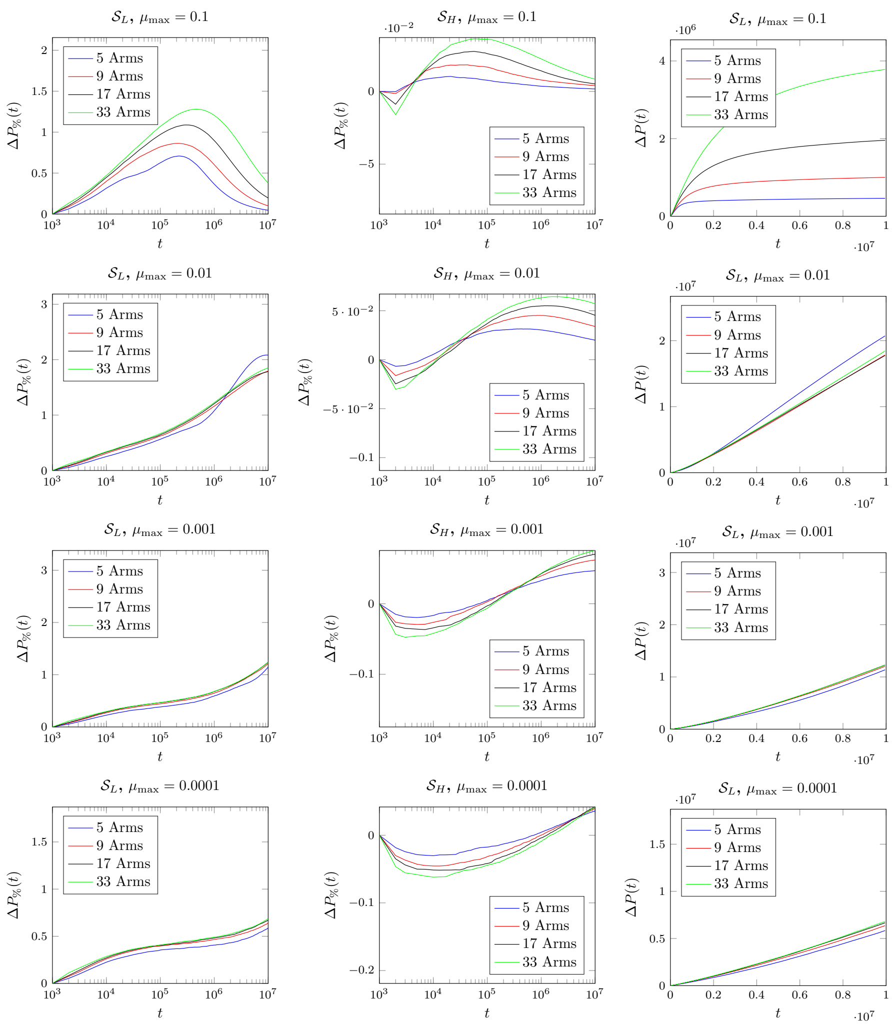

Figure 3: $\Delta P_{\%}(t)$ (first two columns) and $\Delta P(t)$ (third column) obtained with UCB-LM with different configurations.

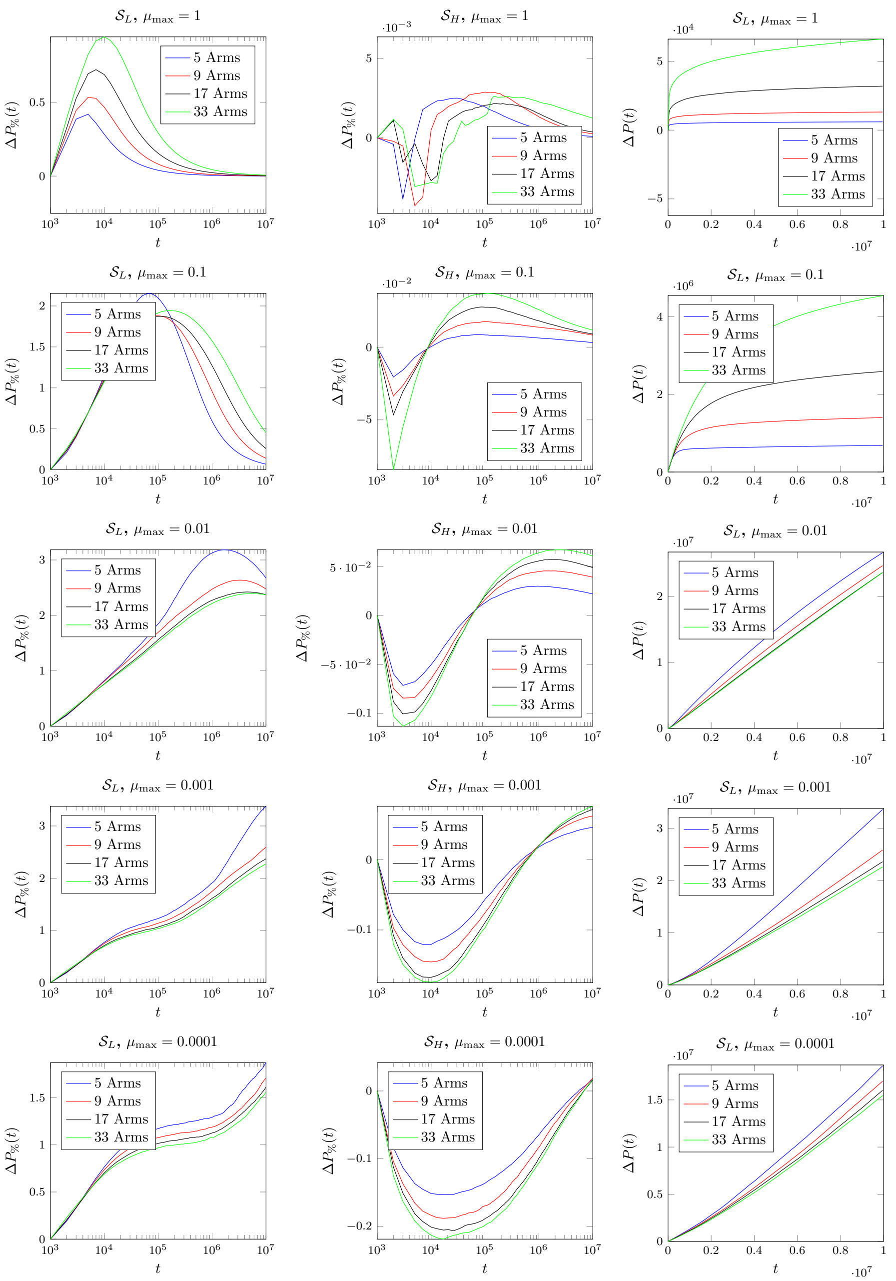

Figure 4: $\Delta P_{\%}(t)$ (first two columns) and $\Delta P(t)$ (third column) obtained with UCBV-M with different configurations.

5.4. Sensitivity Analysis

In this set of experiments, we evaluate the sensitivity of our UCB-L and UCB-ML algorithms with respect to the maximum conversion probability parameter $\mu_{\max}$. This study has been carried out by mis-specifying the $\mu_{\max}$ parameter in the algorithms for the same configurations we described in Section 5.1. More specifically, we evaluate our algorithms for each pair composed of the actual $\mu_{\max}$ and of the $\mu_{\max}$ (mis-)specified in the algorithms. The actual values of $\mu_{\max}$ we use are those used in Section 5.1. The values of $\mu_{\max}$ (mis-)specified in the algorithms are $\bar{\mu}_{\max} \in \{1, 10^{-1}, \dots, 10^{-6}\}$. The performance index is the average $R_{\%}(T)$. The results are obtained by averaging over 100 independent runs of the algorithms.

In Figure 5, the results for the UCB-L algorithm in the configuration with $K = 33$ arms are presented, with different threshold distributions ($\mathcal{S}_L$ and $\mathcal{S}_H$). In the figure, the filled circles denote the experiments in which we correctly specify the maximum conversion rate parameter ($\mu_{\max} = \bar{\mu}_{\max}$). In these cases, the UCB-L is always performing better than UCB1 ($R_{\%}(N) < 1$), except for the case $\mu_{\max} = 1$, in which we know from the theoretical analysis that UCB1 is strictly better than UCB-L. In the figure, it can be observed that the mis-specification of the parameter of $\bar{\mu}_{\max} = \frac{\mu_{\max}}{10}$ provides an improvement in the $R_{\%}(N)$ with respect to the one provided by a correctly specified one. This is because the Chernoff bound, used in the UCB-L algorithm, might not be tight in this specific setting, thus smaller upper bounds might still provide enough confidence on the reward estimates and, at the same time, reduce the regret. Conversely, increasing the value of the parameter with respect to the correct one ($\bar{\mu}_{\max} = 10\mu_{\max}$) the performance of the algorithm worsens, since we are using larger bounds, still providing small improvements over the UCB1 algorithm ones. The use of values $\bar{\mu}_{\max} < \frac{\mu_{\max}}{10}$ in the configuration $\mu_{\max} = 1$ for $\mathcal{S}_L$ and for $\mu_{\max} = 1$, $\mu_{\max} = 0.1$ for $\mathcal{S}_H$ provides results which are even worse than the one obtained with UCB1 ($R_{\%}(N) > 1$). This suggests that the proposed algorithm provides better results than UCB1 as long as the mis-specification stays within one order of magnitude from the real value of the maximum conversion rate.

The same considerations can be drawn in the case we use the UCB-LM algorithm instead of the UCB-L one. In Figure 6, we have a regret behaviour which is similar to the one of UCB-L we analyzed in Figure 5. This suggests that the sensitivity of the method does not change in the case the monotonicity property is taken into account. We do not report the figure corresponding to other configurations ($K \in \{5, 9, 17\}$) since they do not change the conclusion we drew in this section.

6. Experimental Analysis in Nonstationary Environments

We experimentally evaluate the performance of our techniques in an abruptly changing environment, that, as aforementioned, is one of the most common nonstationary settings in e-commerce, e.g., it models the entrance of a new player in the market. We compare the SW-UCB-M algorithm with UCBV-M, as representatives of algorithms exploiting the monotonicity assumption,

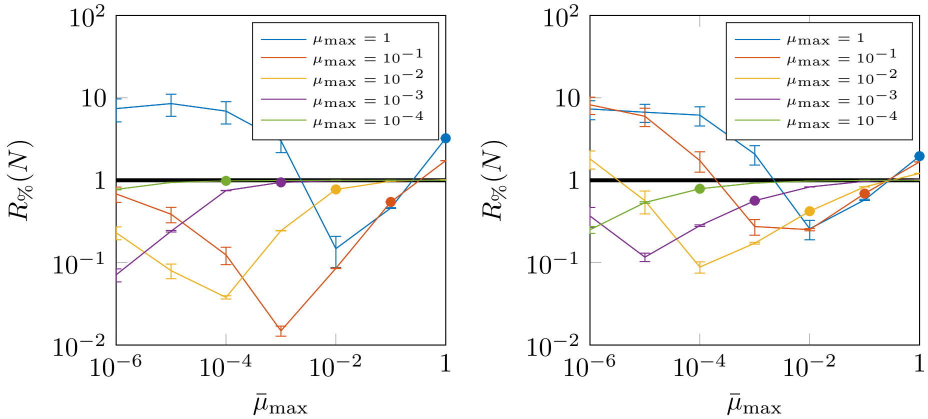

Figure 5: $R_{\%}(N)$ obtained with UCB-L with different configurations: $K = 33$ arms, $\mathcal{S}_L$ (left) and $\mathcal{S}_H$ (right). The error bars represents the 95% confidence intervals for the expected values. The black like represents the performance of the UCB1 algorithm.

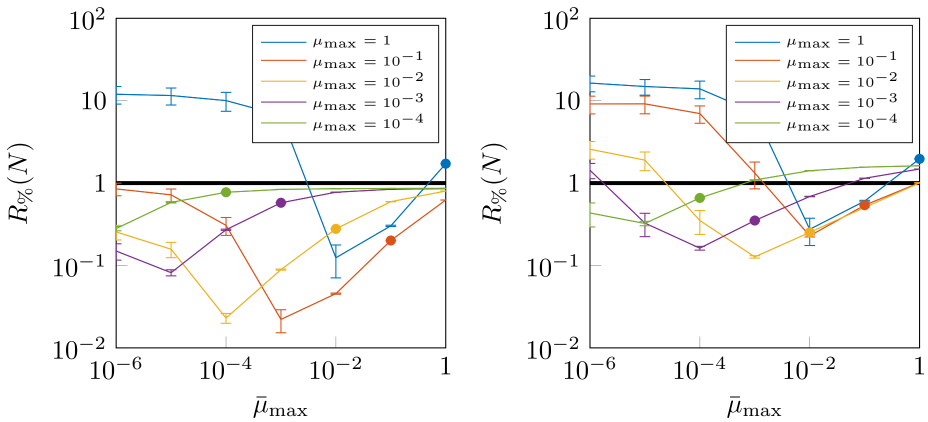

Figure 6: $R_{\%}(N)$ obtained with UCB-LM with different configurations: $K = 33$ arms, $\mathcal{S}_L$ (left) and $\mathcal{S}_H$ (right). The error bars represents the 95% confidence intervals for the expected values. The black like represents the performance of the UCB1 algorithm.

and SW-UCB from [27], as representatives of frequentist MAB designed for nonstationary environments. In addition to the already presented SW-UCB-M, we extend the sliding window approach to other algorithms proposing the SW-UCB-LM and the SW-UCBV-M algorithms, to include the information about the maximum conversion probability and to consider UCBV-like algorithms with the monotonicity information, respectively. The pseudocode of these algorithms is provided in Appendix B.

6.1. Experimental Setting and Performance Indices

The experimental setting considers a number of rounds of $N = 4 \cdot 10^7$ and uses two different abruptly changing pdfs, denoted with $\mathcal{S}_{LHLH}$ and $\mathcal{S}_{HLHL}$, each of which contains three breakpoints at rounds $t = 10^7$, $t = 2 \cdot 10^7$ and $t = 3 \cdot 10^7$. The threshold pdf switches from $\mathcal{S}_L$ to $\mathcal{S}_H$ or *vice versa* for $\mathcal{S}_{LHLH}$ and $\mathcal{S}_{HLHL}$, respectively, where $\mathcal{S}_L$ and $\mathcal{S}_H$ are defined as in Section 5. For instance, $\mathcal{S}_{LHLH}$ starts with $\mathcal{S}_L$ in phase Φ_1, then switches to $\mathcal{S}_H$ in phase Φ_2 and so on. A number of 3 switches demonstrated to be sufficient to show the behavior of the algorithms and, at the same time, tractable in terms of computational effort (more switches would require longer experiments, requiring a higher computational effort). For the sliding window algorithms, we choose a sliding window $\tau = 4\sqrt{N \log(N)}$ and we consider a parameter $\xi = 0.6$ for SW-UCB and SW-UCB-M, as in [27]. We average the results over 100 independent trials.

We redefine the performance indices using SW-UCB as a baseline in place of UCB1 as follows:

$$R_{\%}(t) = \frac{\bar{R}_t(\mathfrak{U})}{\bar{R}_t(\text{SW-UCB})},$$

$$\Delta P(t) = \sum_{t'=1}^{t} \mathbb{E}\left[V_{i_{(\mathfrak{U},t')}}\right] - \sum_{t'=1}^{t} \mathbb{E}\left[V_{i_{(\text{SW-UCB},t')}}\right],$$

$$\Delta P_{\%}(t) = \frac{\Delta P(t)}{\sum_{t'=1}^{t} \mathbb{E}[V_{i_{(\text{SW-UCB},t')}}]},$$

where $\mathfrak{U}$ is a generic policy, $i_{(\mathfrak{U},t)}$ is the index chosen by policy $\mathfrak{U}$ at time t.

6.2. Regret Analysis

The average $R_{\%}(N)$ and the 95% confidence intervals are reported in Table 2 (the results of SW-UCB-L and SW-UCB-LM are omitted for $\mu_{\max} = 1$, their bound being always larger than the one used in SW-UCB). As in the stationary case, we omit the evaluation of $R_{\%}(t)$ for $t < N$, since we provide in the next section a detailed discussion about how the profit provided by the algorithms changes as t changes and we believe this latter evaluation is more significant in practice than the evaluation of the dependency of the regret on time.

The first observation we provide is that, except for some specific cases, the performance in terms of regret of each algorithm is similar in the two configurations $\mathcal{S}_{LHLH}$ and $\mathcal{S}_{HLHL}$. This shows that the switches between L and H and *vice versa* do not significantly affect the performance of the algorithms. Instead, the performance depends on the number of L and H configurations. This holds for all the algorithms, $\mu_{\max}$ values, and the numbers of arms, except for the following special cases:

- UCBV-M: it performs much worse in $\mathcal{S}_{LHLH}$ than in $\mathcal{S}_{HLHL}$ for $\mu_{\max} \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. This result does not depend on the exploitation of the monotonicity, but on the fact that, once UCBV-M has learned a configuration L or H, its bounds do not significantly change after an abrupt change given that it does not exploit any sliding window and the optimal arm in the configuration L has a very small relative reward in configuration H. This does not hold when $\mu_{\max} = 10^{-4}$ since the sliding window is excessively small, and the baseline SW-UCB cannot learn anything.

- SW-UCB-M and SW-UCBV-M: they perform worse in $\mathcal{S}_{LHLH}$ than in $\mathcal{S}_{HLHL}$ for $\mu_{\max} = 1$. This is an anomaly of our algorithms. In this specific case, the cost of exploiting the monotonicity is larger than the gain provided by the algorithm.

Summarily, we can observe that: SW-UCBV is the optimal algorithm for $\mu_{\max} = 1$, SW-UCBV-M is the optimal algorithm for $\mu_{\max} \in \{10^{-1}, 10^{-2}\}$, and UCBV-M is the optimal algorithm for $\mu_{\max} \in \{10^{-3}, 10^{-4}\}$ except in the configuration $\mathcal{S}_{LHLH}$, where, instead, for $\mu_{\max} = 10^{-3}$ SW-UCB-LM is the best one. This is because the exploitation of the monotonicity allows an algorithm to perform better, but it requires a cost, i.e., the one incurred when a union bound over $1 \leq j \leq i$ is performed. When the setting is easy (e.g., $\mu_{\max}$ is very high), the improvement provided by the monotonicity is smaller than the cost needed for its exploitation. Instead, for $\mu_{\max} \in \{10^{-1}, 10^{-2}\}$, the cost required for the exploitation of the monotonicity is much lower than the gain. When $\mu_{\max}$ is smaller, e.g., $\mu_{\max} \in \{10^{-3}, 10^{-4}\}$, the setting is too hard, and we suppose that an optimal solution to the problem would require a sliding window longer than that one used here. Indeed, the fact that UCBV-M is the best algorithms essentially shows that abstaining from learning after the first abrupt change is better than trying to learn the change. In these settings that are so hard, a different approach should be used: for instance, one could identify the abrupt change and employ different stationary MAB policies, one per phase.

Finally, we remark that in every configuration it is possible to outperform the baseline and in many cases the reduction of regret is significant.

6.3. Profit Analysis

The average $\Delta P_{\%}(t)$ and $\Delta P(t)$ for $t \in \{1, \dots, 10^7\}$ obtained with SW-UCB-LM and SW-UCBV-M (with respect to the results obtained with SW-UCB and 5 arms) are reported in Figure 7 and Figure 8, respectively.[6]

The main difference between the results in the stationary settings and those in nonstationary settings concerns the trend of $\Delta P_{\%}(t)$. In the case of the stationary settings, $\Delta P_{\%}(t)$ achieves a maximum and subsequently goes asymptotically to zero, showing that our algorithms provide a gain in the early stages of the learning process. Instead, in the case of nonstationary settings, our algorithms repeatedly provide a gain at each abrupt change. This is showed by the fact that $\Delta P_{\%}(t)$ does not go to zero as t increases. Therefore, the $\Delta P(t)$

[6]Results for SW-UCB-LM in the case $\mu_{\max} = 1$ are not reported since this algorithm requires $\mu_{\max} < \frac{1}{2}$ to be effective.

has an upward trend over time. To summarize, these results provide evidence for a promising application of the proposed SW algorithms in the nonstationary setting.

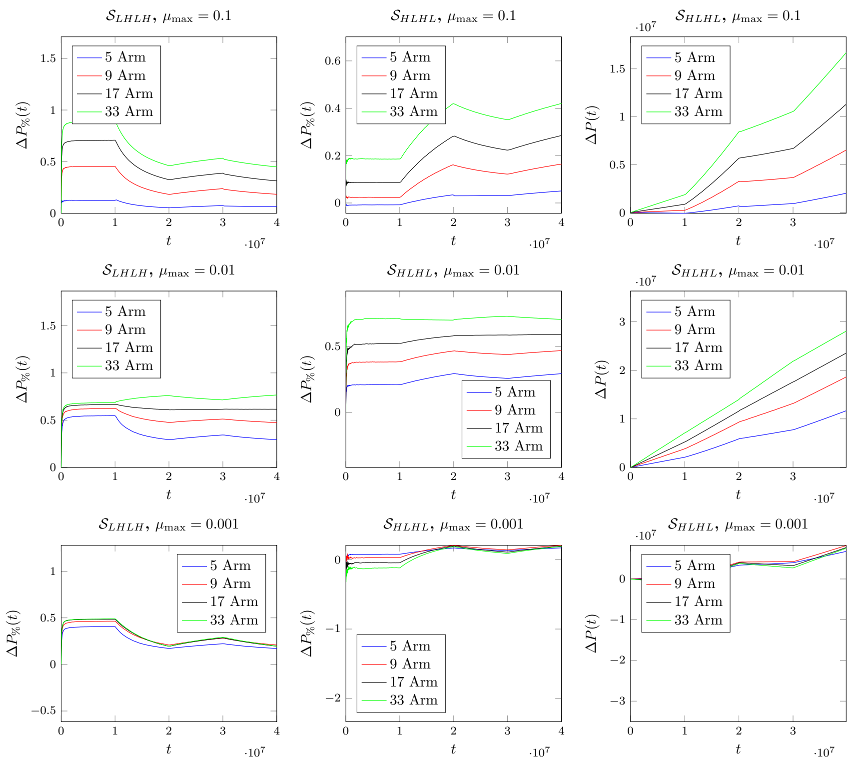

Figure 7: $\Delta P_{\%}(t)$ (first two columns) and $\Delta P(t)$ (third column) obtained with SW-UCB-LM with different configurations.

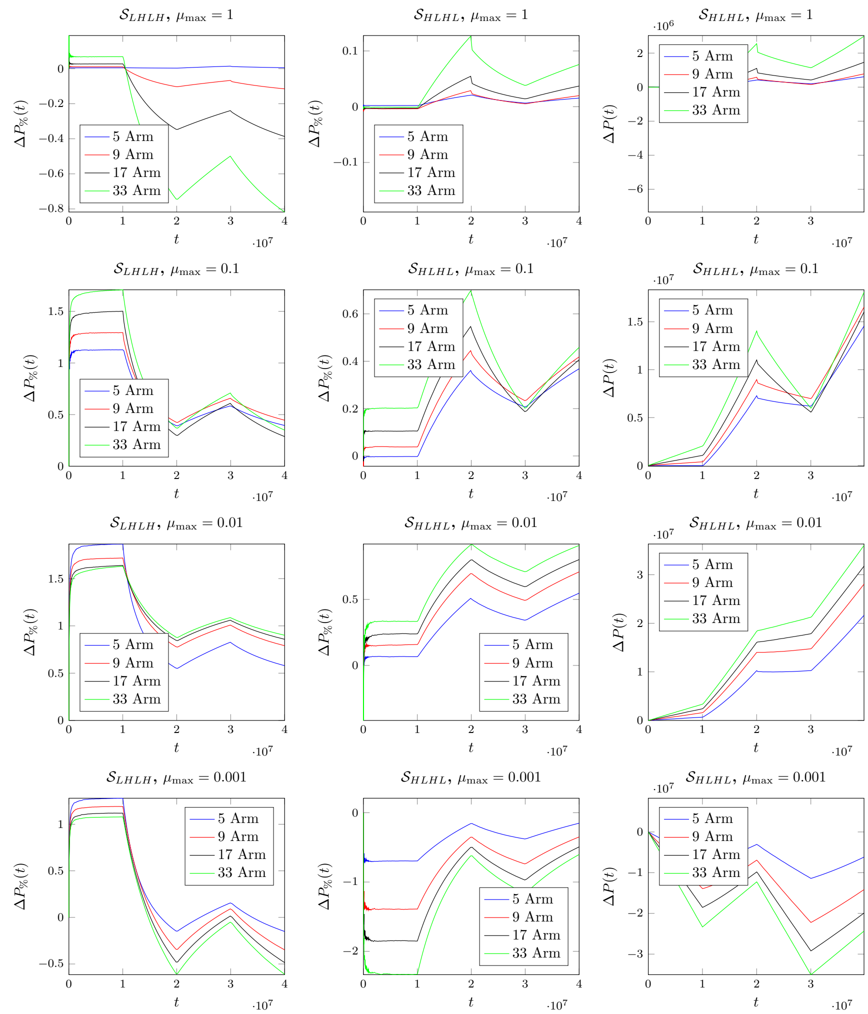

Figure 8: $\Delta P_{\%}(t)$ (first two columns) and $\Delta P(t)$ (third column) obtained with SW-UCBV-M with different configurations.

Table 2: Results concerning $R_{\%}$ in nonstationary settings (averaged values over 100 runs, $\pm$ 95% confidence intervals).

		S_{LHLH}					
$\mu_{\max}$	$\lvert A\rvert$	SW-UCB-M	SW-UCB-L	SW-UCB-LM	SW-UCBV	SW-UCBV-M	UCBV-M
1	5	1.02 ± 0.00	——	——	$\mathbf{0.95 \pm 0.00}$	0.98 ± 0.02	10.67 ± 0.00
	9	1.82 ± 0.26	——	——	$\mathbf{0.94 \pm 0.00}$	1.46 ± 0.18	10.18 ± 0.00
	17	2.29 ± 0.35	——	——	$\mathbf{0.92 \pm 0.00}$	2.44 ± 0.37	9.57 ± 0.00
	33	2.96 ± 0.44	——	——	$\mathbf{0.91 \pm 0.00}$	3.49 ± 0.44	8.60 ± 0.06
10^{-1}	5	0.82 ± 0.01	1.08 ± 0.00	0.88 ± 0.01	0.33 ± 0.00	$\mathbf{0.25 \pm 0.02}$	5.71 ± 0.00
	9	0.72 ± 0.01	1.05 ± 0.00	0.75 ± 0.01	0.51 ± 0.00	$\mathbf{0.40 \pm 0.05}$	4.77 ± 0.00
	17	$\mathbf{0.63 \pm 0.02}$	1.04 ± 0.00	0.64 ± 0.01	0.66 ± 0.00	$\mathbf{0.67 \pm 0.13}$	4.41 ± 0.03
	33	$\mathbf{0.57 \pm 0.02}$	1.04 ± 0.00	$\mathbf{0.57 \pm 0.01}$	0.82 ± 0.00	0.68 ± 0.13	4.01 ± 0.08
10^{-2}	5	0.98 ± 0.00	0.88 ± 0.00	0.79 ± 0.00	0.74 ± 0.00	$\mathbf{0.58 \pm 0.01}$	3.37 ± 0.05
	9	0.98 ± 0.00	0.89 ± 0.00	0.76 ± 0.00	0.88 ± 0.00	$\mathbf{0.59 \pm 0.01}$	3.10 ± 0.03
	17	0.97 ± 0.00	0.90 ± 0.00	0.71 ± 0.00	0.98 ± 0.00	$\mathbf{0.60 \pm 0.01}$	2.97 ± 0.07
	33	0.97 ± 0.00	0.92 ± 0.00	0.69 ± 0.01	1.07 ± 0.00	$\mathbf{0.60 \pm 0.02}$	2.78 ± 0.13
10^{-3}	5	1.10 ± 0.00	0.92 ± 0.00	$\mathbf{0.90 \pm 0.00}$	1.09 ± 0.00	1.08 ± 0.00	3.00 ± 0.11
	9	1.14 ± 0.00	0.93 ± 0.00	$\mathbf{0.92 \pm 0.00}$	1.14 ± 0.00	1.14 ± 0.00	2.65 ± 0.14
	17	1.17 ± 0.00	0.94 ± 0.00	$\mathbf{0.93 \pm 0.00}$	1.19 ± 0.00	1.18 ± 0.00	1.84 ± 0.24
	33	1.19 ± 0.00	0.96 ± 0.00	$\mathbf{0.93 \pm 0.00}$	1.21 ± 0.00	1.20 ± 0.01	1.25 ± 0.23
10^{-4}	5	1.12 ± 0.00	$\mathbf{0.97 \pm 0.00}$	1.04 ± 0.00	1.24 ± 0.00	1.49 ± 0.00	1.34 ± 0.24
	9	1.17 ± 0.00	0.97 ± 0.00	1.08 ± 0.00	1.24 ± 0.00	1.56 ± 0.00	$\mathbf{0.57 \pm 0.01}$
	17	1.21 ± 0.00	0.98 ± 0.00	1.11 ± 0.00	1.26 ± 0.00	1.63 ± 0.00	$\mathbf{0.55 \pm 0.01}$
	33	1.23 ± 0.00	0.98 ± 0.00	1.12 ± 0.00	1.26 ± 0.00	1.64 ± 0.01	$\mathbf{0.54 \pm 0.00}$

		S_{HLHL}					
$\mu_{\max}$	$\lvert A\rvert$	SW-UCB-M	SW-UCB-L	SW-UCB-LM	SW-UCBV	SW-UCBV-M	UCBV-M
1	5	1.01 ± 0.00	——	——	$\mathbf{0.96 \pm 0.00}$	$\mathbf{0.97 \pm 0.01}$	1.44 ± 0.00
	9	0.99 ± 0.00	——	——	$\mathbf{0.95 \pm 0.00}$	$\mathbf{0.96 \pm 0.01}$	1.32 ± 0.00
	17	0.97 ± 0.01	——	——	$\mathbf{0.93 \pm 0.00}$	$\mathbf{0.94 \pm 0.01}$	1.29 ± 0.00
	33	0.92 ± 0.01	——	——	$\mathbf{0.89 \pm 0.00}$	$\mathbf{0.89 \pm 0.01}$	1.18 ± 0.00
10^{-1}	5	0.86 ± 0.01	1.08 ± 0.00	0.90 ± 0.01	0.33 ± 0.00	$\mathbf{0.29 \pm 0.01}$	1.43 ± 0.00
	9	0.75 ± 0.01	1.05 ± 0.00	0.78 ± 0.01	0.51 ± 0.00	$\mathbf{0.43 \pm 0.03}$	1.10 ± 0.00
	17	0.66 ± 0.01	1.04 ± 0.00	0.67 ± 0.01	0.66 ± 0.00	$\mathbf{0.53 \pm 0.07}$	1.04 ± 0.01
	33	$\mathbf{0.59 \pm 0.01}$	1.04 ± 0.00	$\mathbf{0.60 \pm 0.01}$	0.82 ± 0.00	$\mathbf{0.57 \pm 0.07}$	0.95 ± 0.01
10^{-2}	5	0.98 ± 0.00	0.88 ± 0.00	0.79 ± 0.00	0.74 ± 0.00	$\mathbf{0.60 \pm 0.01}$	0.85 ± 0.00
	9	0.98 ± 0.00	0.89 ± 0.00	0.76 ± 0.00	0.88 ± 0.00	$\mathbf{0.64 \pm 0.01}$	0.76 ± 0.01
	17	0.97 ± 0.00	0.90 ± 0.00	0.73 ± 0.00	0.98 ± 0.00	$\mathbf{0.63 \pm 0.01}$	0.74 ± 0.01
	33	0.97 ± 0.00	0.92 ± 0.00	0.71 ± 0.00	1.07 ± 0.00	$\mathbf{0.63 \pm 0.01}$	0.72 ± 0.01
10^{-3}	5	1.10 ± 0.00	0.92 ± 0.00	0.90 ± 0.00	1.09 ± 0.00	1.08 ± 0.00	$\mathbf{0.83 \pm 0.03}$
	9	1.14 ± 0.00	0.93 ± 0.00	0.92 ± 0.00	1.14 ± 0.00	1.14 ± 0.00	$\mathbf{0.76 \pm 0.02}$
	17	1.17 ± 0.00	0.94 ± 0.00	0.93 ± 0.00	1.19 ± 0.00	1.18 ± 0.00	$\mathbf{0.75 \pm 0.02}$
	33	1.19 ± 0.00	0.96 ± 0.00	0.94 ± 0.00	1.21 ± 0.00	1.21 ± 0.00	$\mathbf{0.75 \pm 0.01}$
10^{-4}	5	1.12 ± 0.00	0.97 ± 0.00	1.04 ± 0.00	1.24 ± 0.00	1.49 ± 0.00	$\mathbf{0.85 \pm 0.02}$
	9	1.17 ± 0.00	0.97 ± 0.00	1.08 ± 0.00	1.24 ± 0.00	1.56 ± 0.00	$\mathbf{0.82 \pm 0.01}$
	17	1.21 ± 0.00	0.98 ± 0.00	1.10 ± 0.00	1.26 ± 0.00	1.63 ± 0.00	$\mathbf{0.81 \pm 0.01}$
	33	1.23 ± 0.00	0.98 ± 0.00	1.12 ± 0.00	1.26 ± 0.00	1.65 ± 0.00	$\mathbf{0.80 \pm 0.01}$

7. Conclusions and Future Works

In this paper, we focus on the problem of learning the best gross margin to maximize the seller profit, while minimizing the regret caused by the exploration of sub-optimal gross margins. Previous theoretical works focus on the selection of a finite set of gross margins, while previous heuristic works propose algorithms for specific settings without any theoretical guarantee on the worst-case regret. In this paper, we study how to exploit two properties of the pricing problem to improve the empiric performance of general-purpose bandit algorithms without losing their theoretical guarantees on the regret. The two properties we study hold in general settings: the first one is the (decreasing) monotonicity of the conversion rate on the gross margins, while the second one is the *a priori* information about the maximum conversion rate $\mu_{\max}$. Furthermore, we focus both on stationary settings and nonstationary settings. We provide some algorithms that we summarize in Table 3.

	Stationary		Nonstationary	
	Generic	Monotonic	Generic	Monotonic
$\mu \in [0, 1]$	UCB1, UCB-V	**UCB1-M**, **UCB-V-M**	SW-UCB	**SW-UCB-M**, **SW-UCBV-M**
$\mu \in [0, \mu_{\max}]$	**UCB-L**	**UCB-LM**	**SW-UCB-L**	**SW-UCB-LM**

Table 3: Algorithms for the different assumptions and scenarios analysed in the paper. We use the boldface for the algorithms proposed in this paper.

We provide a wide experimental evaluation of our algorithms, comparing them with other frequentist MAB algorithms with theoretical guarantees that do not exploit the two aforementioned properties. In this way, we evaluate the improvement obtained thanks to the exploitation of the problem characteristics. We elect two algorithms as the best ones: UCBV-M and UCB-LM in stationary settings and SW-UCBV-M and SW-UCB-LM in nonstationary ones. In most of the configurations, our algorithms perform better than the general-purpose ones in the early stages of the learning process. In stationary settings, we observe that our algorithms outperform the other algorithms in terms of profit thanks to a greater gain at the beginning of the learning process. This gain is then kept constant up to the end of the process. As a result, the ratio in time between the profit provided by our algorithms and the baselines achieves a maximum and, subsequently, it asymptotically goes to zero as the time increases. In nonstationary settings, instead, such a ratio does not go asymptotically to zero and thus the difference of profits between our algorithms and the baselines increases as the time increases. This is of paramount importance in practice, potentially allowing a company to dramatically increase its profit.

Future developments of this work may study the exploitation of the monotonicity property in continuous MAB settings. Furthermore, we are also interested in studying a generalized version of the monotonicity property, where the ordering among expected values is only partial. Finally, our goal is to study how the two aforementioned properties can be exploited with Bayesian MAB algorithms such as the Thompson Sampling.

[1] R. Kleinberg, T. Leighton, The value of knowing a demand curve: Bounds on regret for online posted-price auctions, in: FOCS, IEEE Computer Society, 2003, pp. 594–605.

[2] T. L. Lai, H. Robbins, Asymptotically efficient adaptive allocation rules, ADV APPL MATH 6 (1) (1985) 4–22.

[3] N. Cesa-Bianchi, G. Lugosi, Prediction, learning, and games, Cambridge University Press, 2006.

[4] A. Mas-Colell, M. D. Whinston, J. R. Green, et al., Microeconomic theory, Vol. 1, OUP New York, 1995.

[5] H. K. Cheng, Q. C. Tang, Free trial or no free trial: Optimal software product design with network effects, European Journal of Operational Research 205 (2) (2010) 437–447.

[6] N. Gatti, A. Lazaric, F. Trovò, A truthful learning mechanism for contextual multi-slot sponsored search auctions with externalities, in: EC, 2012, pp. 605–622.

[7] A. Piccolboni, C. Schindelhauer, Discrete prediction games with arbitrary feedback and loss, in: COLT, Springer, 2001, pp. 208–223.

[8] W. W. Moe, P. S. Fader, Dynamic conversion behavior at e-commerce sites, MANAGE SCI 50 (3) (2004) 326–335.

[9] R. Combes, A. Proutiere, Unimodal bandits: Regret lower bounds and optimal algorithms, in: ICML, 2014, pp. 521–529.

[10] Y. Y. Jia, S. Mannor, Unimodal bandits, in: ICML-11, 2011, pp. 41–48.

[11] N. Alon, N. Cesa-Bianchi, C. Gentile, Y. Mansour, From bandits to experts: A tale of domination and independence, in: NIPS, 2013, pp. 1610–1618.

[12] S. Mannor, O. Shamir, From bandits to experts: On the value of side-observations, in: NIPS, 2011, pp. 684–692.

[13] G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, C. Szepesvári, Partial monitoring-classification, regret bounds, and algorithms, MATH OPER RES 39 (4) (2014) 967–997.

[14] N. Cesa-Bianchi, G. Lugosi, G. Stoltz, Regret minimization under partial monitoring, MATH OPER RES 31 (3) (2006) 562–580.

[15] G. Bartók, D. Pál, C. Szepesvári, Minimax regret of finite partial-monitoring games in stochastic environments., in: COLT, Vol. 2011, JMLR, 2011, pp. 133–154.

[16] G. Bartók, N. Zolghadr, C. Szepesvári, An adaptive algorithm for finite stochastic partial monitoring, in: ICML, 2012, pp. 1727–1734.

[17] D. P. Foster, A. Rakhlin, No internal regret via neighborhood watch, in: AISTATS, 2012, pp. 382–390.

[18] O. Besbes, A. Zeevi, Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms, OPER RES 57 (6) (2009) 1407–1420.

[19] M. Chhabra, S. Das, Learning the demand curve in posted-price digital goods auctions, in: AAMAS, 2011, pp. 63–70.

[20] J. Broder, P. Rusmevichientong, Dynamic pricing under a general parametric choice model, OPER RES 60 (4) (2012) 965–980.

[21] N. B. Keskin, A. Zeevi, Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies, OPER RES 62 (5) (2014) 1142–1167.

[22] O. Besbes, A. Zeevi, On the (surprising) sufficiency of linear models for dynamic pricing with demand learning, MANAGE SCI 61 (4) (2015) 723–739.

[23] A. Blum, V. Kumar, A. Rudra, F. Wu, Online learning in online auctions, THEOR COMPUT SCI 324 (2) (2004) 137–146.

[24] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, The nonstochastic multiarmed bandit problem, SIAM J COMPUT 32 (1) (2002) 48–77.

[25] R. Cole, T. Roughgarden, The sample complexity of revenue maximization, in: STOC, 2014, pp. 243–252.

[26] J. H. Morgenstern, T. Roughgarden, On the pseudo-dimension of nearly optimal auctions, in: NIPS, 2015, pp. 136–144.

[27] A. Garivier, E. Moulines, On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems, ArXiv e-printsarXiv:0805.3415.

[28] A. Garivier, E. Moulines, On upper-confidence bound policies for switching bandit problems, in: ALT, 2011, pp. 174–188.

[29] D. St-Pierre, J. Liu, Differential evolution algorithm applied to non-stationary bandit problem, in: 2014 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2014, pp. 2397–2403.

[30] E. Kaufmann, N. Korda, R. Munos, Thompson sampling: An asymptotically optimal finite-time analysis, in: ALT, 2012, pp. 199–213.

[31] P. Auer, N. Cesa-Bianchi, P. Fischer, Finite-time analysis of the multiarmed bandit problem, MACH LEARN 47 (2-3) (2002) 235–256.

[32] J. Audibert, R. Munos, C. Szepesvári, Exploration–exploitation tradeoff using variance estimates in multi-armed bandits, THEOR COMPUT SCI 410 (19) (2009) 1876–1902.

[33] H. Chernoff, A note on an inequality involving the normal distribution, ANN PROBAB 9 (3) (1981) 533–535.

[34] W. Hoeffding, Probability inequalities for sums of bounded random variables, J AM STAT ASSOC 58 (301) (1963) 13–30.

[35] F. Chung, L. Lu, Concentration inequalities and martingale inequalities: a survey, Internet Mathematics 3 (1) (2006) 79–127.

[36] Monetate, Monetate ecommerce quarterly, `http://www.monetate.com/resources/research/` (2015).

Appendix A. Proofs of the theorems

Appendix A.1. Proof of Theorem 1

Theorem 1. *If policy UCB1-M is run over a stationary MAB setting with a monotonic set A, the expected regret after N rounds is at most:*

$$\overline{R}_N \leq \sum_{i|a_i \neq a_{i^*}} \frac{8a_i^2 \log(N)}{\Delta_i} + \sum_{i|a_i \neq a_{i^*}} \frac{2a_i^2 \log(K)}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \sum_{i=1}^{K} \Delta_i,$$

where $\Delta_i := a_{i^*}\mu_{i^*} - a_i\mu_i, \forall i \in \{1, \ldots, K\}$.

Proof. Let us remind that we denote with $i^* := \arg\max_{i \in \{1,\ldots,K\}} a_i\mu_i$ the index corresponding to the optimal arm a_{i^*}. Similarly to [31], we want to compute the expected number of times the policy UCB1-M does not pick the optimal arm a_{i^*} or, more formally, $\mathbb{E}[T_i(N)]$, $\forall a_i \neq a_{i^*}$ and compute the regret as:

$$\overline{R}_N = \sum_{i|a_i \neq a_{i^*}} \Delta_i \mathbb{E}[T_i(N)].$$

Consider the round of the learning process at which a specific arm a_i has been selected for s rounds and define:

- $\bar{j}(i,t) := \bar{j}$ (with abuse of notation) as the index $j \in \{1, \ldots, i\}$ minimizing the quantity $\overline{x}_{ji,t} + \sqrt{\frac{4\log(t)+\log(i)}{2T_{ji}(t-1)}}$, i.e., the upper bound of arm a_i;
- $\bar{j}^* := \bar{j}(i^*, t)$ as the index $j \in \{1, \ldots, i^*\}$ minimizing the quantity $\overline{x}_{ji^*,t} + \sqrt{\frac{4\log(t)+\log(i^*)}{2T_{ji^*}(t-1)}}$, i.e., the upper bound of arm a_{i^*};
- $\overline{X}_{i,(s)}$ is the unbiased estimate of μ_i in the case we collected a total of s samples from arm a_i;
- $\overline{X}_{\bar{j}i,(s)}$ is the unbiased estimate of $\mu_{\bar{j}i,t,s} = \mathbb{E}\left[\overline{X}_{\bar{j}i,(s)}\right]$, in the case we collected a total of s samples from arm a_i (and thus we use $s' \geq s$ samples to estimate $\mu_{\bar{j}i,s}$);
- $c_{i,t,s} := \sqrt{\frac{4\log(t)+\log(i)}{2s}}$ as the Hoeffding bound with confidence $\frac{t^{-4}}{i}$ for $\overline{X}_{i,(s)}$ after t rounds;
- $c_{ji,t,s} := \sqrt{\frac{4\log(t)+\log(i)}{2s'}}$ as the Hoeffding bound with confidence $\frac{t^{-4}}{i}$ for $\overline{X}_{ji,(s)}$ after t rounds, in the case arm a_i has been pulled a total of s times and the arms $\{a_j, \ldots, a_i\}$ have been chosen in total $s' > s$ times.

We have that, for each $l > 0$:

$$T_i(N) = 1 + \sum_{t=K+1}^{N} \mathbb{1}\{i_t = i\} \le l + \sum_{t=K+1}^{N} \mathbb{1}\{i_t = i, T_i(t-1) \ge l\}$$
$$\le l + \sum_{t=K+1}^{N} \mathbb{1}\Big\{a_{i^*}\overline{X}_{\bar{j}^* i^*,t} + a_{i^*} c_{\bar{j}^* i^*,t,T_{i^*}(t-1)} \le a_i \overline{X}_{\bar{j}i,t} +$$
$$+ a_i \ c_{\bar{j}i,t,T_i(t-1)}, T_i(t-1) \ge l\Big\}$$
$$\le l + \sum_{t=K+1}^{N} \mathbb{1}\left\{\min_{0<s<t}\left(a_{i^*}\overline{X}_{\bar{j}^* i^*,(s)} + a_{i^*} c_{\bar{j}^* i^*,t,s}\right) \le \max_{l<s_i<t}\left(a_i \overline{X}_{\bar{j}i,(s_i)} + a_i c_{\bar{j}i,t,s_i}\right)\right\}$$
$$\le l + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=l}^{t-1} \mathbb{1}\left\{a_{i^*}\overline{X}_{\bar{j}^* i^*,(s)} + a_{i^*} c_{\bar{j}^* i^*,t,s} \le a_i \overline{X}_{\bar{j}i,(s_i)} + a_i c_{\bar{j}i,t,s_i}\right\}.$$

where we denoted with $\mathbb{1}\{B\}$ the indicator function of the event B.

If we consider:

$$a_{i^*}\overline{X}_{\bar{j}^* i^*,(s)} + a_{i^*} c_{\bar{j}^* i^*,t,s} \le a_i \overline{X}_{\bar{j}i,(s_i)} + a_i c_{\bar{j}i,t,s_i},$$
$$a_{i^*}\overline{X}_{\bar{j}^* i^*,(s)} + a_{i^*} c_{\bar{j}^* i^*,t,s} - a_i \overline{X}_{\bar{j}i,(s_i)} - a_i c_{\bar{j}i,t,s_i} \le 0,$$
$$a_{i^*}\overline{X}_{\bar{j}^* i^*,(s)} - a_{i^*}\mu_{i^*} + a_{i^*} c_{\bar{j}^* i^*,t,s} - a_i \overline{X}_{i,t,(s_i)} + a_i\mu_i + a_i c_{i,t,s_i} +$$
$$+ a_{i^*}\mu_{i^*} - a_i\mu_i + a_i\overline{X}_{i,(s_i)} - a_i \overline{X}_{\bar{j}i,(s_i)} - a_i c_{\bar{j}i,t,s_i} - a_i c_{i,t,s_i} \le 0,$$

we have that that, if the previous inequality is satisfied, at least one of the following inequalities is satisfied:

$$a_{i^*}\overline{X}_{\bar{j}^* i^*,(s)} \le a_{i^*}\mu_{i^*} - a_{i^*} c_{\bar{j}^* i^*,t,s} \tag{A.1}$$
$$a_i\overline{X}_{i,(s_i)} \ge a_i\mu_i + a_i c_{i,t,s_i} \tag{A.2}$$
$$a_{i^*}\mu_{i^*} - a_i\mu_i + a_i\overline{X}_{i,(s_i)} - a_i\overline{X}_{\bar{j}i,(s_i)} - a_i c_{\bar{j}i,t,s_i} - a_i c_{i,t,s_i} \le 0. \tag{A.3}$$

We need to bound the probabilities that the each one of the previous events occurs.

Probability of Event (A.1) By considering the fact that $\overline{X}_{\bar{j}^* i^*,(s)} + c_{\bar{j}^* i^*,t,s}$ is an upper bound for $\mu_{\bar{j}i,t,s}$ and thanks to the monotonicity assumption over μ_{i^*}, we can bound the probability of the events in Equation (A.1) as follows:

$$\mathbb{P}\left(a_{i^*}\overline{X}_{\bar{j}^* i^*,(s)} \le a_{i^*}\mu_{i^*} - a_{i^*} c_{\bar{j}^* i^*,t,s}\right)$$
$$= \mathbb{P}\left(\overline{X}_{\bar{j}^* i^*,(s)} \le \mu_{i^*} - c_{\bar{j}^* i^*,t,s}\right)$$
$$\le \mathbb{P}\left(\overline{X}_{\bar{j}^* i^*,(s)} + c_{\bar{j}^* i^*,t,s} \le \mu_{i^*}\right)$$
$$\le \mathbb{P}\left(\overline{X}_{\bar{j}^* i^*,(s)} + c_{\bar{j}^* i^*,t,s} \le \mu_{\bar{j}i,t,s}\right) \le e^{-4\log t} = t^{-4},$$

where the i term disappeared with the union bound over $\overline{X}_{ji^*,(s)}$ such that $1 \le j \le i^*$.

Probability of Event (A.2) By considering the Hoeffding bound we have that the event Equation A.2 is bounded by:

$$\begin{aligned}&\mathbb{P}(a_i\overline{X}_{i,(s_i)} \geq a_i\mu_i + a_i c_{i,t,s_i})\\&= \mathbb{P}\left(\overline{X}_{i,(s_i)} \geq \mu_i + c_{i,t,s_i}\right) \leq e^{-4\log t - \log i} = \frac{t^{-4}}{i} \leq t^{-4}.\end{aligned}$$

Probability of Event (A.3) Note that since the algorithm chooses the tightest bound among the set $\overline{X}_{ji,(s)} + c_{ji,t,s}$ with $j \leq i$ we have:

$$\begin{aligned}a_i\overline{X}_{\bar{j}i,(s_i)} + a_i c_{\bar{j}i,t,s_i} &\leq a_i\overline{X}_{i,(s_i)} + a_i c_{i,t,s_i},\\ a_i\overline{X}_{\bar{j}i,(s_i)} - a_i\overline{X}_{i,(s_i)} + a_i c_{\bar{j}i,t,s_i} &\leq a_i c_{i,t,s_i},\\ a_i\overline{X}_{i,(s_i)} - a_i\overline{X}_{\bar{j}i,(s_i)} - a_i c_{\bar{j}i,t,s_i} &\geq -a_i c_{i,t,s_i}\end{aligned}$$

If we consider $l = \left\lceil \frac{2a_i^2[4\log(t)+\log(i)]}{\Delta_i^2} \right\rceil$ the event in Equation (A.3) is not possible since:

$$0 \geq a_{i^*}\mu_{i^*} - a_i\mu_i + \underbrace{a_i\overline{X}_{i,(s_i)} - a_i\overline{X}_{\bar{j}i,(s_i)} - a_i c_{\bar{j}i,t,s_i}}_{\geq -a_i c_{i,t,s_i}} - a_i c_{i,t,s_i} \tag{A.4}$$

$$\geq \Delta_i - 2a_i\sqrt{\frac{4\log(t)+\log(i)}{2l}} > \Delta_i - \Delta_i = 0, \tag{A.5}$$

where we recall that $\Delta_i := a_{i^*}\mu_{i^*} - a_i\mu_i$.

Thus, since $\log(t) \leq \log(N)$ and $\log(i) \leq \log(K)$, $\forall i$ we have:

$$\begin{aligned}\mathbb{E}[T_i(N)] &\leq \left\lceil \frac{2a_i^2[4\log(N)+\log(K)]}{\Delta_i^2} \right\rceil + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=l}^{t-1} 2t^{-4}\\ &\leq \frac{8a_i^2\log(N)}{\Delta_i^2} + \frac{2a_i^2\log(K)}{\Delta_i^2} + 1 + \frac{\pi^2}{3}\end{aligned}$$

and the total regret becomes (since $\sum_{i=1}^{K}\mathbb{E}[T_i(N)] = N$):

$$\begin{aligned}\overline{R}_N &= a_{i^*}\mu_{i^*}N - \sum_{i=1}^{K}\mathbb{E}[T_i(N)]a_i\mu_i = \sum_{i=1}^{K}(a_{i^*}\mu_{i^*} - a_i\mu_i)\mathbb{E}[T_i(N)]\\ &\leq \sum_{i|a_i\neq a_{i^*}} \frac{8a_i^2\log(N)}{\Delta_i} + \sum_{i|a_i\neq a_{i^*}} \frac{2a_i^2\log(K)}{\Delta_i} + \left(1+\frac{\pi^2}{3}\right)\sum_{i=1}^{K}\Delta_i,\end{aligned}$$

which concludes the proof.

Appendix A.2. Proof of Theorem 2

Theorem 2. *If policy UCBV-M is run with $\xi = 1.2$ and $c = 1$ over a setting with a monotonic set A, the expected regret after N rounds is at most:*

$$\overline{R}_N \leq \frac{12}{5}\sum_{i|a_i\neq a_{i^*}} a_i^2\left(\frac{\sigma_i^2}{\Delta_i} + \frac{32}{15}\right)\log(N) + \sum_{i|a_i\neq a_{i^*}} \Delta_i\left[1 + a_i^2\left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2}{\Delta_i}\right)\log(K)\right],$$

where $\sigma_i^2 := Var(X_{i,n}),\ \forall i \in \{1,\dots,K\}, \forall\ n \in \{1,\dots,T_i(N)\}$.

Proof. In what follows we make use of the notation used in Theorem 1. By following the proof of Theorem 3 in [32] we would like to bound the number of times a suboptimal arm is played:

$$\mathbb{E}[T_i(N)] \le l_i + \sum_{t=l_i+1}^{N} \underbrace{\sum_{s=l_i}^{t-1} \mathbb{P}\left(a_i \overline{X}_{\bar{j}i,(s)} + a_i c_{\bar{j}i,t,s} \ge a_{i^*}\mu_{i^*}\right)}_{T_{i1}} +$$
$$+ \sum_{t=l_i+1}^{N} \underbrace{\sum_{s=1}^{t-1} \mathbb{P}\left(a_{i^*} \overline{X}_{\bar{j}^* i^*,(s)} + a_{i^*} c_{\bar{j}^* i^*,t,s} \le a_{i^*}\mu_{i^*}\right)}_{T_{i2}},$$

where the inequality is due to Theorem 2 in [32]. Let us consider the two contribution to the regret separately.

Bound over $\mathbf{T_{i1}}$ The first contribution can be bounded as follows:

$$T_{i1} = \sum_{s=l_i}^{t-1} \mathbb{P}\big(a_i \overline{X}_{\bar{j}i,(s)} - a_i \overline{X}_{i,(s)} + a_i c_{\bar{j}i,t,s} - a_{i^*}\mu_{i^*} + a_i\mu_i +$$
$$+ a_i c_{i,t,s} + a_i \overline{X}_{i,(s)} - a_i\mu_i - a_i c_{i,t,s} \ge 0\big)$$
$$\le \sum_{s=l_i}^{t-1} \mathbb{P}\left(a_i \overline{X}_{\bar{j}i,(s)} - a_i \overline{X}_{i,(s)} + a_i c_{\bar{j}i,t,s} - a_{i^*}\mu_{i^*} + a_i\mu_i + a_i c_{i,t,s} > 0\right) + \tag{A.6}$$
$$+ \sum_{s=l_i}^{t-1} \mathbb{P}(a_i \overline{X}_{i,(s)} - a_i\mu_i - a_i c_{i,t,s} \ge 0). \tag{A.7}$$

By considering $s = l_i = \left\lceil 2a_i^2 \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2}{\Delta_i}\right) \max\{c,1\}\xi(log(t)+log(i))\right\rceil$, where $\sigma_i^2 := Var(X_{it}),\ \forall i \in \{1,\dots,K\}, t \in \{1,\dots,N\}$, we have that:

$$0 < \underbrace{a_i \overline{X}_{\bar{j}i,(s)} - a_i^2 \overline{X}_{i,(s)} + a_i c_{\bar{j}i,t,s}}_{\le a_i c_{i,t,s}} - a_{i^*}\mu_{i^*} + a_i\mu_i + a_i c_{i,t,s}$$
$$\le 2a_i c_{i,t,s} - \Delta_i \le \Delta_i - \Delta_i = 0,$$

where we used the fact that, by the choice made by the proposed algorithm, we have $a_i \overline{X}_{\bar{j}i,(s)} + a_i c_{\bar{j}i,t,s} \le a_i \overline{X}_{i,(s)} + a_i c_{i,t,s}$for each $j \in \{1,\dots,i\}$. Thus, the contribution of the term in Equation (A.6) to the regret is null since the aforementioned event is impossible.

The term in Equation (A.7) can be bounded by Theorem 1 in [32] in the following way:

$$\sum_{s=l_i}^{t-1} \mathbb{P}(a_i \overline{X}_{i,(s)} - a_i \mu_i - a_i c_{i,t,s} \geq 0)$$
$$\sum_{s=l_i}^{t-1} \mathbb{P}(a_i \overline{X}_{i,(s)} - a_i \mu_i - a_i c_{i,t,s} \geq 0) \leq \beta(t,c,i) \leq \beta(t,c)$$

where $\beta(t,c,i) := 3\min\{c,1\} \inf_{1<\alpha\leq 3} \left[\left(\min\left\{ \frac{\log(t)}{\log(\alpha)}, t \right\} \right) (ti)^{-\frac{\xi}{\alpha}} \right]$ and $\beta(t,c) := \beta(t,c,1)$.

Bound over $\mathbf{T_{i2}}$ By exploiting the monotonicity, i.e., since $\mu_{i^*} \leq \mu_{\bar{j}^* i^*,t,s}$ and by considering Theorem 1 in [32] we have:

$$\begin{aligned}
T_{i2} &= \sum_{s=1}^{t-1} \mathbb{P}\left(a_{i^*} \overline{X}_{\bar{j}^* i^*,(s)} + a_{i^*} c_{\bar{j}^* i^*,t,s} \leq a_{i^*} \mu_{i^*} \right) \\
&= \sum_{s=1}^{t-1} \mathbb{P}\left(\overline{X}_{\bar{j}^* i^*,(s)} + c_{\bar{j}^* i^*,t,s} \leq \mu_{i^*} \right) \\
&\leq \sum_{s=1}^{t-1} \mathbb{P}\left(\overline{X}_{\bar{j}^* i^*,(s)} + c_{\bar{j}^* i^*,t,s} \leq \mu_{\bar{j}^* i^*} \right) \leq \beta(t,c),
\end{aligned}$$

where for the monotonicity $\mu_{\bar{j}^* i^*} \geq \mu_{i^*}$ and we used a union bound over all the considered bounds ($j \in \{1, \ldots, i\}$).

Regret $\overline{\mathbf{R}}_{\mathbf{N}}$: Summing up, since $\log(t) \leq \log(N)$ and $\log(i) \leq \log(K)$, we have:

$$\begin{aligned}
\overline{R}_N &= \sum_{i=1}^{K} \mathbb{E}[T_i(N)]\Delta_i \leq \sum_{i|a_i \neq a_{i^*}} (l_i + \sum_{t=l_i+1}^{N} T_{i1} + T_{i2}) \\
&\leq \sum_{i|a_i \neq a_{i^*}} \left[1 + 2a_i^2 \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2}{\Delta_i} \right) \max\{c,1\}\xi(log(t) + log(i)) + 2 \sum_{t=l_i+1}^{N} \beta(t,c) \right] \Delta_i \\
&\leq \sum_{i|a_i \neq a_{i^*}} \left[\frac{12}{5} a_i^2 \left(\frac{\sigma_i^2}{\Delta_i} + 2 \right) \log(N) + 4c' \log(N) \right] + \\
&+ \sum_{i|a_i \neq a_{i^*}} \Delta_i \left[1 + a_i^2 \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2}{\Delta_i} \right) \log(K) \right] \\
&\leq \frac{12}{5} \sum_{i|a_i \neq a_{i^*}} a_i^2 \left(\frac{\sigma_i^2}{\Delta_i} + \frac{32}{15} \right) \log(N) + \sum_{i|a_i \neq a_{i^*}} \Delta_i \left[1 + a_i^2 \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2}{\Delta_i} \right) \log(K) \right],
\end{aligned}$$

where by choosing $\xi = 1.2$ and $c = 1$ we have $\sum_{t=l_i+1}^{N} \beta(t,c) \leq c' \frac{2\log(N)}{\Delta_k}$ with $c' \leq 0.08$ (see proof of Theorem 4 in [32] for details). This concludes the proof.

Appendix A.3. Proof of Theorem 4

Let us recall that thanks to the Chernoff's theorem we have:

Theorem 3 (Theorem 4 in [35], Lower tail). *Given a set of $T_i(t-1)$ independent and identically distributed random variables $\{X_{i,1}, \dots, X_{i,T_i(t-1)}\}$ such that $X_{i,s} \sim Be(\mu_i)$, for any $\varepsilon > 0$ we have:*

$$\mathbb{P}(\overline{X}_{i,t} + \varepsilon \le \mu_i) \le e^{-\frac{T_i(t-1)\varepsilon^2}{2\mu_i}}.$$

and also:

Theorem 8 (Theorem 4 in [35], Upper tail). *Given a set of $T_i(t-1)$ independent and identically distributed random variables $\{X_{i,1}, \dots, X_{i,T_i(t-1)}\}$ such that $X_{i,s} \sim Be(\mu_i)$, for any $\varepsilon > 0$ we have:*

$$\mathbb{P}(\overline{X}_{i,t} - \varepsilon \ge \mu_i) \le e^{-\frac{T_i(t-1)\varepsilon^2}{2\mu_i + \frac{\varepsilon}{3}}}.$$

Theorem 4. *If policy UCB-L is run over a stationary MAB setting with a set of arms A in which each arm $a_i \in A$ has outcome $X_{i,t}$ such that $\mathbb{E}[X_{i,t}] = \mu_i \le \mu_{\max} \le \frac{1}{2}$ for each $t \in \{1, \dots, N\}$, the expected regret after N rounds is at most:*

$$\overline{R}_N \le \sum_{i | a_i \neq a_{i^*}} \frac{32\mu_{\max} a_i^2 \log(N)}{\Delta_i} + \left[1 + \frac{\pi^2}{6} + \zeta\left(\frac{10}{7}\right)\right] \sum_{i=1}^{K} \Delta_i,$$

where $\zeta(\cdot)$ is the Riemann zeta function.

Proof. In what follows we make use of the notation used in Theorem 1. Let us recall that $\mu_{\max} \ge \mu_i, \ \forall i \in \{1, \dots, K\}$. By defining:

$$\varepsilon_{i,t,T_i(t-1)} := \sqrt{\frac{8\mu_{max}\log(t)}{T_i(t-1)}},$$

we have that, similarly to what has been derived in Theorem 1, for each $l > 0$:

$$T_i(N) \le l + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=l}^{t-1} \mathbb{1}\{a_{i^*}\overline{X}_{i^*,(s)} + a_{i^*}\varepsilon_{i^*,t,s} \le a_i \overline{X}_{i,(s_i)} + a_i \varepsilon_{i,t,s_i}\}.$$

If we consider the event in the previous inequality, we have:

$$a_{i^*}\overline{X}_{i^*,(s)} + a_{i^*}\varepsilon_{i^*,t,s} \le a_i \overline{X}_{i,(s_i)} + a_i \varepsilon_{i,t,s_i}$$
$$a_{i^*}\overline{X}_{i^*,(s)} - a_{i^*}\mu_{i^*} + a_{i^*}\varepsilon_{i^*,t,s} + a_{i^*}\mu_{i^*} \le a_i \overline{X}_{i,(s_i)} - a_i\mu_i - a_i\varepsilon_{i,t,s_i} + a_i\mu_i + 2a_i\varepsilon_{i,t,s_i}$$
$$a_{i^*}\overline{X}_{i^*,(s)} - a_{i^*}\mu_{i^*} + a_{i^*}\varepsilon_{i^*,t,s} - a_i \overline{X}_{i,(s_i)} + a_i\mu_i + a_i\varepsilon_{i,t,s_i} + a_{i^*}\mu_{i^*} - a_i\mu_i - 2a_i\varepsilon_{i,t,s_i} \le 0,$$

we have that it implies that at least one of the following inequalities is satisfied:

$$a_{i^*}\overline{X}_{i^*,(s)} \le a_{i^*}\mu_{i^*} - a_{i^*}\varepsilon_{i^*,t,s} \tag{A.8}$$
$$a_i\overline{X}_{i,(s_i)} \ge a_i\mu_i + a_i\varepsilon_{i,t,s_i} \tag{A.9}$$
$$a_{i^*}\mu_{i^*} - a_i\mu_i < 2a_i\varepsilon_{i,t,s_i}. \tag{A.10}$$

Let us focus on the event in Equation (A.8). Thanks to Theorem 3 we are able to bound the probability of this event:

$$\mathbb{P}(a_{i^*}\overline{X}_{i^*,(s)} \leq a_{i^*}\mu_{i^*} - a_{i^*}\varepsilon_{i^*,t,s}) = \mathbb{P}\left(\overline{X}_{i^*,(s)} \leq \mu_{i^*} - \varepsilon_{i^*,t,s}\right)$$
$$\leq e^{-\frac{s(\varepsilon_{i^*,t,s})^2}{2\mu^*}} \leq e^{-\frac{s(\varepsilon_{i^*,t,s})^2}{2\mu_{\max}}} = e^{-4\log t} = t^{-4}.$$

By relying on the upper tail of the Chernoff's bound, as described in Theorem 8 (cited in this appendix) we can bound the probability of the event in Equation (A.9):

$$\mathbb{P}(a_i\overline{X}_{i,(s_i)} \geq a_i\mu_i + a_i\varepsilon_{i,t,s_i}) = \mathbb{P}\left(\overline{X}_{i,(s_i)} \geq \mu_i + \varepsilon_{i,t,s_i}\right)$$
$$\leq \exp\left\{-\frac{s_i(\varepsilon_{i,t,s_i})^2}{2\mu_i + \frac{\varepsilon_{i,t,s_i}}{3}}\right\} \leq e^{-\frac{s_i(\varepsilon_{i,t,s_i})^2}{\frac{7}{3}\mu_{\max}}} \leq t^{-\frac{24}{7}},$$

where we consider $\varepsilon_{i,t,s_i} \leq \mu_{\max}$ and $\mu_i \leq \mu_{\max} \leq \frac{1}{2}$. At last, if we focus on the event in Equation (A.10) and we consider $l = \left\lceil \frac{32\mu_{\max}a_i^2\log(t)}{\Delta_i^2} \right\rceil$, where $\Delta_i = a_{i^*}\mu_{i^*} - a_i\mu_i$, the event in Equation (A.3) is not possible since:

$$0 \geq a_{i^*}\mu_{i^*} - a_i\mu_i - 2a_i\varepsilon_{i,t,s_i}$$
$$\underbrace{\geq}_{s_i \geq l} a_{i^*}\mu_{i^*} - a_i\mu_i - 2a_i\varepsilon_{i,t,l} \geq a_{i^*}\mu_{i^*} - a_i\mu_i - a_{i^*}\mu_{i^*} - a_i\mu_i = 0.$$

Finally we have:

$$\mathbb{E}[T_i(N)] \leq \left\lceil \frac{32\mu_{\max}a_i^2\log(t)}{\Delta_i^2} \right\rceil + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=l}^{t-1}(t^{-4} + t^{-\frac{24}{7}})$$
$$\leq \frac{32\mu_{\max}a_i^2\log(N)}{\Delta_i^2} + 1 + \frac{\pi^2}{6} + \zeta\left(\frac{10}{7}\right)$$

where $\zeta(\cdot)$ is the Riemann zeta function. The total regret becomes (since $\sum_{i=1}^{K}\mathbb{E}[T_i(N)] = N$):

$$\overline{R}_N = a_{i^*}\mu_{i^*}N - \sum_{i=1}^{K}\mathbb{E}[T_i(N)]a_i\mu_i = \sum_{i=1}^{K}(a_{i^*}\mu_{i^*} - a_i\mu_i)\mathbb{E}[T_i(N)]$$
$$\leq \sum_{i|a_i \neq a_{i^*}} \frac{32\mu_{\max}a_i^2\log(N)}{\Delta_i} + \left[1 + \frac{\pi^2}{6} + \zeta\left(\frac{10}{7}\right)\right]\sum_{i=1}^{K}\Delta_i,$$

which concludes the proof.

Appendix A.4. Proof of Theorem 5

Theorem 5. *If policy UCB-LM is run over a stationary MAB setting with a monotonic set A in which each arm $a_i \in A$ has outcome $X_{i,t}$ such that* $\mathbb{E}[X_{i,t}] =$

$\mu_i \leq \mu_{\max} \leq \frac{1}{2}$ *for each* t*, the expected regret after* N *rounds is at most:*

$$\overline{R}_N \leq \sum_{i|a_i \neq a_{i^*}} \frac{32\mu_{\max} a_i^2 \log(N)}{\Delta_i} + \sum_{i|a_i \neq a_{i^*}} \frac{8\mu_{\max} a_i^2 \log(K)}{\Delta_i} + \left[1 + \frac{\pi^2}{6} + \zeta\left(\frac{10}{7}\right)\right] \sum_{i=1}^{K} \Delta_i,$$

where $\zeta(\cdot)$ *is the Riemann zeta function.*

Proof. The proof is a straightforward combination of the arguments used for the UCB1-M and UCB-L ones. Consider the round of the learning process at which a specific arm a_i has been selected for s rounds and define:

- $\bar{j}(i,t) := \bar{j}$ (with abuse of notation) as the index $j \in \{1, \dots, i\}$ minimizing the quantity $\overline{x}_{ji,t} + \sqrt{\frac{2\mu_{\max}[4\log(t)+\log(i)]}{T_{ji}(t-1)}}$, i.e., the upper bound of arm a_i;
- $\bar{j}^* := \bar{j}(i^*, t)$ as the index $j \in \{1, \dots, i^*\}$ minimizing the quantity $\overline{x}_{ji^*,t} + \sqrt{\frac{2\mu_{\max}[4\log(t)+\log(i^*)]}{T_{ji^*}(t-1)}}$, i.e., the upper bound of arm a_{i^*};
- $\overline{X}_{i,(s)}$ is the unbiased estimate of μ_i in the case we collected a total of s samples from arm a_i;
- $\overline{X}_{\bar{j}i,(s)}$ is the unbiased estimate of $\mu_{\bar{j}i,t,s} = \mathbb{E}\left[\overline{X}_{\bar{j}i,(s)}\right]$, in the case we collected a total of s samples from arm a_i (and thus we have $s' \geq s$ samples to estimate $\mu_{\bar{j}i,s}$);
- $c_{i,t,s} := \sqrt{\frac{2\mu_{\max}[4\log(t)+\log(i)]}{s}}$ as the Hoeffding bound with confidence $\frac{t^{-4}}{i}$ for $\overline{X}_{i,(s)}$ after t rounds;
- $c_{ji,t,s} := \sqrt{\frac{2\mu_{\max}[4\log(t)+\log(i)]}{s'}}$ as the Hoeffding bound with confidence $\frac{t^{-4}}{i}$ for $\overline{X}_{ji,(s)}$ after t rounds, in the case arm a_i has been pulled a total of s times and the arms $\{a_j, \dots, a_i\}$ have been chosen in total $s' > s$ times.

We have that, for each $l > 0$:

$$T_i(N) \leq l + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=l}^{t-1} \mathbb{1}\left\{a_{i^*}\overline{X}_{\bar{j}^*i^*,(s)} + a_{i^*}c_{\bar{j}^*i^*,t,s} \leq a_i\overline{X}_{\bar{j}i,(s_i)} + a_i c_{\bar{j}i,t,s}\right\}$$

and consequently we only need to bound the probability of these three events:

$$a_{i^*}\overline{X}_{\bar{j}^*i^*,(s)} \leq a_{i^*}\mu_{i^*} - a_{i^*}c^{\bar{j}^*i^*,t,s} \tag{A.11}$$

$$a_i\overline{X}_{i,(s_i)} \geq a_i\mu_i + a_i c_{i,t,s_i} \tag{A.12}$$

$$a_{i^*}\mu_{i^*} - a_i\mu_i + a_i\overline{X}_{i,(s_i)} - a_i\overline{X}_{\bar{j}i,(s_i)} - a_i c_{\bar{j}i,t,s_i} - a_i c_{i,t,s_i} \leq 0. \tag{A.13}$$

Similarly to what has been done for Theorem 1, the probability of the event in Equation (A.11) can be bounded by t^{-4} by using the monotonicity assumption and Theorem 3, the one corresponding to the event in Equation (A.12) is bounded by $t^{\frac{24}{7}}$ by using the Chernoff theorem (Theorem 8, which considers the upper tail) and the event in Equation (A.13) is not possible if we choose $l = \left\lceil \frac{8a_i^2\mu_{\max}[4\log(t)+\log(i)]}{\Delta_i^2} \right\rceil$. Thus, by considering that $\log(t) \leq \log(N)$ and $\log(i) \leq \log(K)$, $\forall i$, we have:

$$\overline{R}_N = a_{i^*}\mu_{i^*}N - \sum_{i=1}^{K} \mathbb{E}[T_i(N)]a_i\mu_i = \sum_{i=1}^{K}(a_{i^*}\mu_{i^*} - a_i\mu_i)\mathbb{E}[T_i(N)]$$

$$\leq \sum_{i|a_i \neq a_{i^*}} \frac{32\mu_{\max}a_i^2\log(N)}{\Delta_i} + \sum_{i|a_i \neq a_{i^*}} \frac{8\mu_{\max}a_i^2\log(K)}{\Delta_i} + \left[1 + \frac{\pi^2}{6} + \zeta\left(\frac{10}{7}\right)\right]\sum_{i=1}^{K}\Delta_i,$$

which concludes the proof.

Appendix A.5. Proof of Theorem 7

Theorem 7. *If policy SW-UCB-M is run over a nonstationary MAB setting $\mathcal{S}^{(B)}$, for any $\tau \in \mathbb{N}$ and $\xi > \frac{1}{2}$, the expected regret after N rounds is at most:*

$$\overline{R}_N \leq \sum_{i=1}^{K}\left[\frac{N}{\tau}\frac{4a_i^2\xi[\log(i)+\log(\tau)]}{\Delta_i} + a_i\Upsilon_N\tau + \frac{2N}{\tau}\left\lceil\frac{\log(\tau)}{\log\left(1+4\sqrt{1-\frac{1}{2\xi}}\right)}\right\rceil\right],$$

where Υ_N is the number of breakpoints before N and

$$\Delta_i := \min_{\phi\in\{1,\ldots,\Upsilon_N\}}\left(a_{i^*_\phi}\mu_{i^*_\phi,\phi} - a_i\mu_{i,\phi}\right)\mathbb{1}\{i \neq i^*_\phi\} \quad \forall i \in \{1,\ldots,K\},$$

denotes the minimum, over all the phases Φ_ϕ in which the arm a_i is not optimal, of the difference of the expected reward $a_{i^*_\phi}\mu_{i^*_\phi,\phi}$ of the best arm $a_{i^*_\phi}$ and the expected reward $a_i\mu_{i,\phi}$ of the arm a_i.

Proof. Consider the phases $\phi \in \{1,\ldots,\Upsilon_N\}$ introduced in Section 2. Let us define:

$$A_{i,\phi}(\tau) = \frac{4a_i^2\xi[\log(i)+\log(\tau)]}{\Delta_{i,\phi}^2},$$

where $\Delta_{i,\phi} = a_{i^*_\phi}\mu_{i^*_\phi,\phi} - a_i\mu_{i,\phi}$, $\forall i \in \{1,\ldots,K\} \setminus \{i^*_\phi\}$.

Let us denote with $T_i(\Phi'_\phi)$ the number of times an arm a_i, with $i \in \{1,\ldots,K\} \setminus \{i^*_\phi\}$, has been played when it was not the best arm during the rounds $t \in \Phi'_\phi := \{t | b_{\phi-1} + \tau \leq t < b_\phi\}$. We consider $\tau < N_\phi$, i.e., τ is smaller than the number of rounds in each phase.[7]

[7]We make this assumption for ease of notation. In the case $\exists \tau > N_\phi$, it is straightforward to extend the analysis.

We can bound the number of times we are pulling an arm as:

$$T_i(N) = \sum_{\phi=1}^{\Upsilon_N} T_i(\Phi_\phi) \leq \sum_{\phi=1}^{\Upsilon_N} \tau + T_i(\Phi'_\phi)$$

where we assume that $\tau > K$.

Let us focus on a single phase Φ_ϕ. Consider the number of times a suboptimal arm $a_i \neq a_{i^*_\phi}$ has been pulled, we have:

$$\begin{aligned} T_i(\Phi'_\phi) &= \sum_{t \in \Phi'_\phi} \mathbb{1}\{i_t = i\} \\ &\leq \sum_{t \in \Phi'_\phi} \mathbb{1}\{i_t = i, T_i(t-1,\tau) < A_{i,\phi}(\tau)\} + \sum_{t \in \Phi'_\phi} \mathbb{1}\{i_t = i, T_i(t-1,\tau) \geq A_{i,\phi}(\tau)\} \end{aligned} \tag{A.14}$$

where i_t is the index of the arm a_{i_t} selected at round t by policy SW-UCB-M with a window of size τ.

By using Lemma 25 in [27], we can bound the first term of Equation (A.14), we have:

$$\begin{aligned} T_i(\Phi'_\phi) &\leq \left\lceil \frac{|\Phi'_\phi|}{\tau} \right\rceil A_{i,\phi}(\tau) + \sum_{t \in \Phi'_\phi} \mathbb{1}\{i_t = i, T_i(t-1,\tau) \geq A_{i,\phi}(\tau)\} \\ &\leq \left\lceil \frac{N_\phi - \tau}{\tau} \right\rceil A_{i,\phi}(\tau) + \sum_{t \in \Phi'_\phi} \mathbb{1}\{i_t = i, T_i(t-1,\tau) \geq A_{i,\phi}(\tau)\} \\ &\leq \frac{N_\phi}{\tau} A_{i,\phi}(\tau) + \sum_{t \in \Phi'_\phi} \mathbb{1}\{i_t = i, T_i(t-1,\tau) \geq A_{i,\phi}(\tau)\}. \end{aligned} \tag{A.15}$$

Let us focus on the second term of the last expression. The event $i_t = i$ occurs when:

$$\begin{aligned} &a_{i^*_\phi}\overline{X}_{\bar{j}^* i^*_\phi,t,\tau} + a_{i^*_\phi}\varepsilon_{i^*_\phi,t,T_{\bar{j}^* i^*}(t-1),\tau} \leq a_i\overline{X}_{\bar{j}i,t,\tau} + a_i\varepsilon_{i,t,T_{\bar{j}i}(t-1),\tau} \\ &a_{i^*_\phi}\overline{X}_{\bar{j}^* i^*_\phi,t,\tau} - a_{i^*_\phi}\mu_{i^*_\phi,\phi} + a_{i^*_\phi}\varepsilon_{i^*_\phi,t,T_{\bar{j}^* i^*}(t-1),\tau} - a_i\overline{X}_{i,t,\tau} + a_i\mu_{i,\phi} + a_i\varepsilon_{i,t,T_i(t-1),\tau} + \\ &+ a_{i^*_\phi}\mu_{i^*_\phi,\phi} - a_i\mu_{i,\phi} + a_i\overline{X}_{i,t,\tau} - a_i\overline{X}_{\bar{j}i,t,\tau} - a_i\varepsilon_{i,t,T_i\bar{j}i(t-1),\tau} - a_i\varepsilon_{i,t,T_i(t-1),\tau} \end{aligned}$$

where $\varepsilon_{i,t,T_{ji}(t-1),\tau} := \sqrt{\dfrac{\xi[\log(i) + \log(\min\{t,\tau\})]}{T_{ji}(t-1)}} = \sqrt{\dfrac{\xi[\log(i) + \log(\tau)]}{T_{ji}(t-1)}}$, since $t \in \Phi'_\phi \Rightarrow t > \tau$ and it is contained in the union of the following three events:

$$a_{i^*_\phi}\overline{X}_{\bar{j}^* i^*_\phi,t,\tau} \leq a_{i^*_\phi}\mu_{i^*_\phi,\phi} - a_{i^*_\phi}\varepsilon_{i^*_\phi,t,T_{\bar{j}^* i^*}(t-1),\tau}; \tag{A.16}$$

$$a_i\overline{X}_{i,t,\tau} \geq a_i\mu_{i,\phi} + a_i\varepsilon_{i,t,T_i(t-1),\tau}; \tag{A.17}$$

$$a_{i^*_\phi}\mu_{i^*_\phi,\phi} - a_i\mu_{i,\phi} + a_i\overline{X}_{i,t,\tau} - a_i\overline{X}_{\bar{j}i,t,\tau} - a_i\varepsilon_{i,t,T_{\bar{j}i}(t-1),\tau} - a_i\varepsilon_{i,t,T_i(t-1),\tau} \leq 0;. \tag{A.18}$$

Let us define $\delta = \varepsilon_{i,t,T_i(t-1),\tau}\sqrt{T_i(t-1,\tau)} = \sqrt{\xi[\log(i)+\log(\tau)]}$ and consider the probability of the event in Equation (A.17), we have:

$$\mathbb{P}\left(a_i \overline{X}_{i,t,\tau} \geq a_i \mu_{i,\phi} + a_i \frac{\delta}{\sqrt{T_i(t-1,\tau)}}\right) = \mathbb{P}\left(\overline{X}_{i,t,\tau} - \mu_{i,\phi} \geq \frac{\delta}{\sqrt{T_i(t-1,\tau)}}\right)$$
$$\leq \mathbb{P}\left(\overline{X}_{i,t,\tau} - \mu_{i,\phi} \geq \frac{\delta}{\sqrt{T_i(t-1,\tau)}}\right) = \mathbb{P}\left(\frac{T_i(t-1,\tau)\left(\overline{X}_{i,t,\tau} - \mu_{i,\phi}\right)}{\sqrt{T_i(t-1,\tau)}} \geq \delta\right).$$

By applying Corollary 21 in [27] we have that for all $\eta > 0$:

$$\mathbb{P}\left(\frac{T_i(t-1,\tau)\left(\overline{X}_{i,t,\tau} - \mu_{i,\phi}\right)}{\sqrt{T_i(t-1,\tau)}} \geq \delta\right) \leq \left\lceil \frac{\log(\tau)}{\log(1+\eta)} \right\rceil \exp\left(-2\delta^2\left(1-\frac{\eta^2}{16}\right)\right)$$
$$\leq \left\lceil \frac{\log(\tau)}{\log(1+\eta)} \right\rceil \exp\left(-2\xi[\log(i)+\log(\tau)]\left(1-\frac{\eta^2}{16}\right)\right)$$
$$= \left\lceil \frac{\log(\tau)}{\log(1+\eta)} \right\rceil (i\tau)^{-2\xi\left(1-\frac{\eta^2}{16}\right)}$$

where we consider the events of choosing arms a_i as the sequence of previsible variables.

Similarly, by exploiting the monotonicity property, we have that for each $j \leq i^*_\phi$ and by defining $\delta = \varepsilon_{i,t,T_{ji^*_\phi}(t-1),\tau}\sqrt{T_{ji^*_\phi}(t-1,\tau)}$:

$$\mathbb{P}\left(a_{i^*_\phi}\overline{X}_{ji^*_\phi,t,\tau} \leq a_{i^*_\phi}\mu_{i^*_\phi,\phi} - a_{i^*_\phi}\varepsilon_{i^*_\phi,t,T_{ji^*}(t-1),\tau}\right)$$
$$\leq \mathbb{P}\left(a_{i^*_\phi}\overline{X}_{ji^*_\phi,t,\tau} \leq a_{i^*_\phi}\mu_{ji^*_\phi,\phi} - a_{i^*_\phi}\varepsilon_{i^*_\phi,t,T_{ji^*}(t-1),\tau}\right)$$
$$= \mathbb{P}\left(a_{i^*_\phi}\overline{X}_{ji^*_\phi,t,\tau} \geq a_{i^*_\phi}\mu_{ji^*_\phi,\phi} + a_{i^*_\phi}\varepsilon_{i^*_\phi,t,T_{ji^*}(t-1),\tau}\right)$$
$$= \mathbb{P}\left(\overline{X}_{ji^*_\phi,t,\tau} - \mu_{ji^*_\phi,\phi} \geq \varepsilon_{i^*_\phi,t,T_{ji^*}(t-1),\tau}\right)$$
$$\left\lceil \frac{\log(\tau)}{\log(1+\eta)} \right\rceil \exp\left(-2\delta^2\left(1-\frac{\eta^2}{16}\right)\right)$$
$$\leq \left\lceil \frac{\log(\tau)}{\log(1+\eta)} \right\rceil \exp\left(-2\xi[\log(i)+\log(\tau)]\left(1-\frac{\eta^2}{16}\right)\right)$$
$$= \left\lceil \frac{\log(\tau)}{\log(1+\eta)} \right\rceil (i\tau)^{-2\xi\left(1-\frac{\eta^2}{16}\right)}$$
$$= \left\lceil \frac{\log(\tau)}{\log(1+\eta)} \right\rceil (\tau)^{-2\xi\left(1-\frac{\eta^2}{16}\right)},$$

where first equality sign is due to the symmetry of the Bernoulli distribution, the event of choosing an arm among the set $\{a_j, \dots a_{i^*_\phi}\}$ has been chosen as the sequence of previsible Bernoulli variables.

Thus, the probability of the event in Equation (A.16) can be bounded by:

$$\begin{aligned}
&\mathbb{P}\left(a_{i^*_\phi}\overline{X}_{\bar{j}^* i^*_\phi,t,\tau} \leq a_{i^*_\phi}\mu_{i^*_\phi,\phi} - a_{i^*_\phi}\varepsilon_{i^*_\phi,t,T_{\bar{j}^* i^*}(t-1),\tau}\right) \\
&= \left\lceil \frac{\log(\tau)}{\log(1+\eta)} \right\rceil (\tau)^{-2\xi\left(1-\frac{\eta^2}{16}\right)},
\end{aligned}$$

by resorting to an union bound over all $j \leq i$.

Finally, consider the event in Equation (A.18) and that $T_i(t-1,\tau) \geq A_{i,\phi}(\tau)$:

$$\begin{aligned}
0 &\geq \Delta_{i,\phi} + \underbrace{a_i\overline{X}_{i,t,\tau} - a_i\overline{X}_{\bar{j}i,t,\tau} - a_i\varepsilon_{i,t,T_{\bar{j}i}(t-1),\tau}}_{\geq -a_i\varepsilon_{i,t,T_i(t-1),\tau}} - a_i\varepsilon_{i,t,T_i(t-1),\tau} \\
&\geq \Delta_{i,\phi} - 2a_i\varepsilon_{i,t,T_i(t-1),\tau} > 0;
\end{aligned}$$

where the inequality is given from the fact that the SW-UCB-M algorithm chooses the tightest bound among the $a_i\overline{X}_{ji,t,\tau} + a_i\varepsilon_{i,t,T_{ji}(t-1),\tau}$ with $1 \leq j \leq i$. Since the last expression is a contradiction, the considered event does not occur.

By choosing $\eta = 4\sqrt{1-\frac{1}{2\xi}}$ we have $2\xi\left(1-\frac{\eta^2}{16}\right) = 1$ and we get:

$$\begin{aligned}
\mathbb{E}[T_i(\Phi'_\phi)] &\leq \frac{N_\phi}{\tau}A_{i,\phi}(\tau) + 2\sum_{t\in\Phi'_\phi} \frac{\left\lceil \frac{\log(\tau)}{\log(1+\eta)} \right\rceil}{\tau} \\
&= \frac{N_\phi}{\tau}A_{i,\phi}(\tau) + \frac{2|\Phi'_\phi|}{\tau}\frac{\log(\tau)}{\log(1+\eta)} \\
&\leq \frac{N_\phi}{\tau}\frac{4a_i^2\xi[\log(i)+\log(\tau)]}{\Delta_{i,\phi}^2} + \frac{2N_\phi}{\tau}\left\lceil \frac{\log(\tau)}{\log\left(1+4\sqrt{1-\frac{1}{2\xi}}\right)} \right\rceil
\end{aligned}$$

The total regret becomes:

$$\begin{aligned}
\overline{R}_N &= \sum_{\phi=1}^{\Upsilon_N}\left(a_{i^*,\phi}\mu_{i^*,\phi}N_\phi - \sum_{i=1}^{K}a_i\mu_{i,\phi}\mathbb{E}[T_i(\Phi_\phi)]\right) = \sum_{\phi=1}^{\Upsilon_N}\left(\sum_{i=1}^{K}(a_{i^*,\phi}\mu_{i^*,\phi} - a_i\mu_{i,\phi})\mathbb{E}[T_i(\Phi_\phi)]\right) \\
&= \sum_{i=1}^{K}\left(\sum_{\phi=1}^{\Upsilon_N}(a_{i^*,\phi}\mu_{i^*,\phi} - a_i\mu_{i,\phi})\mathbb{E}[T_i(\Phi_\phi)]\right) \\
&\leq \sum_{i=1}^{K}\left(\sum_{\phi=1}^{\Upsilon_N}\Delta_{i,\phi}\mathbb{E}[T_i(\Phi_\phi)]\right) \leq \sum_{i=1}^{K}\left[\sum_{\phi=1}^{\Upsilon_N}\Delta_{i,\phi}\left(\tau + \mathbb{E}[T_i(\Phi_\phi)]\right)\right] \\
&\leq \sum_{i=1}^{K}\left[a_i\Upsilon_N\tau + \sum_{\phi=1}^{\Upsilon_N}\Delta_{i,\phi}\left(\frac{N_\phi}{\tau}\frac{4a_i^2\xi[\log(i)+\log(\tau)]}{\Delta_{i,\phi}^2} + \frac{2N_\phi}{\tau}\left\lceil \frac{\log(\tau)}{\log\left(1+4\sqrt{1-\frac{1}{2\xi}}\right)} \right\rceil\right)\right]
\end{aligned}$$

Considering Δ_i as defined in in the theorem statement, we obtain:

$$\overline{R}_N \leq \sum_{i=1}^{K} \left[\frac{N}{\tau} \frac{4a_i^2 \xi [\log(i) + \log(\tau)]}{\Delta_i} + a_i \Upsilon_N \tau + \frac{2N}{\tau} \left[\frac{\log(\tau)}{\log\left(1 + 4\sqrt{1 - \frac{1}{2\xi}}\right)} \right] \right],$$

which concludes the proof.

Appendix B. Sliding Window Algorithms

In this section, we report the algorithm used in the experimental analysis of the nonstationary case. While the algorithm SW-UCB has been proposed in [27] and is used here as baseline, the other presented algorithms are the straightforward application of the developed bounds in the sliding windows paradigm.

We recall that the expected value of the outcome μ_i over the last $\min\{\tau, t\}$ rounds is:

$$\overline{X}_{i,t,\tau} = \frac{1}{T_i(t-1,\tau)} \sum_{s=T_i(\max\{t-\tau,1\})}^{T_i(t-1)} X_{i,s},$$

where $T_i(t,\tau) = T_i(t) - T_i(\max\{t-\tau+1, 1\})$ is the number of rounds the arm a_i has been selected in the last $\min\{\tau, t\}$ ones and its realization is:

$$\overline{x}_{i,t,\tau} = \frac{1}{T_i(t-1,\tau)} \sum_{s=T_i(\max\{t-\tau,1\})}^{T_i(t-1)} x_{i,s}.$$

Moreover, we recall that $\overline{X}_{ji,t,\tau}$ is the following convex linear combination of the sample means $\overline{X}_j, \ldots, \overline{X}_i$:

$$\overline{X}_{ji,t,\tau} = \frac{1}{T_{ji}(t-1,\tau)} \sum_{k=j}^{i} \sum_{s=T_k(\max\{t-\tau,1\})}^{T_k(t-1)} X_{k,s},$$

where $T_{ji}(t,\tau) = \sum_{k=j}^{i} T_k(t-1) - T_k(\max\{t-\tau, 1\})$ is the number of rounds one of the arms in $\{a_j, \ldots, a_i\}$ has been selected in the last $\min\{\tau, t\}$ ones and the realization of $\overline{X}_{ji,t,\tau}$ is denoted as follows:

$$\overline{x}_{ji,t,\tau} = \frac{1}{T_{ji}(t-1,\tau)} \sum_{k=j}^{i} \sum_{s=T_k(\max\{t-\tau,1\})}^{T_k(t-1)} x_{k,s}.$$

At last, the variances $\overline{V}_{i,t,\tau}$ and $\overline{V}_{ji,t,\tau}$ of the two aforementioned random variables $\overline{X}_{i,t,\tau}$ and $\overline{X}_{ji,t,\tau}$ is:

$$\overline{V}_{i,t,\tau} = \frac{\sum_{s=T_i(\max\{t-\tau,1\})}^{T_i(t-1)} \left(X_{i,s} - \overline{X}_{i,t,\tau}\right)^2}{T_i(t,\tau)}$$

$$\overline{V}_{ji,t,\tau} = \frac{\sum_{s=T_i(\max\{t-\tau,1\})}^{T_i(t-1)} \left(X_{k,s} - \overline{X}_{ji,t,\tau}\right)^2}{T_i(t,\tau)},$$

respectively, and their realizations $\overline{v}_{i,t,\tau}$ and $\overline{v}_{ji,t,\tau}$:

$$\overline{v}_{i,t,\tau} = \frac{\sum_{s=T_i(\max\{t-\tau,1\})}^{T_i(t-1)} \left(x_{i,s} - \overline{x}_{i,t,\tau}\right)^2}{T_{ji}(t-1,\tau)}$$

$$\overline{v}_{ji,t,\tau} = \frac{\sum_{s=T_i(\max\{t-\tau,1\})}^{T_i(t-1)} \left(x_{k,s} - \overline{x}_{ji,t,\tau}\right)^2}{T_{ji}(t-1,\tau)},$$

respectively.

In what follows, the algorithms derived from the bound in [27] consider a parameter $\xi > 0$. For ease of comparison with [27], in the experimental section we set it to $\xi = 0.6$.

Appendix B.1. SW-UCB

ALGORITHM 6: SW-UCB

Initialization
Input: ξ
for $t \in \{1, \dots, K\}$ **do**
  Play arm a_t and observe $x_{t,1}$
Loop
for $t \in \{K+1, \dots, N\}$ **do**
  for $i \in \{1, \dots, K\}$ **do**
    Compute:

$$u_{i,t}^{(\text{SW-UCB})} = \overline{x}_{i,t,\tau} + \sqrt{\frac{\xi \log(\min\{t,\tau\})}{T_i(t-1,\tau)}}$$

  Play arm a_{i_t} such that $i_t = \arg\max_{i\in\{1,\dots,K\}} a_i u_{i,t}^{(\text{SW-UCB})}$ and observe $x_{i_t,T_{i_t}(t)}$

Appendix B.2. SW-UCB1-M

ALGORITHM 7: SW–UCB1–M

Initialization
for $t \in \{1, \dots, K\}$ **do**
  Play arm a_t and observe $x_{t,1}$
Loop
for $t \in \{K+1, \dots, N\}$ **do**
  for $i \in \{1, \dots, K\}$ **do**
    Compute:

$$u_{i,t}^{(\text{SW–UCB1–M})} = \min_{j \in \{1,\dots,i\}} \left\{ \overline{x}_{ji,t,\tau} + \sqrt{\frac{4\log(\min\{t,\tau\}) + \log(i)}{2T_{ji}(t-1,\tau)}} \right\}$$

  Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\dots,K\}} a_i u_{i,t}^{(\text{SW–UCB1–M})}$and observe $x_{i_t, T_{i_t}(t)}$

Appendix B.3. SW-UCB-L

ALGORITHM 8: SW-UCB-L

Initialization
Input: $\mu_{\max}$
for $t \in \{1, \dots, K\}$ **do**
  Play arm a_t and observe $x_{t,1}$
Loop
for $t \in \{K+1, \dots, N\}$ **do**
  for $i \in \{1, \dots, K\}$ **do**
    Compute:

$$u_{i,t}^{(\text{SW-UCB-L})} = \overline{x}_{i,t,\tau} + \sqrt{\frac{8\mu_{\max}\log(\min\{t,\tau\})}{T_i(t-1,\tau)}}$$

  Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\dots,K\}} a_i u_{i,t}^{(\text{SW-UCB-L})}$ and observe $x_{i_t, T_{i_t}(t)}$

Appendix B.4. SW-UCB-LM

ALGORITHM 9: SW–UCB–LM

Initialization
Input: $\mu_{\max}$
for $t \in \{1, \ldots, K\}$ **do**
  Play arm a_t and observe $x_{t,1}$
Loop
for $t \in \{K+1, \ldots, N\}$ **do**
  for $i \in \{1, \ldots, K\}$ **do**
    Compute:

$$u_{i,t}^{(\text{SW-UCB-LM})} = \min_{j \in \{1,\ldots,i\}} \left\{ \overline{x}_{ji,t,\tau} + \sqrt{\frac{2\mu_{\max}[\log(\min\{t,\tau\}) + \log(i)]}{T_{ji}(t-1,\tau)}} \right\}$$

  Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\ldots,K\}} a_i u_{i,t}^{(\text{SW-UCB-LM})}$ and observe $x_{i_t,T_{i_t}(t)}$

Appendix B.5. SW-UCBV

ALGORITHM 10: SW-UCBV

Initialization
Input: ξ, c
for $t \in \{1, \ldots, K\}$ **do**
  Play arm a_t and observe $x_{t,1}$
Loop
for $t \in \{K+1, \ldots, N\}$ **do**
  for $i \in \{1, \ldots, K\}$ **do**
    Compute:

$$u_{i,t}^{(\text{SW-UCBV})} = \overline{x}_{i,t,\tau} + \sqrt{\frac{2\overline{v}_{i,t,\tau}\xi \log(\min\{t,\tau\})}{T_i(t-1,\tau)}} + \\ + \frac{3c\xi \log(\min\{t,\tau\})}{T_i(t-1,\tau)}$$

  Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\ldots,K\}} a_i u_{i,t}^{(\text{SW-UCBV})}$ and observe $x_{i_t,T_{i_t}(t)}$

Appendix B.6. SW-UCBV-M

ALGORITHM 11: SW-UCBV-M

Initialization

Input: ξ, c

for $t \in \{1, \ldots, K\}$ **do**

Play arm a_t and observe $x_{t,1}$

Loop

for $t \in \{K+1, \ldots, N\}$ **do**

for $i \in \{1, \ldots, K\}$ **do**

Compute:

$$u_{i,t}^{(\text{SW-UCBV-M})} = \min_{j \in \{1,\ldots,i\}} \left\{ \overline{x}_{ji,t,\tau} + \sqrt{\frac{2\overline{v}_{ji,t,\tau}[\xi \log(\min\{t,\tau\}) + \log(i)]}{T_{ji}(t-1,\tau)}} + \frac{3c[\xi \log(\min\{t,\tau\}) + \log(i)]}{T_{ji}(t-1,\tau)} \right\}$$

Play arm a_{i_t} such that $i_t = \arg\max_{i \in \{1,\ldots,K\}} a_i u_{i,t}^{(\text{SW-UCBV-M})}$ and observe $x_{i_t, T_{i_t}(t)}$

Appendix C. Experimental Results (Detailed Tables)

Table C.4: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_L$ and $\mu_{\max} = 1$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds							
		4	5	6	7	8	10	11	17
5	UCB1M	0.114	0.112	0.103	0.095	0.086	0.074	0.069	0.049
	UCBV	0.407	0.402	0.375	0.345	0.317	0.273	0.254	0.179
	UCBV-M	0.423	0.417	0.388	0.356	0.327	0.281	0.261	0.184
9	UCB1M	0.277	0.293	0.291	0.282	0.271	0.243	0.231	0.169
	UCBV	0.372	0.419	0.436	0.439	0.435	0.406	0.392	0.308
	UCBV-M	0.506	0.535	0.539	0.529	0.515	0.471	0.451	0.346
17	UCB1M	0.416	0.472	0.500	0.503	0.507	0.490	0.479	0.392
	UCBV	0.297	0.360	0.403	0.418	0.437	0.449	0.448	0.402
	UCBV-M	0.623	0.693	0.722	0.720	0.719	0.686	0.668	0.549
33	UCB1M	0.538	0.608	0.651	0.694	0.714	0.731	0.736	0.699
	UCBV	0.252	0.305	0.339	0.380	0.405	0.444	0.460	0.490
	UCBV-M	**0.741**	**0.823**	**0.871**	**0.912**	**0.933**	**0.944**	**0.946**	**0.868**

Table C.5: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_L$ and $\mu_{\max} = 10^{-1}$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds																		
		46	67	72	98	130	149	177	225	231	268	296	346	362	467	551	587	737	1,211	2,846
5	UCB1-M	0.346	0.343	0.340	0.316	0.297	0.292	0.286	0.280	0.280	0.268	0.262	0.250	0.246	0.220	0.199	0.192	0.162	0.106	0.047
	UCB-L	0.180	0.234	0.243	0.295	0.347	0.376	0.408	0.449	0.454	0.466	0.474	0.478	0.477	0.459	0.433	0.421	0.371	0.260	0.125
	UCB-LM	0.502	0.554	0.566	0.623	0.667	0.683	0.701	0.707	0.709	0.699	0.692	0.669	0.660	0.604	0.555	0.535	0.460	0.311	0.145
	UCBV	2.027	2.108	2.111	2.075	1.984	1.919	1.831	1.676	1.658	1.548	1.476	1.355	1.318	1.123	0.991	0.943	0.776	0.494	0.219
	UCBV-M	**2.111**	**2.170**	**2.168**	**2.115**	**2.014**	**1.947**	1.855	1.694	1.676	1.563	1.489	1.365	1.327	1.130	0.997	0.948	0.781	0.496	0.220
9	UCB1-M	0.508	0.545	0.550	0.565	0.575	0.578	0.579	0.580	0.578	0.571	0.564	0.549	0.543	0.508	0.481	0.470	0.427	0.320	0.161
	UCB-L	0.103	0.129	0.136	0.167	0.201	0.220	0.243	0.276	0.279	0.298	0.312	0.333	0.337	0.366	0.377	0.380	0.384	0.346	0.217
	UCB-LM	0.718	0.772	0.780	0.821	0.845	0.855	0.861	0.866	0.865	0.860	0.853	0.840	0.836	0.797	0.764	0.750	0.693	0.541	0.294
	UCBV	1.426	1.583	1.607	1.683	1.714	1.717	1.703	1.665	1.659	1.616	1.583	1.523	1.504	1.383	1.294	1.258	1.126	0.829	0.423
	UCB-VM	1.783	1.861	1.868	1.885	1.871	1.853	1.817	1.753	1.744	1.690	1.649	1.578	1.556	1.421	1.325	1.287	1.148	0.840	0.427
17	UCB1-M	0.598	0.651	0.662	0.702	0.738	0.753	0.769	0.784	0.786	0.794	0.795	0.797	0.797	0.790	0.777	0.772	0.746	0.648	0.398
	UCB-L	0.071	0.088	0.092	0.112	0.132	0.144	0.160	0.183	0.186	0.202	0.212	0.230	0.236	0.266	0.285	0.293	0.319	0.353	0.298
	UCB-LM	0.810	0.894	0.909	0.963	1.013	1.035	1.060	1.077	1.078	1.088	1.090	1.088	1.087	1.066	1.042	1.031	0.983	0.838	0.525
	UCBV	1.043	1.202	1.231	1.356	1.451	1.487	1.525	1.552	1.555	1.563	1.559	1.545	1.539	1.494	1.448	1.429	1.351	1.127	0.687
	UCBV-M	1.757	1.833	1.840	1.867	1.872	1.867	1.855	1.816	1.812	1.786	1.760	1.716	1.703	1.620	1.554	1.527	1.426	1.169	0.701
33	UCB1-M	0.651	0.719	0.732	0.787	0.836	0.857	0.885	0.919	0.923	0.944	0.957	0.973	0.977	0.992	0.999	0.999	0.995	0.945	0.733
	UCB-L	0.049	0.060	0.062	0.073	0.086	0.093	0.102	0.117	0.119	0.129	0.137	0.150	0.154	0.176	0.194	0.200	0.223	0.279	0.335
	UCB-LM	0.867	0.967	0.985	1.059	1.120	1.147	1.179	1.217	1.223	1.242	1.253	1.266	1.270	1.277	1.274	1.272	1.251	1.162	0.871
	UCBV	0.778	0.904	0.930	1.048	1.155	1.205	1.268	1.346	1.354	1.394	1.418	1.449	1.456	1.479	1.485	1.482	1.462	1.354	1.007
	UCBV-M	1.747	1.847	1.859	1.906	1.933	1.939	**1.941**	**1.934**	**1.932**	**1.920**	**1.908**	**1.884**	**1.876**	**1.820**	**1.779**	**1.760**	**1.687**	**1.489**	**1.058**

Table C.6: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_L$ and $\mu_{\max} = 10^{-2}$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds												
		1,729	1,842	3,405	4,289	4,439	5,508	5,877	8,471	9,434	9,997	9,998	9,999	10,000
5	UCB1-M	0.363	0.366	0.382	0.385	0.386	0.387	0.389	0.371	0.363	0.358	0.358	0.358	0.358
	UCB-L	0.917	0.955	1.387	1.548	1.571	1.701	1.736	1.876	1.900	1.907	1.907	1.907	1.908
	UCB-LM	1.380	1.415	1.773	1.889	1.905	1.988	2.010	2.080	2.085	2.083	2.083	2.083	2.083
	UCBV	3.143	3.144	3.079	3.025	3.014	2.943	2.920	2.755	2.700	2.666	2.666	2.666	2.666
	UCBV-M	**3.177**	**3.175**	**3.098**	**3.039**	**3.029**	**2.955**	**2.931**	**2.763**	**2.707**	**2.673**	**2.673**	**2.673**	**2.673**
9	UCB1-M	0.471	0.475	0.502	0.519	0.522	0.539	0.544	0.570	0.576	0.578	0.578	0.578	0.578
	UCB-L	0.512	0.533	0.786	0.898	0.916	1.032	1.065	1.252	1.302	1.330	1.330	1.330	1.330
	UCB-LM	1.331	1.350	1.538	1.606	1.615	1.677	1.691	1.770	1.787	1.795	1.795	1.794	1.794
	UCBV	2.407	2.428	2.545	2.555	2.555	2.545	2.538	2.481	2.458	2.444	2.444	2.444	2.444
	UCBV-M	2.584	2.595	2.638	2.629	2.626	2.604	2.594	2.521	2.493	2.477	2.477	2.477	2.477
17	UCB1-M	0.528	0.534	0.586	0.608	0.612	0.634	0.641	0.677	0.688	0.694	0.694	0.694	0.694
	UCB-L	0.322	0.335	0.486	0.560	0.572	0.649	0.672	0.828	0.878	0.905	0.905	0.905	0.905
	UCB-LM	1.370	1.390	1.571	1.637	1.645	1.694	1.708	1.764	1.777	1.783	1.783	1.783	1.784
	UCBV	1.910	1.940	2.164	2.220	2.227	2.259	2.267	2.286	2.284	2.282	2.282	2.282	2.282
	UCBV-M	2.356	2.366	2.418	2.425	2.426	2.420	2.419	2.393	2.381	2.375	2.375	2.375	2.375
33	UCB1-M	0.563	0.570	0.633	0.660	0.664	0.693	0.701	0.752	0.767	0.775	0.775	0.776	0.776
	UCB-L	0.205	0.213	0.304	0.348	0.355	0.405	0.421	0.526	0.561	0.581	0.581	0.581	0.581
	UCB-LM	1.402	1.421	1.603	1.668	1.676	1.732	1.748	1.825	1.844	1.853	1.853	1.853	1.853
	UCBV	1.448	1.481	1.789	1.885	1.899	1.977	1.997	2.092	2.113	2.122	2.122	2.122	2.122
	UCBV-M	2.314	2.321	2.382	2.392	2.393	2.396	2.396	2.384	2.377	2.373	2.373	2.373	2.373

Table C.7: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_L$ and $\mu_{\max} = 10^{-3}$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds			
		9,989	9,992	9,998	10,000
5	UCB1-M	0.345	0.345	0.345	0.345
	UCB-L	0.670	0.670	0.670	0.670
	UCB-LM	1.146	1.146	1.146	1.146
	UCBV	3.280	3.280	3.280	3.280
	UCBV-M	**3.374**	**3.374**	**3.374**	**3.374**
9	UCB1-M	0.405	0.405	0.405	0.405
	UCB-L	0.391	0.391	0.391	0.391
	UCB-LM	1.211	1.211	1.211	1.211
	UCBV	2.259	2.259	2.260	2.260
	UCBV-M	2.608	2.608	2.608	2.609
17	UCB1-M	0.428	0.428	0.428	0.428
	UCB-L	0.251	0.251	0.252	0.252
	UCB-LM	1.234	1.235	1.235	1.235
	UCBV	1.606	1.606	1.606	1.607
	UCBV-M	2.376	2.376	2.376	2.376
33	UCB1-M	0.437	0.437	0.437	0.437
	UCB-L	0.165	0.165	0.165	0.165
	UCB-LM	1.234	1.234	1.234	1.234
	UCBV	1.152	1.152	1.153	1.153
	UCBV-M	2.267	2.268	2.268	2.268

Table C.8: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_L$ and $\mu_{\max} = 10^{-4}$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds						
		9,893	9,979	9,981	9,991	9,997	9,998	10,000
5	UCB1-M	0.344	0.344	0.344	0.344	0.344	0.344	0.344
	UCB-L	0.157	0.158	0.158	0.158	0.158	0.158	0.158
	UCB-LM	0.580	0.582	0.582	0.582	0.583	0.583	0.583
	UCBV	1.348	1.353	1.353	1.354	1.354	1.354	1.354
	UCBV-M	**1.869**	**1.871**	**1.872**	**1.872**	**1.872**	**1.872**	**1.872**
9	UCB1-M	0.392	0.392	0.392	0.392	0.392	0.392	0.392
	UCB-L	0.107	0.107	0.107	0.107	0.107	0.107	0.107
	UCB-LM	0.632	0.633	0.633	0.633	0.633	0.633	0.633
	UCBV	0.910	0.912	0.912	0.913	0.913	0.913	0.913
	UCBV-M	1.701	1.703	1.703	1.703	1.703	1.703	1.703
17	UCB1-M	0.406	0.406	0.406	0.406	0.406	0.406	0.406
	UCB-L	0.074	0.074	0.074	0.074	0.074	0.074	0.074
	UCB-LM	0.668	0.670	0.670	0.670	0.671	0.671	0.671
	UCBV	0.706	0.707	0.707	0.707	0.707	0.707	0.707
	UCBV-M	1.614	1.618	1.618	1.619	1.619	1.619	1.619
33	UCB1-M	0.409	0.409	0.409	0.409	0.409	0.409	0.409
	UCB-L	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	UCB-LM	0.678	0.678	0.678	0.678	0.678	0.678	0.678
	UCBV	0.590	0.591	0.591	0.591	0.591	0.591	0.591
	UCBV-M	1.554	1.556	1.556	1.557	1.557	1.557	1.557

Table C.9: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_H$ and $\mu_{\max} = 1$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds									
		2	5	9	23	24	69	96	146	272	303
5	UCB1-M	0.000	0.001	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000
	UCBV	−0.001	**0.001**	**0.002**	0.003	0.003	0.002	0.002	0.002	0.001	0.001
	UCBV-M	−0.000	0.000	0.002	0.003	0.003	0.002	0.002	0.002	0.001	0.001
9	UCB1-M	−0.000	0.000	0.001	−0.000	−0.000	−0.000	−0.000	−0.000	0.000	0.000
	UCBV	−0.000	−0.005	−0.000	**0.003**	**0.003**	**0.004**	**0.003**	**0.003**	**0.003**	**0.003**
	UCBV-M	−0.000	−0.005	−0.001	0.002	0.002	0.003	0.003	0.003	0.003	0.002
17	UCB1-M	0.001	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000
	UCBV	0.000	−0.001	−0.003	0.002	0.002	0.002	0.003	0.003	0.002	0.002
	UCBV-M	0.001	−0.000	−0.003	0.001	0.001	0.002	0.002	0.002	0.002	0.002
33	UCB1-M	**0.002**	−0.001	−0.001	−0.000	−0.000	0.001	0.000	0.000	0.000	0.000
	UCBV	0.001	−0.001	−0.001	−0.001	−0.001	0.001	0.002	0.003	0.003	0.003
	UCBV-M	0.001	−0.003	−0.002	−0.000	−0.000	0.001	0.001	0.002	0.002	0.002

Table C.10: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_H$ and $\mu_{\max} = 10^{-1}$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds																	
		21	37	40	42	43	57	74	78	88	91	101	105	128	142	151	421	595	940
5	UCB1-M	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.000
	UCB-L	0.009	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.007	0.005	0.004	0.004
	UCB-LM	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.005	0.005	0.004
	UCBV	0.003	0.006	0.006	0.007	0.007	0.007	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.007	0.007	0.007
	UCBV-M	0.007	0.008	0.008	0.008	0.008	0.009	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.008	0.007	0.007	0.007
9	UCB1-M	0.005	0.007	0.007	0.008	0.008	0.007	0.007	0.007	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.005	0.004	0.003
	UCB-L	0.011	0.012	0.012	0.012	0.012	0.011	0.012	0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.010	0.009	0.008	0.007
	UCB-LM	0.018	0.018	0.018	0.018	0.018	0.018	0.017	0.017	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.011	0.009	0.008
	UCBV	−0.004	0.004	0.005	0.005	0.006	0.008	0.010	0.010	0.011	0.012	0.013	0.013	0.013	0.013	0.014	0.015	0.014	0.014
	UCBVM	0.012	0.016	0.016	0.016	0.017	0.017	0.017	0.017	0.018	0.018	0.018	0.018	0.017	0.017	0.017	0.016	0.015	0.014
17	UCB1-M	0.009	0.013	0.014	0.013	0.014	0.015	0.015	0.016	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.012	0.010	0.009
	UCB-L	0.010	0.011	0.011	0.011	0.011	0.012	0.012	0.012	0.013	0.013	0.012	0.012	0.012	0.012	0.012	0.010	0.010	0.009
	UCB-LM	0.027	0.028	0.028	0.028	0.028	0.027	0.027	0.027	0.026	0.026	0.026	0.026	0.025	0.025	0.024	0.019	0.017	0.014
	UCBV	−0.014	−0.005	−0.004	−0.003	−0.003	0.002	0.006	0.006	0.007	0.008	0.009	0.009	0.011	0.012	0.012	0.015	0.015	0.015
	UCBV-M	0.019	0.024	0.025	0.025	0.026	0.027	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.023	0.021	0.019
33	UCB1-M	0.016	0.021	0.022	0.022	0.023	0.023	0.024	0.024	0.024	0.024	0.025	0.025	0.025	0.025	0.025	0.021	0.020	0.017
	UCB-L	0.011	0.012	0.012	0.012	0.012	0.013	0.013	0.013	0.013	0.013	0.014	0.014	0.014	0.014	0.014	0.013	0.012	0.011
	UCB-LM	**0.032**	**0.035**	**0.035**	**0.036**	**0.036**	**0.036**	0.035	0.036	0.036	0.036	0.036	0.036	0.035	0.035	0.034	0.029	0.026	0.023
	UCBV	−0.028	−0.018	−0.016	−0.015	−0.015	−0.009	−0.005	−0.004	−0.002	−0.001	0.000	0.001	0.004	0.005	0.006	0.014	0.015	0.016
	UCBV-M	0.025	0.033	0.034	0.034	0.035	0.036	**0.037**	**0.037**	**0.037**	**0.037**	**0.037**	**0.037**	**0.037**	**0.037**	**0.037**	**0.032**	**0.030**	**0.026**

Table C.11: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_H$ and $\mu_{\max} = 10^{-2}$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds																		
		416	628	967	975	1,034	1,097	1,324	2,128	2,239	2,396	2,551	3,005	4,064	5,650	6,126	7,755	9,496	9,957	9,984
5	UCB1-M	−0.005	−0.003	−0.001	−0.001	−0.001	−0.000	0.000	0.002	0.002	0.002	0.002	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002
	UCB-L	0.030	0.030	0.030	0.030	0.029	0.029	0.029	0.027	0.027	0.027	0.026	0.026	0.025	0.023	0.022	0.021	0.020	0.020	0.020
	UCB-LM	0.032	0.031	0.031	0.031	0.030	0.030	0.029	0.028	0.027	0.027	0.027	0.026	0.025	0.023	0.022	0.021	0.020	0.020	0.020
	UCBV	0.026	0.029	0.030	0.030	0.030	0.030	0.030	0.029	0.029	0.028	0.028	0.028	0.027	0.025	0.025	0.024	0.023	0.022	0.022
	UCBV-M	0.028	0.030	0.030	0.030	0.030	0.030	0.030	0.029	0.029	0.028	0.028	0.028	0.026	0.025	0.024	0.023	0.022	0.022	0.022
9	UCB1-M	−0.009	−0.005	−0.001	−0.001	−0.000	−0.000	0.001	0.005	0.005	0.005	0.006	0.006	0.007	0.008	0.008	0.008	0.008	0.008	0.008
	UCB-L	0.037	0.038	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.038	0.038	0.038	0.037	0.036	0.035	0.034	0.033	0.033	0.033
	UCB-LM	0.043	0.044	0.045	0.045	0.045	0.045	0.045	0.043	0.043	0.043	0.042	0.042	0.040	0.038	0.037	0.036	0.034	0.034	0.034
	UCBV	0.027	0.033	0.038	0.039	0.039	0.040	0.041	0.043	0.043	0.043	0.043	0.043	0.043	0.042	0.041	0.041	0.039	0.039	0.039
	UCBV-M	0.041	0.043	0.045	0.045	0.045	0.045	0.045	0.045	0.046	0.045	0.045	0.045	0.044	0.042	0.042	0.041	0.040	0.039	0.039
17	UCB1M	−0.009	−0.004	0.002	0.002	0.003	0.003	0.005	0.009	0.010	0.010	0.011	0.012	0.014	0.015	0.015	0.016	0.016	0.016	0.016
	UCB-L	0.036	0.038	0.041	0.041	0.041	0.041	0.042	0.042	0.042	0.043	0.043	0.042	0.042	0.042	0.041	0.040	0.040	0.039	0.039
	UCB-LM	0.052	0.054	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.054	0.054	0.053	0.052	0.050	0.049	0.047	0.046	0.045	0.045
	UCBV	0.016	0.025	0.033	0.033	0.034	0.035	0.037	0.042	0.043	0.043	0.044	0.045	0.046	0.046	0.046	0.046	0.045	0.045	0.045
	UCBV-M	0.049	0.053	0.056	0.056	0.056	0.056	0.057	0.057	0.057	0.057	0.057	0.056	0.055	0.053	0.052	0.051	0.049	0.049	0.049
33	UCB1-M	−0.008	−0.002	0.005	0.005	0.006	0.007	0.009	0.015	0.016	0.016	0.017	0.019	0.021	0.023	0.024	0.024	0.025	0.025	0.025
	UCB-L	0.032	0.035	0.038	0.039	0.039	0.039	0.040	0.043	0.043	0.043	0.044	0.044	0.045	0.045	0.045	0.045	0.044	0.044	0.044
	UCB-LM	**0.059**	**0.062**	**0.063**	**0.063**	0.063	0.063	0.064	0.064	0.064	0.064	0.064	0.064	0.062	0.061	0.060	0.059	0.057	0.057	0.057
	UCBV	−0.002	0.009	0.020	0.020	0.022	0.023	0.027	0.036	0.037	0.038	0.039	0.041	0.044	0.047	0.047	0.048	0.048	0.048	0.048
	UCBV-M	0.053	0.059	0.063	0.063	**0.064**	**0.064**	**0.066**	**0.067**	**0.067**	**0.067**	**0.067**	**0.067**	**0.066**	**0.065**	**0.064**	**0.063**	**0.061**	**0.061**	**0.061**

Table C.12: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_H$ and $\mu_{\max} = 10^{-3}$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds						
		1	9,931	9,982	9,990	9,991	9,994	10,000
5	UCB1-M	0.000	−0.018	−0.018	−0.018	−0.018	−0.018	−0.018
	UCB-L	0.000	0.046	0.046	0.046	0.046	0.046	0.046
	UCB-LM	0.000	0.047	0.047	0.047	0.047	0.047	0.047
	UCBV	0.000	0.045	0.045	0.045	0.045	0.045	0.045
	UCBV-M	0.000	0.046	0.046	0.046	0.046	0.046	0.046
9	UCB1-M	0.000	−0.028	−0.028	−0.028	−0.028	−0.028	−0.028
	UCB-L	0.000	0.055	0.055	0.055	0.055	0.055	0.055
	UCB-LM	0.000	0.062	0.062	0.062	0.062	0.062	0.062
	UCBV	0.000	0.052	0.053	0.053	0.053	0.053	0.053
	UCBV-M	0.000	0.062	0.062	0.062	0.062	0.062	0.062
17	UCB1-M	0.000	−0.033	−0.033	−0.033	−0.033	−0.033	−0.033
	UCB-L	0.000	0.055	0.055	0.055	0.055	0.055	0.055
	UCB-LM	0.000	0.070	0.070	0.070	0.070	0.070	0.070
	UCBV	0.000	0.045	0.045	0.045	0.045	0.045	0.045
	UCBV-M	0.000	0.071	0.071	0.071	0.071	0.071	0.071
33	UCB1-M	0.000	−0.035	−0.035	−0.035	−0.035	−0.035	−0.035
	UCB-L	0.000	0.050	0.050	0.050	0.050	0.050	0.050
	UCB-LM	0.000	**0.076**	**0.076**	**0.076**	**0.076**	**0.076**	**0.076**
	UCBV	0.000	0.030	0.030	0.030	0.030	0.030	0.030
	UCBV-M	0.000	0.075	0.076	0.076	0.076	0.076	0.076

Table C.13: $\Delta P_{\%}(t)$ (with respect to UCB1 with 5 arms) obtained by different algorithms. The configuration is: $\mathcal{S}_H$ and $\mu_{\max} = 10^{-4}$. Rounds are in thousands and each number of rounds reported in the table corresponds to the maximum of $\Delta P_{\%}(t)$ for some algorithm (potentially more than one).

		Rounds		
		1	9,981	10,000
5	UCB1-M	0.000	−0.036	−0.036
	UCB-L	0.000	0.035	0.035
	UCB-LM	0.000	0.037	0.037
	UCBV	0.000	0.009	0.009
	UCBV-M	0.000	0.018	0.018
9	UCB1-M	0.000	−0.055	−0.055
	UCB-L	0.000	0.035	0.035
	UCB-LM	0.000	0.040	0.040
	UCBV	0.000	−0.004	−0.004
	UCBV-M	0.000	0.019	0.019
17	UCB1-M	0.000	−0.066	−0.066
	UCB-L	0.000	0.031	0.031
	UCB-LM	0.000	0.042	0.042
	UCBV	0.000	−0.021	−0.021
	UCBV-M	0.000	0.017	0.017
33	UCB1-M	0.000	−0.072	−0.072
	UCB-L	0.000	0.026	0.026
	UCB-LM	0.000	**0.041**	**0.041**
	UCBV	0.000	−0.041	−0.041
	UCBV-M	0.000	0.019	0.019