Stochastic Rising Bandits: A Best Arm Identification Approach

Alessandro Montenegro
Politecnico di Milano, Milan
alessandro.montenegro@polimi.it

Marco Mussi
Politecnico di Milano, Milan
marco.mussi@polimi.it

Francesco Trovó
Politecnico di Milano, Milan
francesco1.trovo@polimi.it

Marcello Restelli
Politecnico di Milano, Milan
marcello.restelli@polimi.it

Alberto Maria Metelli
Politecnico di Milano, Milan
albertomaria.metelli@polimi.it

Abstract

Stochastic Rising Bandits (SRBs) model sequential decision-making problems in which the expected reward of the available options increases every time they are selected. This setting captures a wide range of scenarios in which the available options are *learning entities* whose performance improves (in expectation) over time. While previous works addressed the regret minimization problem, this paper focuses on the *fixed-budget Best Arm Identification* (BAI) problem for SRBs. In this scenario, given a fixed budget of rounds, we are asked to provide a recommendation about the best option at the end of the identification process. We propose two algorithms to tackle the above-mentioned setting, namely R-UCBE, which resorts to a UCB-like approach, and R-SR, which employs a successive reject procedure. Then, we prove that, with a sufficiently large budget, they provide guarantees on the probability of properly identifying the optimal option at the end of the learning process. Furthermore, we derive a lower bound on the error probability, matched by our R-SR (up to logarithmic factors), and illustrate how the need for a sufficiently large budget is unavoidable in the SRB setting. Finally, we numerically validate the proposed algorithms in synthetic and real-world environments and compare them with the currently available BAI strategies.

1 Introduction

Multi-Armed Bandits (MAB, Lattimore and Szepesvári, 2020) are a well-known framework that effectively solves learning problems requiring sequential decisions. Given a time horizon, the learner chooses, at each round, a single option (a.k.a. arm) and observes the corresponding noisy reward, which is a realization of an unknown distribution. The MAB problem is commonly studied in two flavours: *regret minimization* (Auer et al., 2002) and *best arm identification* (Bubeck et al., 2009). In regret minimization, the goal is to control the cumulative loss w.r.t. the optimal arm over a time horizon. Conversely, in best arm identification, the goal is to provide a recommendation about the best arm at the end of the time horizon. Specifically, we are interested in the fixed-budget scenario, where we seek to minimize the error probability of recommending the wrong arm at the end of the time budget, no matter the loss incurred during learning.

16th European Workshop on Reinforcement Learning (EWRL 2023).

This work focuses on the *Stochastic Rising Bandits* (SRB), a specific instance of the *rested* bandit (Tekin and Liu, 2012) setting in which the expected reward of an arm increases according to the number of times it has been pulled. Online learning in such a scenario has been recently addressed from a regret minimization perspective by Metelli et al. (2022), in which the authors provide no-regret algorithms for the SRB setting in both the rested and restless cases. The SRB setting models several real-world scenarios where arms improve their performance over time. A classic example is the so-called *Combined Algorithm Selection and Hyperparameter optimization* (CASH, Thornton et al., 2013; Kotthoff et al., 2017; Erickson et al., 2020; Li et al., 2020; Zöller and Huber, 2021), a problem of paramount importance in *Automated Machine Learning* (AutoML, Feurer et al., 2015; Yao et al., 2018; Hutter et al., 2019; Mussi et al., 2023). In CASH, the goal is to identify the *best learning algorithm* together with the *best hyperparameter* configuration for a given ML task (e.g., classification or regression). In this problem, every arm represents a hyperparameter tuner acting on a specific learning algorithm. A pull corresponds to a unit of time/computation in which we improve (on average) the hyperparameter configuration (via the tuner) for the corresponding learning algorithm. CASH was handled in a bandit *Best Arm Identification* (BAI) fashion in Li et al. (2020) and Cella et al. (2021). The former handles the problem by considering rising rested bandits with *deterministic* rewards, failing to represent the intrinsic uncertain nature of such processes. Instead, the latter, while allowing stochastic rewards, assumes that the expected rewards evolve according to a *known* parametric functional class, whose parameters have to be learned.

Original Contributions In this paper, we address the design of algorithms to solve the BAI task in the rested SRB setting when a *fixed budget* is provided.[1] More specifically, we are interested in algorithms guaranteeing a sufficiently large probability of recommending the arm with the largest expected reward *at the end* of the time budget (as if only this arm were pulled from the beginning). The main contributions of the paper are summarized as follows:[2]

- We propose two *algorithms* to solve the BAI problem in the SRB setting: `R-UCBE` (an optimistic approach, Section 4) and `R-SR` (a phases-based rejection algorithm, Section 5). First, we introduce specifically designed estimators required by the algorithms (Section 3). Then, we provide guarantees on the error probability of the misidentification of the best arm.
- We derive the first error probability *lower bound* for the SRB setting, matched by our `R-SR` algorithm up to logarithmic factors, which highlights the complexity of the problem and the need for a sufficiently large time budget (Section 6).
- Finally, we conduct *numerical simulations* on synthetically generated data and a real-world online best model selection problem. We compare the proposed algorithms with the ones available in the bandit literature to tackle the SRB problem (Section 8).

2 Problem Formulation

In this section, we revise the Stochastic Rising Bandits (SRB) setting (Heidari et al., 2016; Metelli et al., 2022). Then, we formulate our best arm identification problem, introduce the definition of error probability, and provide a preliminary characterization of the problem.

Setting We consider a rested Multi-Armed Bandit problem $\mathfrak{p}\ _i\mathfrak{q}_{i\in\mathrm{J}K\mathrm{K}}$ with a finite number of arms K.[3] Let $T \in \mathbb{N}$ be the time budget of the learning process. At every round $t \in \mathrm{J}T\mathrm{K}$, the agent selects an arm $I_t \in \mathrm{J}K\mathrm{K}$, plays it, and observes a reward $x_t \sim {}_{I_t}\mathfrak{p}N_{I_t,t}\mathfrak{q}$, where ${}_{I_t}\mathfrak{p}N_{I_t,t}\mathfrak{q}$ is the reward distribution of the chosen arm I_t at round t and depends on the number of pulls performed so far $N_{i,t} := \sum_{s=1}^{t-1}\mathbb{1}\{I_s = i\}$ (i.e., rested). The rewards are stochastic, formally $x_t := {}_{I_t}\mathfrak{p}N_{I_t,t}\mathfrak{q} + \epsilon_t$, where ${}_{I_t}\mathfrak{p}\,\mathfrak{q}$ is the expected reward of arm I_t and ϵ_t is a zero-mean σ^2-subgaussian noise, conditioned to the past.[4] As customary in the bandit literature, we assume that the rewards are bounded in expectation, formally ${}_i\mathfrak{p}n\mathfrak{q} \in [0, 1], \forall i \in \mathrm{J}K\mathrm{K}, n \in \mathrm{J}T\mathrm{K}$. As in (Metelli et al., 2022), we focus on a particular family of rested bandits in which the expected rewards are monotonic *non-decreasing* and *concave* in expectation.

[1] We focus on the rested setting only and, thus, from now on, we will omit "rested" in the setting name.

[2] Additional motivating examples are discussed in Appendix A. The proofs of all the statements in this work are provided in Appendix C.

[3] Let $y, z \in \mathbb{N}$, we denote with $\mathrm{J}z\mathrm{K} := \{1, \dots, z\}$, and with $\mathrm{J}y, z\mathrm{K} := \{y, \dots, z\}$.

[4] A zero-mean random variable x is σ^2-subgaussian if it holds $\mathbb{E}_x[e^{\lambda x}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for every $\lambda \in \mathbb{R}$.

Assumption 2.1 (Non-decreasing and concave expected rewards)**.** *Let be a rested MAB, defining* $\Delta_i(n) := \mu_i(n) - \mu_i(n-1)$, *for every* $n \in \mathbb{N}$ *and every arm* $i \in [K]$ *the rewards are non-decreasing and concave, formally:*

$$\textit{Non-decreasing:} \quad \Delta_i(n) \geq 0; \qquad\qquad \textit{Concave:} \quad \Delta_i(n+1) \leq \Delta_i(n).$$

Intuitively, the $\Delta_i(n)$ represents the *increment* of the real process $\mu_i(\cdot)$ evaluated at the n^{th} pull. Notice that concavity emerges in several settings, such as the best model selection and economics, representing the decreasing marginal returns (Lehmann et al., 2001; Heidari et al., 2016).[5]

Learning Problem The goal of BAI in the SRB setting is to select the arm providing the largest expected reward with a large enough probability given a fixed budget $T \in \mathbb{N}$. Unlike the stationary BAI problem (Audibert et al., 2010), in which the optimal arm is not changing, in this setting, we need to decide *when* to evaluate the optimality of an arm. We define optimality by considering the largest expected reward at time T. Formally, given a time budget T, the optimal arm $i^*(T) \in [K]$, which we assume unique, satisfies:

$$i^*(T) := \arg\max_{i \in [K]} \mu_i(T),$$

where we highlighted the dependence on T as, with different values of the budget, $i^*(T)$ may change. Let $i \in [K] \setminus \{i^*(T)\}$ be a suboptimal arm, we define the suboptimality gap as $\Delta_i(T) := \mu_{i^*(T)}(T) - \mu_i(T)$. We employ the notation (i) to denote the i^{th} best arm at time T (arbitrarily breaking ties), i.e., we have $\mu_{(2)}(T) \geq \cdots \geq \mu_{(K)}(T)$. Given an algorithm $\mathcal{A}$ that recommends $\hat{I}(T) \in [K]$ at the end of the learning process, we measure its performance with the *error probability*, i.e., the probability of recommending a suboptimal arm at the end of the time budget T:

$$e_T(\mathcal{A}) := \mathbb{P}_{\mathcal{A}}(\hat{I}(T) \neq i^*(T)).$$

Problem Characterization We now provide a characterization of a specific class of polynomial functions to upper bound the increments $\Delta_i(n)$.

Assumption 2.2 (Bounded $\Delta_i(n)$)**.** *Let be a rested MAB, there exist* $c > 0$ *and* $\alpha > 1$ *such that for every arm* $i \in [K]$ *and number of pulls* $n \in [0; T]$ *it holds that* $\Delta_i(n) \leq cn^{-\alpha}$.

We anticipate that, even if our algorithms will not require such an assumption, it will be used for deriving the lower bound and for providing more human-readable error probability guarantees. Furthermore, we observe that our Assumption 2.2 is fulfilled by a strict superset of the functions employed in Cella et al. (2021).

3 Estimators

In this section, we introduce the estimators of the arm expected reward employed by the proposed algorithms.[6] A visual representation of such estimators is provided in Figure 1.

Let $\gamma \in (0; 1)$ be the fraction of samples collected up to the current time t we use to build estimators of the expected reward. We employ an *adaptive arm-dependent window* size $h(N_{i,t-1}) := \gamma N_{i,t-1}$ to include the most recent samples collected only, avoiding the use of samples that are no longer representative. We define the set of the last $h(N_{i,t-1})$ rounds in which the i^{th} arm was pulled as:

$$T_{i,t} := \{s \in [t] : I_s = i \wedge N_{i,s} > N_{i,t-1} - h(N_{i,t-1}) \wedge I_s \in [0]; h(N_{i,t-1}) - 1\}.$$

Furthermore, the set of the pairs of rounds s and s_1 belonging to the sets of the last and second-last $h(N_{i,t-1})$-wide windows of the i^{th} arm is defined as:

$$S_{i,t} := \{(s, s_1) \in [t] \times [t] : I_s = I_{s_1} = i \wedge N_{i,s} > N_{i,t-1} - h(N_{i,t-1}) \wedge N_{i,s_1} > N_{i,t-1} - 2h(N_{i,t-1}) \wedge s > s_1\}.$$

In the following, we design a *pessimistic* estimator and an *optimistic* estimator of the expected reward of each arm at the end of the budget time T, i.e., $\mu_i(T)$.[7]

[5] A discussion on the non-learnability of the problem when concavity assumption does not hold is provided in Appendix E.

[6] The estimators are adaptations of those presented by Metelli et al. (2022) to handle a fixed time budget T.

[7] Naïvely computing the estimators from their definition requires $O(h(N_{i,t-1}))$ number of operations. An efficient way to incrementally update them, using $O(1)$ operations, is provided in Appendix B.

Pessimistic Estimator The *pessimistic* estimator $\hat{\mu}_i(N_{i,t-1})$ is a negatively biased estimate of $\mu_i(T)$ obtained assuming that the function $\mu_i(\cdot)$ remains constant up to time T. This corresponds to the minimum admissible value under Assumption 2.1 (due to the *Non-decreasing* constraint). This estimator is an average of the last $h(N_{i,t-1})$ observed rewards collected from the i^{th} arm, formally:

$$\hat{\mu}_i(N_{i,t-1}) := \frac{1}{h(N_{i,t-1})} \sum_{s \in T_{i,t}} X_{i,s} \tag{1}$$

The estimator enjoys the following concentration property.

Lemma 3.1 (Concentration of $\hat{\mu}_i$)**.** *Under Assumption 2.1, for every $a > 0$, simultaneously for every arm $i \in [K]$ and number of pulls $n \in [0, T]$, with probability at least $1 - 2TKe^{-a^2}$ it holds that:*

$$\hat{\mu}_i(n) - \mu_i(n) \le \hat{\mu}_i(n) - \mu_i(n) \le \hat{\mu}_i(n) - \mu_i(n),$$

where $\hat{\mu}_i(n) := \sqrt{\frac{a}{h(n)}}$ *and* $\hat{\mu}_i(n) := \frac{1}{2}(2T - n - h(n) - 1)\,\mu_i(n) - \mu_i(h(n) - 1)$.

As supported by intuition, we observe that the estimator is affected by a negative bias that is represented by $\hat{\mu}_i(n)$ that vanishes as $n \to \infty$ under Assumption 2.1 with a rate that depends on the increment functions $\mu_i(\cdot)$. Considering also the term $\hat{\mu}_i(n)$ and recalling that $h(n) = O(n)$, under Assumption 2.2, the overall concentration rate is $O(n^{-1/2} + cT^{-n})$.

Optimistic Estimator The *optimistic* estimator $B_i^T(N_{i,t-1})$ is a positively biased estimation of $\mu_i(T)$ obtained assuming that function $\mu_i(\cdot)$ linearly increases up to time T. This corresponds to the maximum value admissible under Assumption 2.1 (due to the *Concavity* constraint). The estimator is constructed by adding to the pessimistic estimator $\hat{\mu}_i(N_{i,t-1})$ an estimate of the increment occurring in the next step up to T. The latter uses the last $2h(N_{i,t-1})$ samples to obtain an upper bound of such growth thanks to the concavity assumption, formally:

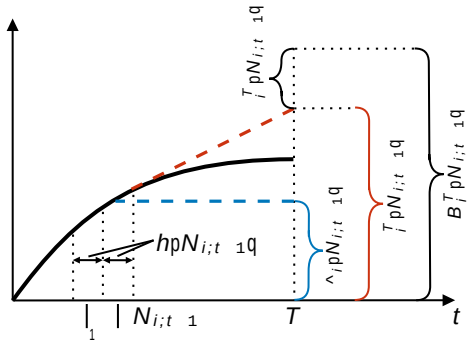


Figure 1: Graphical representation of the pessimistic $\hat{\mu}_i(N_{i,t-1})$ and the optimistic $B_i^T(N_{i,t-1})$ estimators.

$$B_i^T(N_{i,t-1}) := \hat{\mu}_i(N_{i,t-1}) + \sum_{(j,k) \in S_{i,t}} (T - j)\frac{x_j - x_k}{h(N_{i,t-1})^2}. \tag{2}$$

The estimator displays the following concentration guarantee.

Lemma 3.2 (Concentration of B_i^T)**.** *Under Assumption 2.1, for every $a > 0$, simultaneously for every arm $i \in [K]$ and number of pulls $n \in [0, T]$, with probability at least $1 - 2TKe^{-a^{10}}$ it holds that:*

$$B_i^T(n) - \mu_i(n) \le B_i^T(n) - \mu_i(n) \le B_i^T(n) - \mu_i(n),$$

where $B_i^T(n) := (T - n - h(n) - 1)\sqrt{\frac{a}{h(n)^3}}$ *and* $B_i^T(n) := \frac{1}{2}(2T - n - h(n) - 1)\,\mu_i(n) - \mu_i(2h(n) - 1)$.

Differently from the pessimistic estimation, the optimistic one displays a positive vanishing bias $B_i^T(n)$. Under Assumption 2.2, we observe that the overall concentration rate is $O(T^{-n^{3/2}} + cT^{-n})$.

4 Optimistic Algorithm: `Rising Upper Confidence Bound Exploration`

In this section, we introduce and analyze `Rising Upper Confidence Bound Exploration` (`R-UCBE`) an *optimistic* error probability minimization algorithm for the SRB setting with a fixed budget. The algorithm explores by means of a UCB-like approach and, for this reason, makes use of the optimistic estimator B_i^T plus a bound to account for the uncertainty of the estimation. In `R-UCBE`, the choice of considering the optimistic estimator is natural and obliged since the pessimistic estimator is affected by negative bias and cannot be used to deliver optimistic estimates.

Algorithm The algorithm, whose pseudo-code is reported in Algorithm 1, requires as input an exploration parameter $a \ge 0$, the window size $\rho \in (0, 1/2]$, the time budget T, and the number of

arms K. At first, it initializes to zero the counters $N_{i,0}$, and sets to $+\infty$ the upper bounds $B_i^T(N_{i,0})$ of all the arms (Line 2). Subsequently, at each time $t \in \{1,\dots,T\}$, the algorithm selects the arm I_t with the largest upper confidence bound (Line 4):

$$I_t \in \arg\max_{i \in \{1,\dots,K\}} B_i^T(N_{i,t-1}) : \quad B_i^T(N_{i,t-1}) \neq B_i^T(N_{i,t-1}), \tag{3}$$

$$\text{with:} \quad B_i^T(N_{i,t-1}) : \quad \hat{\mu}(N_{i,t-1}) + \beta(T - N_{i,t-1}, N_{i,t-1}) + \sqrt{\frac{a}{N_{i,t-1}}}, \tag{4}$$

where $\beta_i^T(N_{i,t-1})$ represents the exploration bonus (a graphical representation is reported in Figure 1). Once the arm is chosen, the algorithm plays it and observes the feedback x_t (Line 5). Then, the optimistic estimate $\hat{\mu}_{I_t}^T(N_{I_t,t})$ and the exploration bonus $\beta_{I_t}^T(N_{I_t,t})$ of the selected arm I_t are updated (Lines 8-9). The procedure is repeated until the algorithm reaches the time budget T. The final recommendation of the best arm is performed using the last computed values of the bounds $B_i^T(N_{i,T})$, returning the arm $\hat{I}(T)$ corresponding to the largest upper confidence bound (Line 12).

Bound on the Error Probability of R-UCBE We now provide bounds on the error probability for R-UCBE. We start with a general analysis that makes no assumption on the increments $\beta(\cdot,\cdot)$ and, then, we provide a more explicit result under Assumption 2.2. The general result is formalized as follows.

Theorem 4.1. *Under Assumption 2.1, let a^* be the largest positive value of a satisfying:*

$$T \geq \sum_{i \neq i^*} \gamma_i(a), \tag{5}$$

where for every $i \in \{1,\dots,K\}$, $\gamma_i(a)$ is the largest integer for which it holds:

$$\underbrace{T - \gamma_i(a)\Delta_i + 2\gamma_i(a)\beta(T - \gamma_i(a), \gamma_i(a))}_{(A)} \geq \underbrace{T - \sqrt{\frac{a}{\gamma_i(a)}}}_{(B)}. \tag{6}$$

If a^* exists, then for every $a \in (0, a^*]$ the error probability of R-UCBE is bounded by:

$$e_T(\text{R-UCBE}) \leq 2TK \exp\left(-\frac{a}{10}\right). \tag{7}$$

Some comments are in order. First, a^* is defined implicitly, depending on the constants α, T, the increments $\beta(\cdot,\cdot)$, and the suboptimality gaps $\Delta_i(T)$. In principle, there might exist no $a^* > 0$ fulfilling condition in Equation (5) (this can happen, for instance, when the budget T is not large enough), and, in such a case, we are unable to provide theoretical guarantees on the error probability of R-UCBE. Second, the result presented in Theorem 4.1 holds for generic increasing and concave expected reward functions. This result shows that, as expected, the error probability decreases when the exploration parameter a increases. However, this behavior stops when we reach the threshold a^*. Intuitively, the value of a^* sets the maximum amount of exploration we should use for learning.

Under Assumption 2.2, i.e., using the knowledge on the increment $\beta(\cdot,\cdot)$ upper bound, we derive a result providing conditions on the time budget T under which a^* exists and an explicit value for a^*.

Corollary 4.2. *Under Assumptions 2.1 and 2.2, if the time budget T satisfies:*

$$T \geq \begin{cases} c^{\frac{1}{1-\alpha}}(1-\alpha)^{\frac{1}{1-\alpha}} H_{1,1-\alpha}(T)(\log K)^{\frac{1}{1-\alpha}} & \text{if } \alpha \in (1/2, 3/4) \\ c^{\frac{2}{3}}(1-\alpha)^{\frac{2}{3}} H_{1,2/3}(T)(\log K)^{3} & \text{if } \alpha \in (0, 1/2] \end{cases}; \tag{8}$$

there exists $a^* > 0$ defined as:

$$a^* = \begin{cases} \frac{3}{4}\left(\frac{T^{1-(1-\alpha)}}{H_{1,1-\alpha}(T)}\right)^{2} (c(1-\alpha))^{2} & \text{if } \alpha \in (1/2, 3/4) \\ \frac{3}{4}\left(\frac{T^{1-2/3}}{H_{1,2/3}(T)}\right)^{3/2} (c(1-\alpha))^{2} & \text{if } \alpha \in (0, 1/2] \end{cases};$$

where $H_{1,\alpha}(T) := \sum_{i \neq i^*} \frac{1}{\Delta_i(T)^{1/\alpha}}$ for $\alpha > 0$. Then, for every $a \in (0, a^*]$, the error probability of R-UCBE is bounded by:

$$e_T(\text{R-UCBE}) \leq 2TK \exp\left(-\frac{a}{10}\right).$$

Algorithm 1: R-UCBE.

Input : Time budget T, Number of arms K, Window size ε, Exploration parameter a

Initialize $N_{i,0} \leftarrow 0$,
  $B_i^T(0) \leftarrow \infty \; ; \forall i \in \{1,\dots,K\}$
for $t \in \{1,\dots,T\}$ **do**
  Compute $I_t \in \arg\max_{i \in \{1,\dots,K\}} B_i^T(N_{i,t-1})$
  Pull arm I_t and observe x_t
  $N_{I_t,t} \leftarrow N_{I_t,t-1} + 1$
  $N_{i,t} \leftarrow N_{i,t-1}; \; \forall i \neq I_t$
  Update $\hat{\mu}_{I_t}^T(N_{I_t,t})$
  Update $\hat{\mu}_{I_t}^T(N_{I_t,t})$
  Compute $B_{I_t}^T(N_{I_t,t}) = \hat{\mu}_{I_t}^T(N_{I_t,t}) + \hat{\mu}_{I_t}^T(N_{I_t,t})$
end
Recommend $J(T) \in \arg\max_{i \in \{1,\dots,K\}} B_i^T(N_{i,T})$

Algorithm 2: R-SR.

Input : Time budget T, Number of arms K, Window size ε

Initialize $t \leftarrow 1$; $N_0 \leftarrow 0$; $X_0 \leftarrow \{1,\dots,K\}$
for $j \in \{1,\dots,K-1\}$ **do**
  for $i \in X_{j-1}$ **do**
    for $l \in \{N_{j-1}+1,\dots,N_j\}$ **do**
      Pull arm i and observe x_t
      $t \leftarrow t+1$
    end
    Update $\hat{\mu}_i(N_j)$
  end
  Define $\bar{I}_j \in \arg\min_{i \in X_{j-1}} \hat{\mu}_i(N_j)$
  Update $X_j \leftarrow X_{j-1} \setminus \{\bar{I}_j\}$
end
Recommend $J(T) \in X_{K-1}$ (unique)

First of all, we notice that the error probability $e_T(\text{R-UCBE})$ presented in Theorem 4.2 holds under the condition that the time budget T fulfills Equation (8). We defer a more detailed discussion on this condition to Remark 5.1, where we show that the existence of a finite value of T fulfilling Equation (8) is ensured under mild conditions.

Let us remark that term $H_{1/2}(T)$ characterizes the complexity of the SRB setting, corresponding to term H_1 of Audibert et al. (2010) for the classical BAI problem when $\rho = 2$. As expected, in the small-ρ regime (i.e., $\rho \in (1, 3/2]$), looking at the dependence of $H_{1/\rho}(T)$ on ρ, we realize that the complexity of a problem decreases as the parameter ρ increases. Indeed, the larger ρ, the faster the expected reward reaches a stationary behavior. Nevertheless, even in the large-ρ regime (i.e., $\rho > 3/2$), the complexity of the problem is governed by $H_{1/2}(T)$, leading to an error probability larger than the corresponding one for BAI in standard bandits (Audibert et al., 2010). This can be explained by the fact that R-UCBE uses the optimistic estimator that, as shown in Section 3, enjoys a slower concentration rate compared to the standard sample mean, even for stationary bandits.

This two-regime behavior has an interesting interpretation when comparing Corollary 4.2 with Theorem 4.1. Indeed, $\rho = 3/2$ is the break-even threshold in which the two terms of the l.h.s. of Equation (6) have the same convergence rate. Specifically, the term $A(\rho)$ takes into account the expected rewards growth (i.e., the bias in the estimators), while $B(\rho)$ considers the uncertainty in the estimations of the R-UCBE algorithm (i.e., the variance). Intuitively, when the expected reward function displays a slow growth (i.e., $\mu(n) \propto n^{\rho}$ with $\rho < 3/2$), the bias term $A(\rho)$ dominates the variance term $B(\rho)$ and the value of a changes accordingly. Conversely, when the variance term $B(\rho)$ is the dominant one (i.e., $\mu(n) \propto n^{\rho}$ with $\rho > 3/2$), the threshold a is governed by the estimation uncertainty, being the bias negligible.

As common in optimistic algorithms for BAI (Audibert et al., 2010), setting a theoretically sound value of exploration parameter a (i.e., computing a), requires additional knowledge of the setting, namely the complexity index $H_{1/2}(T)$.[8] In the next section, we propose an algorithm that relaxes this requirement.

5 Phase-Based Algorithm: Rising Successive Rejects

In this section, we introduce the Rising Successive Rejects (R-SR), a phase-based solution inspired by the one proposed by Audibert et al. (2010), which overcomes the drawback of R-UCBE of requiring knowledge of $H_{1/2}(T)$.

Algorithm R-SR, whose pseudo-code is reported in Algorithm 2, takes as input the time budget T and the number of arms K. At first, it initializes the set of the active arms X_0 with all the available arms (Line 1). This set will contain the arms that are still eligible candidates to be recommended. The entire process proceeds through $K-1$ phases. More specifically, during the j^{th} phase, the arms

[8] We defer the empirical study of the sensitivity of a to Section 8.

still remaining in the active arms set X_{j-1} are played (Line 5) for $N_j - N_{j-1}$ times each, where:

$$N_j := \left\lceil \frac{1}{\overline{\log}(K)} \frac{T-K}{K+1-j} \right\rceil, \tag{9}$$

and $\overline{\log}(K) := \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$. At the end of each phase, the arm with the smallest value of the pessimistic estimator $\hat{\nu}_i(N_j)$ is discarded from the set of active arms (Line 11). At the end of the $(K-1)^{\text{th}}$ phase, the algorithm recommends the (unique) arm left in X_{K-1} (Line 13).

It is worth noting that R-SR makes use of the pessimistic estimator $\hat{\nu}_i(n)$. Even if both estimators defined in Section 3 are viable for R-SR, the choice of using the pessimistic estimator is justified by its better concentration rate $O(n^{-1/2})$ compared to that of the optimistic estimator $O(Tn^{-3/2})$, being $n \leq T$ (see Section 3).

Note that the phase lengths are the ones adopted by Audibert et al. (2010). This choice allows us to provide theoretical results without requiring domain knowledge (still under a large enough budget). An optimized version of N_j may be derived assuming full knowledge of the gaps $\Delta_i(T)$, but, unfortunately, such a hypothetical approach would have similar drawbacks as R-UCBE.

Bound on the Error Probability of R-SR The following theorem provides the guarantee on the error probability for the R-SR algorithm.

Theorem 5.1. *Under Assumptions 2.1 and 2.2, if the time budget T satisfies:*

$$T \geq 2^{\frac{1}{1}} c^{-1} \overline{\log}(K) \max_{i \in \{2,\dots,K\}} \left(i^{-1} \Delta_{\sigma(i)}(T)^{-1} \right), \tag{10}$$

then, the error probability of R-SR is bounded by:

$$e_T(\text{R-SR}) \leq \frac{K(K-1)}{2} \exp\left(-\frac{1}{8} \frac{T-K}{\overline{\log}(K) H_2(T)} \right),$$

where $H_2(T) := \max_{i \in \{1,\dots,K\}} i \Delta_{\sigma(i)}(T)^{-2}$ and $\overline{\log}(K) := \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$.

Similar to the R-UCBE, the complexity of the problem is characterized by term $H_2(T)$ that, for the standard MAB setting, reduces to the H_2 term of Audibert et al. (2010). Furthermore, when the condition of Equation (10) on the time budget T is satisfied, the error probability coincides with that of the SR algorithm for standard MABs (apart for constant terms). The following remark elaborates on the conditions of Equations (8) and (10) about the minimum requested time budget.

Remark 5.1 (About the minimum time budget T)**.** *To satisfy the e_T bounds presented in Corollary 4.2 and Theorem 5.1, R-UCBE and R-SR require the conditions provided by Equations* (8) *and* (10) *about the time budget T, respectively. First, let us notice that if the suboptimal arms converge to an expected reward different from that of the optimal arm as $T \to \infty$, it is always possible to find a finite value of T such that these conditions are fulfilled. Formally, assume that there exists $T_0 < \infty$ and that for every $T \geq T_0$ we have that for all suboptimal arms $i \neq i^*(T)$ it holds that $\Delta_i(T) \geq \Delta_{\min} > 0$. In such a case, the l.h.s. of Equations* (8) *and* (10) *are upper bounded by a function of $\Delta_{\min}$ and are independent on T. Instead, if a suboptimal arm converges to the same expected reward as the optimal arm when $T \to \infty$, the identification problem is more challenging and, depending on the speed at which the two arms converge as a function of T, might slow down the learning process arbitrarily. This should not surprise as the BAI problem becomes non-learnable even in standard (stationary) MABs when multiple optimal arms are present (Heide et al., 2021).*

6 Lower Bound

In this section, we investigate the complexity of the BAI problem for SRBs with a fixed budget.

Minimum time budget T We show that, under Assumptions 2.1 and 2.2, any algorithm requires a minimum time budget T to be guaranteed to identify the optimal arm, even in a deterministic setting.

Theorem 6.1. *For every algorithm $\mathcal{A}$, there exists a deterministic SRB satisfying Assumptions 2.1 and 2.2 with $K \geq 8^{1/(p-1)}$ such that the optimal arm $i^*(T)$ cannot be identified for some time budgets T unless:*

$$T \geq H_{1;1/(p-1)}(T) \left(\sum_{i \neq i^*(T)} \frac{1}{\Delta_i(T)^{\frac{1}{1}}} \right). \tag{11}$$

	Error Probability $e_T(\pi)$	Time Budget T
SRB	$\frac{1}{4}\exp\left(-\frac{8T}{2\sum_{i\neq i^*(T)}\frac{1}{\Delta_i^2}}\right)$	$\sum_{i\neq i^*(T)}\frac{1}{\Delta_i^2\log(K)}$
R-UCBE	$2TK\exp\left(-\frac{a}{10}\right)$	$\begin{cases} c^{-1}(1-2\varepsilon)^{-1}\sum_{i\neq i^*(T)}\frac{1}{(1-\Delta_i)^2}\log(K)^{-1} & \text{if } \varepsilon\in[1,3-\{2] \\ c^{-\frac{2}{3}}(1-2\varepsilon)^{-\frac{2}{3}}\sum_{i\neq i^*(T)}\frac{1}{\Delta_i^{2\{3}}}\log(K)^{3} & \text{if } \varepsilon\in(3\{2,8] \end{cases}$
R-SR	$\frac{K(K-1)}{2}\exp\left(-\frac{T-K}{\log(K)\max_{i\in[K]}i\Delta_{(i)}^{-2}}\right)$	$2^{-1}c^{-1}\log(K)\max_{i\in[2,K]}i\Delta_{(i)}^{-2})$

Table 1: Bounds on the time budget and error probability: lower for the setting and upper for the algorithms.

Theorem 6.1 formalizes the intuition that any of the suboptimal arms must be pulled a sufficient number of times to ensure that, if pulled further, it cannot become the optimal arm. Indeed, in the proof of Theorem 6.1, we show that each suboptimal arm $i \neq i^*(T)$ has to be pulled at least an expected number of times of order $\mathbb{E}[N_i(T)] \geq (1-\varepsilon)\Delta_i^{-2}\log(T)$ (see Equation 36). It is worth comparing this bound on the time budget with the corresponding conditions on the minimum time budget requested by Equations (8) and (10) for R-UCBE and R-SR, respectively. Regarding R-UCBE, we notice that the minimum admissible time budget in the small-ε regime is of order $H_{1;1\{}\log(T)^{\{}$ which is larger than term $H_{1;1-\varepsilon}\log(T)$ of Equation (11).[9] Similarly, in the large-ε regime (i.e., $\varepsilon > 3\{2$), the R-UCBE requirement is of order $H_{1;2\{3}\log(T)^3 \geq H_{1;2}\log(T)$ which is larger than the term of Theorem 6.1 since $1\{(1-\varepsilon)^2$. Concerning R-SR, it is easy to show that $H_{1;1-\varepsilon}\log(T) = \max_{i\in[2;K]}i\Delta_{(i)}^{-2}\log(T)$, apart from logarithmic terms, by means of the argument provided by (Audibert et al., 2010, Section 6.1). Thus, up to logarithmic terms, Equation (10) provides a tight condition on the minimum budget.

Error Probability Lower Bound We now present a lower bound on the error probability that every algorithm performing BAI in the SRB setting suffers.

Theorem 6.2. *For every algorithm* $\mathcal{A}$ *run with a time budget* T *fulfilling Equation* (11)*, there exists a SRB satisfying Assumptions 2.1 and 2.2 with* $K \geq 8^{1\{(1-\varepsilon)}$ *such that the error probability is lower bounded by:*

$$e_T(\mathcal{A}) \geq \frac{1}{4}\exp\left(-\frac{8T}{\log(K)H_{1;2}\log(T)}\right); \quad \textit{where} \quad H_{1;2}\log(T) := \sum_{i\neq i^*(T)}\frac{1}{\Delta_i^2\log(T)}.$$

Some comments are in order. First, we stated the lower bound for the case in which the minimum time budget satisfies the inequality of Theorem 6.1, which is a necessary condition for identifying the optimal arm. Second, the lower bound on the error probability matches, up to logarithmic factors, that of our R-SR, suggesting the superiority of this algorithm compared to R-UCBE. Finally, provided that the identificability condition of Equation (11), such a result corresponds to that of the standard (stationary) MABs (Audibert et al., 2010; Kaufmann et al., 2016). A summary of all the bounds provided in the paper is presented in Table 1.

7 Related Works

In this section, we summarize the relevant literature related both to the works focusing on the best arm identification problem and rested bandits. The SRB setting was proposed by Heidari et al. (2016) and analyzed by Metelli et al. (2022) from the regret minimization perspective.

Best Arm Identification in Stochastic Rising Bandits As highlighted in Section 1, the works mostly related to ours are the ones by Li et al. (2020) and Cella et al. (2021). They both focus on the BAI problem in the rested setting, given a fixed-budget. More specifically, Li et al. (2020) consider rising rested bandits in which the reward function of each arm increases as it is pulled. However, they limit to deterministic arms and, thus, fail to deal with the intrinsic stochasticity of the real-world

[9] See Lemma C.11.

processes they want to model. Instead, Cella et al. (2021) deal with the problem of identifying the arm with the smallest loss in a setting where the losses incurred by selecting an arm decrease over time. It is easy to show that such a setting can be transformed straightforwardly in the SRB one. However, the authors develop two algorithms whose theoretical guarantees hold under the assumption that the expected loss follows a specific known parametric functional form, whose parameters are to be estimated. This constitutes a major limitation to the presented work since checking such an assumption is not feasible in real-world settings.

Best Arm Identification The pure exploration and BAI problems have been first introduced by Bubeck et al. (2009), while algorithms able to learn in such a setting have been provided by Audibert et al. (2010). The work by Gabillon et al. (2012) proposes a unified approach to deal with stochastic best arm identification problems by having either a fixed budget or fixed confidence. However, the stochastic algorithms developed in this line of research only provide theoretical guarantees in settings where the expected reward is stationary over the pulls. Abbasi-Yadkori et al. (2018) propose a method able to handle both the stochastic and adversarial cases, but they do not make explicit use of the properties (e.g., increasing nature) of the expected reward. Finally, (Garivier and Kaufmann, 2016; Kaufmann et al., 2016; Carpentier and Locatelli, 2016) analyze the problem of BAI from the lower bound perspective.

Rested Bandits Bandit settings in which the evolution of an arm reward depends on the number of times the arm has been pulled, such as the one analyzed in our paper, are generally referred to as *rested*. A first general formulation of the rested bandit setting appeared in the work by Tekin and Liu (2012) and was further discussed by Mintz et al. (2020) and Seznec et al. (2020). In these works, the evolution of the expected reward of each arm is regulated by a Markovian process that is assumed to visit the same state multiple times. This is not the case for the rising bandits, where the arm expected rewards continuously increase over the time budget. Finally, a specific instance of the rested bandits is constituted by the *rotting* bandits (Levine et al., 2017; Seznec et al., 2019, 2020), in which the expected payoff for a given arm decreases with the number of pulls. However, as pointed out by Metelli et al. (2022), techniques developed for this setting cannot be directly translated into ours, due to the inherently different nature of the problem.

8 Numerical Validation

In this section, we provide a numerical validation of `R-UCBE` and `R-SR`. We compare them with state-of-the-art bandit baselines designed for stationary and non-stationary BAI in a synthetic setting, and we evaluate the sensitivity of `R-UCBE` to its exploration parameter a. Additional details about the experiments presented in this section are available in Appendix F. Additional experimental results on both synthetic settings and in a real-world experiment are available in Appendix G. The code to reproduce the experiments can be found at https://github.com/MontenegroAlessandro/BestArmIdSRB.

Baselines We compare our algorithms against a wide range of solutions for BAI:

- `RR`: uniformly pulls all the arms until the budget ends in a *round-robin* fashion and, in the end, makes a recommendation based on the empirical mean of their reward over the collected samples;
- `RR-SW`: makes use of the same exploration strategy as `RR` to pull arms but makes a recommendation based on the empirical mean over the last $\frac{T}{K}$ collected samples from an arm.[10]
- `UCB-E` and `SR` (Audibert et al., 2010): algorithms for the stationary BAI problem;
- `Prob-1` (Abbasi-Yadkori et al., 2018): an algorithm dealing with the adversarial BAI setting;
- `ETC` and `Rest-Sure` (Cella et al., 2021): algorithms developed for the decreasing loss BAI setting, that we converted through a linear transformation of the reward.

The hyperparameters required by the above methods have been set as prescribed in the original papers. For both our algorithms and `RR-SW`, we set $\varepsilon = 0.25$.

Setting To assess the quality of the recommendation $\hat{\imath}(T)$ provided by our algorithms, we consider a synthetic SRB setting with $K = 5$ and $\sigma = 0.01$. Figure 2 shows the evolution of the expected values of the arms w.r.t. the number of pulls. In this setting, the optimal arm changes depending on whether $T \in [1, 185]$ or $T \in [185, 8]$. Thus, when the time budget is close to that value, the problem is more challenging since the optimal and second-best arms expected rewards are close to each other. For this reason, the BAI algorithms are less likely to provide a correct recommendation

[10] The formal description of this baseline, as well as its theoretical analysis, is provided in Appendix D.

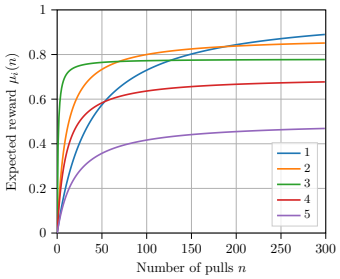


Figure 2: Expected values $\mu_i(n)$ for the arms of the synthetic setting.

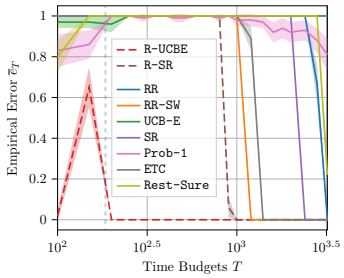


Figure 3: Empirical error rate for the synthetically generated setting (100 runs, mean 95% c.i.).

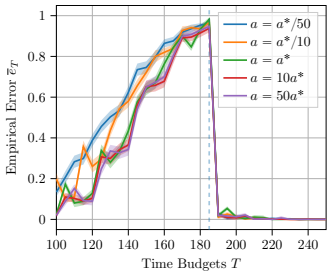


Figure 4: Empirical error rate for the R-UCBE at different a (1000 runs, mean 95% c.i.).

than for time budgets for which the two expected rewards are well separated. We compare the analyzed algorithms A in terms of empirical error $\bar{e}_T(A)$ (the smaller, the better), i.e., the empirical counterpart of $e_T(A)$ averaged over 100 runs, considering time budgets $T \in [100, 3200]$.

Results The empirical error probability provided by the analyzed algorithms in the synthetically generated setting is presented in Figure 3. We report with a dashed vertical blue line at $T = 185$, i.e., the budgets after which the optimal arm no longer changes. Before such a budget, all the algorithms provide large errors (i.e., $e_T(A) > 0.2$). However, R-UCBE outperforms the others by a large margin, suggesting that an optimistic estimator might be advantageous when the time budget is small. Shortly after $T = 185$, R-UCBE starts providing the correct suggestion consistently. R-SR begins to identify the optimal arm (i.e., with $e_T(\text{R-SR}) < 0.05$) for time budgets $T > 1000$. Nonetheless, both algorithms perform significantly better than the baseline algorithms used for comparison.

Sensitivity Analysis for the Exploration Parameter of R-UCBE We perform a sensitivity analysis on the exploration parameter a of R-UCBE. Such a parameter should be set to a value less or equal to a^*, and the computation of the latter is challenging. We tested the sensitivity of R-UCBE to this hyperparameter by looking at the error probability for $a \in \{50, a^*, 10, a^*/10, 50, a^*/50\}$. Figure 4 shows the empirical errors of R-UCBE with different parameters a, where the blue dashed vertical line denotes the last time the optimal arm changes over the time budget. It is worth noting how, even in this case, we have two significantly different behaviors before and after such a time. Indeed, if $T \geq 185$, we have that misspecification with larger values than a^* does not significantly impact the performance of R-UCBE, while smaller values slightly decrease the performance. Conversely, for $T < 185$ learning with different values of a seems not to impact the algorithm performance significantly. This corroborates the previous results about the competitive performance of R-UCBE.

9 Discussion and Conclusions

This paper introduces the BAI problem with a fixed budget for the Stochastic Rising Bandits setting. Notably, such setting models many real-world scenarios in which the reward of the available options increases over time, and the interest is on the recommendation of the one having the largest expected rewards after the time budget has elapsed. In this setting, we presented two algorithms, namely R-UCBE and R-SR providing theoretical guarantees on the error probability. R-UCBE is an optimistic algorithm requiring an exploration parameter whose optimal value requires prior information on the setting. Conversely, R-SR is a phase-based solution that only requires the time budget to run. We established lower bounds for the error probability an algorithm suffers in such a setting, which is matched by our R-SR, up to logarithmic factors. Furthermore, we showed how a requirement on the minimum time budget is unavoidable to ensure the identifiability of the optimal arm. Finally, we validate the performance of the two algorithms in both synthetically generated and real-world settings. A possible future line of research is to derive an algorithm balancing the tradeoff between theoretical guarantees on the e_T and the chance of providing such guarantees with lower time budgets.

Acknowledgements

This paper is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

References

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Proceedings of the Algorithmic Learning Theory Conference (ALT)*, volume 5809, pages 23–37, 2009.

Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transaction on Information Theory*, 58(8):5588–5611, 2012.

Alberto Maria Metelli, Francesco Trovò, Matteo Pirola, and Marcello Restelli. Stochastic rising bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 15421–15457, 2022.

Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the ACM Conference of Knowledge Discovery and Data Mining (SIGKDD)*, pages 847–855, 2013.

Lars Kotthoff, Chris Thornton, Holger H Hoos, Frank Hutter, and Kevin Leyton-Brown. Auto-weka 2.0: Automatic model selection and hyperparameter optimization in weka. *Journal of Machine Learning Research*, 18:1–5, 2017.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Yang Li, Jiawei Jiang, Jinyang Gao, Yingxia Shao, Ce Zhang, and Bin Cui. Efficient automatic CASH via rising bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4763–4771, 2020.

Marc-André Zöller and Marco F Huber. Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research*, 70:409–472, 2021.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2962–2970, 2015.

Quanming Yao, Mengshuo Wang, Yuqiang Chen, Wenyuan Dai, Yu-Feng Li, Wei-Wei Tu, Qiang Yang, and Yang Yu. Taking human out of learning applications: A survey on automated machine learning. *arXiv preprint arXiv:1810.13306*, 2018.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

Marco Mussi, Davide Lombarda, Alberto Maria Metelli, Francesco Trovó, and Marcello Restelli. Arlo: A framework for automated reinforcement learning. *Expert Systems with Applications*, 224: 119883, 2023.

Leonardo Cella, Massimiliano Pontil, and Claudio Gentile. Best model identification: A rested bandit formulation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 1362–1372, 2021.

Hoda Heidari, Michael J. Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *Proceeding of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1562–1570, 2016.

Benny Lehmann, Daniel Lehmann, and Noam Nisan. Combinatorial auctions with decreasing marginal utilities. In *ACM Proceedings of the Conference on Electronic Commerce (EC)*, pages 18–28, 2001.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 41–53, 2010.

Rianne De Heide, James Cheshire, Pierre Ménard, and Alexandra Carpentier. Bandits with many optimal arms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 22457–22469, 2021.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1:1–1:42, 2016.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3221–3229, 2012.

Yasin Abbasi-Yadkori, Peter L. Bartlett, Victor Gabillon, Alan Malek, and Michal Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 75, pages 918–949, 2018.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 49, pages 998–1027, 2016.

Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 49, pages 590–604, 2016.

Yonatan Mintz, Anil Aswani, Philip Kaminsky, Elena Flowers, and Yoshimi Fukuoka. Nonstationary bandits with habituation and recovery dynamics. *Operations Research*, 68(5):1493–1516, 2020.

Julien Seznec, Pierre Ménard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 3784–3794, 2020.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3074–3083, 2017.

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 2564–2572, 2019.

A Additional Motivating Examples

In this appendix, we provide two additional motivating examples to better understand and appreciate the SRB setting.

Selection of Athletes for Competitions Consider the role of a professional trainer for a team, having several athletes (i.e., our arms) to train in order to increase their performances. The final goal is to select a single athlete to represent the team in a competition. The performances of athletes increase when the trainer properly follows them. However, a trainer can follow just one athlete at a time. The trainer can be modeled as an agent performing best-arm (athlete) identification, and the athletes represent the arms that increase their payoffs (i.e., performance) when pulled (i.e., when the trainer follows them).

Online Best Model Selection Suppose we have to choose among a set of algorithms to maximize a given index (e.g., accuracy) over a training set. In this setting, we expect that all the algorithms progressively increase (on average) the index value and converge to their optimum value with different convergence rates. Therefore, we want to identify which candidate algorithm is the most likely to reach optimal performances, given the budget, and assign the available resources (e.g., computational power or samples). In summary, this problem reduces to the identification, with the largest probability, of the algorithm that converges faster to the optimum. A real-world example of such a scenario is provided in Figure 8.

B Estimators Efficient Update

In this appendix, we describe how to implement an efficient version (i.e., fully online) of the estimators we presented in the main paper. We resort to the update developed by Metelli et al. (2022). This update provides a way to achieve an $O(1)$ computational complexity at each step for the update of the estimates for the pessimistic estimator $\wedge_i(t)$ and optimistic estimator $\overline{\wedge}_i(t)$.

With a slight abuse of notation, only in this appendix, for the sake of simplicity, we denote with $x_{i;n}$ the reward collected at the n^{th} pull from the arm i and with $h_{i;t} = h(N_{i;t-1})$ the window size. Differently, from what we use in the paper, here the reward subscript indicates the arm i and the number of pulls of that arm n instead of the total number of pulls t we used in the definition of x_t.

More specifically, the *pessimistic* estimator $\wedge_i$ can be written as:

$$\wedge_i(t) = \frac{\bar{a}_i}{h_{i;t}};$$

where the quantity $\bar{a}_i$ is updated as follows:

$$\bar{a}_i \leftarrow \begin{cases} \bar{a}_i - x_{i;N_{i;t}} + x_{i;N_{i;t}-h_{i;t}} & \text{if } h_{i;t} = h_{i;t-1}; \\ \bar{a}_i - x_{i;N_{i;t}} & \text{otherwise} \end{cases}$$

and $\bar{a}_i = 0$ as the algorithm starts.

Instead, the *optimistic* estimator $\overline{\wedge}_i(t)$, is updated using:

$$\overline{\wedge}_i(t) = \frac{1}{h_{i;t}} \bar{a}_i + \frac{T(\bar{a}_i - \bar{b}_i)}{h_{i;t}} - \frac{\bar{c}_i - \bar{d}_i}{h_{i;t}}.$$

Where the quantity $\bar{a}_i$ is defined and updated above and $\bar{b}_i$, $\bar{c}_i$, and $\bar{d}_i$ are updated as follows:

$$\bar{b}_i \leftarrow \begin{cases} \bar{b}_i - x_{i;N_{i;t}-h_{i;t}} + x_{i;N_{i;t}-2h_{i;t}} & \text{if } h_{i;t} = h_{i;t-1}; \\ \bar{b}_i - x_{i;N_{i;t}-2h_{i;t}-1} & \text{otherwise} \end{cases}$$

$$\bar{c}_i \leftarrow \begin{cases} \bar{c}_i - N_{i;t} x_{i;N_{i;t}} + (N_{i;t} - h_{i;t}) x_{i;N_{i;t}-h_{i;t}} & \text{if } h_{i;t} = h_{i;t-1}; \\ \bar{c}_i - N_{i;t} x_{i;N_{i;t}} & \text{otherwise} \end{cases}$$

$$\bar{d}_i \leftarrow \begin{cases} \bar{d}_i - N_{i;t} x_{i;N_{i;t}-h_{i;t}} + (N_{i;t} - h_{i;t}) x_{i;N_{i;t}-2h_{i;t}} & \text{if } h_{i;t} = h_{i;t-1}; \\ \bar{d}_i - (N_{i;t} - h_{i;t}) x_{i;N_{i;t}-2h_{i;t}-1} - \bar{b}_i & \text{otherwise} \end{cases}.$$

Similarly to what is presented above, the quantities are initialized to 0 as the algorithms start.

C Proofs and Derivations

In this appendix, we provide all the proofs omitted in the main paper. For the sake of generality, we will provide the derivations for a generic choice of the window size of the estimators $h_{i,t}$ which depends on the arm $i \in [K]$ and on the round $t \in [T]$. When needed, we will particularize the choice for the case in which the window size depends on the number of pulls only $h_{i,t} = h(N_{i,t-1})$.

C.1 Proofs of Section 3

Lemma C.1. *Under Assumption 2.1, for every $i \in [K]$, $j, k \in \mathbb{N}$ with $k > j$, it holds that:*

$$\mu_i(j) - \mu_i(k) \leq \frac{\mu_i(j) - \mu_i(k)}{j - k}.$$

Proof. Using Assumption 2.1, we get:

$$\mu_i(j) - \frac{1}{j-k}\sum_{l=k}^{j-1}\mu_i(l) \geq \mu_i(j) - \frac{1}{j-k}\sum_{l=k}^{j-1}\mu_i(l) - \frac{1}{j-k}\sum_{l=k}^{j-1}(\mu_i(l) - \mu_i(l+1)) = \frac{\mu_i(j) - \mu_i(k)}{j-k},$$

where the first inequality comes from the concavity of the expected reward function (Assumption 2.1), and the second equality comes from the definition of increment. □

Lemma C.2. *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h_{i,t} \leq t - N_{i,t-1}$, let us define:*

$$r_i^T(N_{i,t}) := \frac{1}{h_{i,t}}\sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \mu_i(l) + (T-t)\frac{\mu_i(l) - \mu_i(l-h_{i,t})}{h_{i,t}};$$

otherwise if $h_{i,t} = 0$, we set $r_i^T(N_{i,t}) := +\infty$. Then, $r_i^T(N_{i,t}) \geq \mu_i(T)$ and, if $N_{i,t-1} \geq 2$, it holds that:

$$r_i^T(N_{i,t}) - \mu_i(T) \leq \frac{1}{2}(2T - 2N_{i,t-1} + h_{i,t} - 1)(\mu_i(N_{i,t-1} - 2h_{i,t} + 1).$$

Proof. Following the derivation provided above, we have for every $l \in \{2, \dots, N_{i,T-1}\}$:

$$\begin{aligned}
\mu_i(T) - \mu_i(l) &\leq \sum_{j=l}^{T-1}(\mu_i(j) - \mu_i(j+1)) \\
&\leq (\mu_i(l) - \mu_i(l-1))(T-l) \qquad (12) \\
&\leq (\mu_i(l) - \mu_i(l-h_{i,t}))\frac{T-l}{h_{i,t}}; \qquad (13)
\end{aligned}$$

where Equation (12) follows from Assumption 2.1, and Equation (13) is obtained from Lemma C.1. By averaging over the most recent $1 \leq h_{i,t} \leq t - N_{i,t-1}$ pulls, we get:

$$\mu_i(T) - \frac{1}{h_{i,t}}\sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \mu_i(l) \leq (T-t)\frac{\mu_i(l) - \mu_i(l-h_{i,t})}{h_{i,t}} \leq r_i^T(N_{i,t}) - \mu_i(T).$$

For the bias bound, when $N_{i,t-1} \geq 2$, we retrieve:

$$r_i^T(N_{i,t}) - \mu_i(T) \leq r_i^T(N_{i,t}) - \frac{1}{h_{i,t}}\sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \mu_i(l) + (T-t)\frac{\mu_i(l) - \mu_i(l-h_{i,t})}{h_{i,t}} \leq r_i^T(N_{i,t}) - \mu_i(T) \qquad (14)$$

$$\begin{aligned}
&\leq \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \left(T-l\right)\frac{\mu_i(l)-\mu_i(l:h_{i,t})}{h_{i,t}} \\
&\quad \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \left(T-l\right)\frac{1}{h_{i,t}} \sum_{j=l-h_{i,t}}^{l-1} \mu_i(j) \\
&\leq \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \left(T-l\right)\left(\mu_i(l)-\mu_i(l:h_{i,t})\right) \qquad (15)\\
&\leq \frac{1}{2}(2T-2N_{i,t-1}+h_{i,t}-1)\left(\mu_i(N_{i,t-1})-\mu_i(h_{i,t}-1)\right); \qquad (16)
\end{aligned}$$

where Equation (14) follows from Assumption 2.1 applied as $\mu_i(l) \leq \mu_i(N_{i,t})$, Equation (15) follows from Assumption 2.1 and bounding $\frac{1}{h_{i,t}} \sum_{j=l-h_{i,t}}^{l-1} \mu_i(j) \leq \mu_i(l-h_{i,t})$, and Equation (16) is derived still from Assumption 2.1, $\mu_i(l-h_{i,t}) \leq \mu_i(N_{i,t-1}-h_{i,t})$ and computing the summation. □

Lemma C.3. *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq t-N_{i,t-1}+2$, let us define:*

$$\begin{aligned}
&\Gamma_i^T(N_{i,t}) := \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} X_i(l) \left(T-l\right)\frac{X_i(l)-X_i(l:h_{i,t})}{h_{i,t}} \; ; \\
&\Gamma_i^T(N_{i,t}) \left(T-N_{i,t-1}+h_{i,t}-1\right)\sqrt{\frac{a}{h_{i,t}^3}};
\end{aligned}$$

where $X_i(l)$ denotes the reward collected from arm i when pulled for the l-th time. Otherwise, if $h_{i,t}=0$, we set $\Gamma_i^T(t) := \infty$ and $\Gamma_i^T(t) := -\infty$. Then, if the window size depends on the number of pulls only $h_{i,t}=h(N_{i,t-1})$, it holds that:

$$\mathbb{P}\left(\exists t \in [T] : \Gamma_i^T(N_{i,t}) - \tilde{\Gamma}_i^T(N_{i,t}) > \Gamma_i^T(N_{i,t})\right) \leq 2T\exp\left(-\frac{a}{10}\right).$$

Proof. Before starting the proof, it is worth noting that under the event $\{h_{i,t}=0\}$, it holds that $\Gamma_i^T(t) - \tilde{\Gamma}_i^T(t) \leq \Gamma_i^T(t) = \infty$. Thus, under the convention that $\infty - \infty = 0$, then $0 > \Gamma_i^T(t)$ is not satisfied. For this reason, we need to perform our analysis under the event $\{h_{i,t} \geq 1\}$.

The first thing to do is to remove the dependence on the number of pulls that, in a generic time instant, represents a random variable. So, we can write:

$$\begin{aligned}
&\mathbb{P}\left(\exists t \in [T] : \Gamma_i^T(N_{i,t}) - \tilde{\Gamma}_i^T(N_{i,t}) > \Gamma_i^T(N_{i,t})\right) \\
&\quad \leq \mathbb{P}\left(\exists n \in [0,T] \text{ s.t. } h_{i,n} \geq 1 : \left|\Gamma_i^T(n) - \tilde{\Gamma}_i^T(n)\right| > \Gamma_i^T(n)\right) \\
&\quad \leq \sum_{n \in [0,T]: h_{i,n} \geq 1} \mathbb{P}\left(\left|\Gamma_i^T(n) - \tilde{\Gamma}_i^T(n)\right| > \Gamma_i^T(n)\right); \qquad (17)
\end{aligned}$$

where Equation (17) follows form a union bound over the possible values of $N_{i,t}$.

Now that we have a fixed value of n, consider a generic time t in which arm i has been pulled. We will observe a reward x_t composed by the mean of the process $\mu_i(N_{i,t})$ plus some noise. The noise will be equal to $\mu_i(N_{i,t}) - x_t = \mu_i(N_{i,t}) - \mu_i(N_{i,t})$, i.e., as the difference (not known) between the observed value for the arm i at time t and its real value at the same time. Let us rewrite the quantity to be bounded as follows for every n:

$$\begin{aligned}
&h_{i,n} \; \Gamma_i^T(n) - \tilde{\Gamma}_i^T(n) \\
&\quad \sum_{l=n-h_{i,n}+1}^{n} \mu_i(l) \left(T-l\right)\frac{\mu_i(l)-\mu_i(l:h_{i,n})}{h_{i,n}}
\end{aligned}$$

$$\sum_{l=n-h_{i,n}+1}^{n}\left(1-\frac{T-l}{h_{i,n}}\right)\mu_i(l)-\sum_{l=n-h_{i,n}+1}^{n}\frac{T-l}{h_{i,n}}\mu_i(l-h_{i,n}).$$

Here, notice that all the quantities $\mu_i(l)$ and $\mu_i(l-h_{i,n})$ are independent since the number of pulls l is fully determined by n and $h_{i,n}$, that now are non-random quantities.

Now, we apply the Azuma-Hoëffding's inequality of Lemma C.5 from Metelli et al. (2022) for weighted sums of subgaussian martingale difference sequences. To this purpose, we compute the sum of the square weights:

$$\begin{aligned}
&\sum_{l=n-h_{i,n}+1}^{n}\left(1-\frac{T-l}{h_{i,n}}\right)^2+\sum_{l=n-h_{i,n}+1}^{n}\left(\frac{T-l}{h_{i,n}}\right)^2\\
&\leq h_{i,n}\left(1-\frac{T-n+h_{i,n}-1}{h_{i,n}}\right)^2+h_{i,n}\left(\frac{T-n+h_{i,n}-1}{h_{i,n}}\right)^2\\
&\leq \frac{5(T-n+h_{i,n}-1)}{h_{i,n}}.
\end{aligned}$$

Given the previous argument, we have, for a fixed n:

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{\mu}_i^T(n)-\tilde{\mu}_i^T(n)\right|\geq\varepsilon_i^T(n)\right)\\
&\leq \mathbb{P}\left(\sum_{l=n-h_{i,n}+1}^{n}\left(1-\frac{T-l}{h_{i,n}}\right)\mu_i(T)-\sum_{l=n-h_{i,n}+1}^{n}\frac{T-l}{h_{i,n}}\mu_i(T-h_{i,n})\geq h_{i,n}\,\varepsilon_i^T(t)\right)\\
&\leq 2\exp\left(-\frac{h_{i,n}^2\,\varepsilon_i^T(n)^2}{2\cdot 2\cdot\frac{5(T-n+h_{i,n}-1)}{h_{i,n}}}\right)\\
&\leq 2\exp\left(-\frac{a}{10}\right).
\end{aligned}$$

By replacing the obtained result into Equation (17) we get:

$$\sum_{n\in[0,T]:h_{i,n}\geq 1}2\exp\left(-\frac{a}{10}\right)\leq\sum_{n=1}^{t}2\exp\left(-\frac{a}{10}\right)\leq 2T\exp\left(-\frac{a}{10}\right).$$

□

Lemma C.4. *For every arm $i\in[K]$, every round $t\in[T]$, and window width $1\leq h_{i,t}\leq N_{i,t-1}$, let us define:*

$$\mu_i(N_{i,t}):=\frac{1}{h_{i,t}}\sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}}\mu_i(l);$$

otherwise, if $h_{i,t}=0$, we set $\mu_i(N_{i,t}):=0$. Then, $\hat{\mu}_i^T(N_{i,t})=\mu_i(T)$ and, if $N_{i,t-1}\geq 2$, it holds that:

$$\mu_i(T)-\mu_i(N_{i,t})=\frac{1}{2}\left(2T-2N_{i,t-1}-h_{i,t}+1\right)\left(\mu_i(N_{i,t-1})-\mu_i(N_{i,t-1}-h_{i,t}+1)\right).$$

Proof. Following the derivation provided above, we have for every $l\in\{2,\dots,N_{i,T-1}\}$:

$$\mu_i(T)-\mu_i(l)=\sum_{j=l}^{T-1}\left(\mu_i(j)-\mu_i(l)\right).\tag{18}$$

Thus, by averaging over the most recent $1 \leq h_{i,t} \leq t N_{i,t-1}\{2u$ pulls, we get:

$$
\begin{aligned}
{}_i\mathrm{p}T\mathrm{q} &\quad \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} {}_i\mathrm{p}l\mathrm{q} \sum_{j=l}^{T-1} {}_i\mathrm{p}j\mathrm{q} \\
&\quad {}_i\mathrm{p}N_{i,t}\mathrm{q} \quad \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \sum_{j=l}^{T-1} {}_i\mathrm{p}j\mathrm{q} \\
&\leq {}_i\mathrm{p}N_{i,t}\mathrm{q} \quad \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \sum_{j=l}^{T-1} {}_i\mathrm{p}j\mathrm{q} \\
&\leq {}_i\mathrm{p}N_{i,t}\mathrm{q} \quad \frac{1}{2}\mathrm{p}2T \quad 2N_{i,t-1} \quad h_{i,t} \quad 1\mathrm{q}\ {}_i\mathrm{p}N_{i,t-1} \quad h_{i,t} \quad 1\mathrm{q};
\end{aligned}
$$

where we used Assumption 2.1. □

Lemma C.5. *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq t N_{i,t-1}\{2u$, let us define:*

$$
\begin{aligned}
\hat{\mu}_i^T\mathrm{p}N_{i,t}\mathrm{q} &: \quad \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} X_i\mathrm{p}l\mathrm{q}; \\
\hat{\mu}_i^T\mathrm{p}N_{i,t}\mathrm{q} &\quad \mathrm{c}\sqrt{\frac{a}{h_{i,t}}};
\end{aligned}
$$

where $X_i\mathrm{p}l\mathrm{q}$ denotes the reward collected from arm i when pulled for the l-th time. Otherwise, if $h_{i,t} = 0$, we set $\hat{\mu}_i^T\mathrm{p}t\mathrm{q} := \infty$ and $\hat{\mu}_i^T\mathrm{p}t\mathrm{q} := \infty$. Then, if the window size depends on the number of pulls only $h_{i,t} = h\mathrm{p}N_{i,t-1}\mathrm{q}$, it holds that:

$$
\mathbb{P}\left[\exists t \in [T] : |\hat{\mu}_i\mathrm{p}N_{i,t}\mathrm{q} \quad {}_i\mathrm{p}N_{i,t}\mathrm{q}| \geq \hat{\mu}_i\mathrm{p}N_{i,t}\mathrm{q}\right] \leq 2T \exp\left(-\frac{a}{2}\right).
$$

Proof. Before starting the proof, it is worth noting that under the event $\{h_{i,t} = 0u$, it holds that $\hat{\mu}_i^T\mathrm{p}t\mathrm{q} = {}_i^T\mathrm{p}t\mathrm{q} + \hat{\mu}_i^T\mathrm{p}t\mathrm{q} = \infty$. Thus, under the convention that $\infty \geq \infty = 0$, then $0 \geq \hat{\mu}_i^T\mathrm{p}t\mathrm{q}$ is not satisfied. For this reason, we need to perform our analysis under the event $\{h_{i,t} \geq 1u$.

The first thing to do is to remove the dependence on the number of pulls that, in a generic time instant, represents a random variable. So, we can write:

$$
\begin{aligned}
&\mathbb{P}\left[\exists t \in [T] : |\hat{\mu}_i\mathrm{p}N_{i,t}\mathrm{q} \quad {}_i\mathrm{p}N_{i,t}\mathrm{q}| \geq \hat{\mu}_i\mathrm{p}N_{i,t}\mathrm{q}\right] \\
&\quad \leq \mathbb{P}\left[\exists n \in [0; T] \text{ s.t. } h_{i,n} \geq 1 : |\hat{\mu}_i\mathrm{p}n\mathrm{q} \quad {}_i\mathrm{p}n\mathrm{q}| \geq \hat{\mu}_i\mathrm{p}n\mathrm{q}\right] \\
&\quad \leq \sum_{n \in [0;T] : h_{i,n} \geq 1} \mathbb{P}\left[|\hat{\mu}_i\mathrm{p}n\mathrm{q} \quad {}_i\mathrm{p}n\mathrm{q}| \geq \hat{\mu}_i\mathrm{p}n\mathrm{q}\right];
\end{aligned} \tag{19}
$$

where Equation (19) follows form a union bound over the possible values of $N_{i,t}$.

Now that we have a fixed value of n, consider a generic time t in which arm i has been pulled. We will observe a reward x_t composed by the mean of the process ${}_i\mathrm{p}N_{i,t}\mathrm{q}$ plus some noise. The noise will be equal to ${}_i\mathrm{p}N_{i,t}\mathrm{q} \quad x_t = {}_i\mathrm{p}N_{i,t}\mathrm{q}$, i.e., as the difference (not known) between the observed value for the arm i at time t and its real value at the same time. Let us rewrite the quantity to be bounded as follows, for every n:

$$
h_{i,n}\,\mathrm{p}\hat{\mu}_i\mathrm{p}n\mathrm{q} \quad {}_i\mathrm{p}n\mathrm{q}\mathrm{q} \sum_{l=n-h_{i,n}+1}^{n} {}_i\mathrm{p}l\mathrm{q}.
$$

Here we can note that all the quantities ${}_i\mathrm{p}l\mathrm{q}$ and ${}_i\mathrm{p}l = h_{i,n}\mathrm{q}$ are independent since the number of pulls l is fully determined by n and $h_{i,n}$, that now are non-random quantities.

Now, we apply the Azuma-Hoëffding's inequality of Lemma C.5 from Metelli et al. (2022) for sums of subgaussian martingale difference sequences. For a fixed n, we have:

$$\begin{aligned}
\mathbb{P}\left[\left|\hat{\mu}_i(n) - \mu_i(n)\right| \geq \hat{c}_i(n)\right] &\leq \mathbb{P}\left[\bigcup_{l=n}^{t} \mu_i(T_l) \geq h_{i,n}\, \hat{\mu}_i^T(t)\right] \\
&\leq 2\exp\left(-\frac{h_{i,n}\, \hat{\mu}_i^T(n)^2}{2\,\sigma^2}\right) \\
&\leq 2\exp\left(-\frac{a}{2}\right).
\end{aligned}$$

By replacing the obtained result into Equation (19) we get:

$$\sum_{n\in[0,T]:\, h_{i,n}\geq 1} 2\exp\left(-\frac{a}{2}\right) \leq \sum_{n=1}^{t} 2\exp\left(-\frac{a}{2}\right) \leq 2T\exp\left(-\frac{a}{2}\right).$$

□

Lemma 3.1 (Concentration of $\hat{\mu}_i$)**.** *Under Assumption 2.1, for every $a > 0$, simultaneously for every arm $i \in [K]$ and number of pulls $n \in [0, T]$, with probability at least $1 - 2TKe^{-a/2}$ it holds that:*

$$\hat{\mu}_i(n) - \hat{c}_i(n) \leq \mu_i(n) \leq \hat{\mu}_i(n) + \hat{c}_i(n),$$

where $\hat{c}_i(n) := \sqrt{\frac{a}{h_i(n)}}$ *and* $\hat{c}_i(n) := \frac{1}{2}(2T - n - h_i(n) + 1)\,\mu_i(n) - h_i(n) + 1)$.

Proof. The proof simply combines Lemmas C.4 and C.5 and a union bound over the arms. □

Lemma 3.2 (Concentration of $\hat{\mu}_i^T$)**.** *Under Assumption 2.1, for every $a > 0$, simultaneously for every arm $i \in [K]$ and number of pulls $n \in [0, T]$, with probability at least $1 - 2TKe^{-a/10}$ it holds that:*

$$\hat{\mu}_i^T(n) - \hat{c}_i^T(n) \leq \mu_i^T(n) \leq \hat{\mu}_i^T(n) + \hat{c}_i^T(n),$$

where $\hat{c}_i^T(n) := (T - n - h_i(n) + 1)\sqrt{\frac{a}{h_i(n)^3}}$ *and* $\hat{c}_i^T(n) := \frac{1}{2}(2T - n - h_i(n) + 1)\,\mu_i(n) - 2h_i(n) + 1)$.

Proof. The proof simply combines Lemmas C.2 and C.3 and a union bound over the arms. □

C.2 Proofs of Section 4

In this appendix, we provide the proofs we have omitted in the main paper for what concerns the theoretical results about R-UCBE. All the lemma below are assuming that the strategy we use for selecting the arm is R-UCBE.

Let us define the *good event* $\mathcal{E}$ corresponding to the scenario in which all (over the rounds and over the arms) the bounds $B_i^T(n)$ hold for the projection up to time T of the real reward expected value $\mu_i(n)$, formally:

$$\mathcal{E} := \left\{\forall i \in [K], \forall t \in [T] : \left|\hat{\mu}_i^T(t) - \tilde{\mu}_i^T(t)\right| \leq \hat{c}_i^T(t)\right\},$$

where $\tilde{\mu}_i^T(t)$ is the deterministic counterpart of $\hat{\mu}_i^T(t)$ considering the expected payoffs $\mu_i(\cdot)$ instead of the realizations, formally:

$$\tilde{\mu}_i^T(N_{i,t}) := \frac{1}{h_{i,t}} \sum_{l=N_{i,t}-h_{i,t}+1}^{N_{i,t}} \mu_i(l) \leq \frac{T - l}{h_{i,t}}\left(\mu_i(l) - \mu_i(l - h_{i,t})\right).$$

Lemma C.6. *Under Assumption 2.1 and assuming that the good event $\mathcal{E}$ holds, the maximum number of pulls $N_{i,T}$ of a sub-optimal arm ($i \neq i^*(T)$) performed by the R-UCBE is upper bounded by the*

maximum integer $y_i(a)$ *which satisfies the following condition:*

$$T\ _i(t)\left(1 \quad 2\varepsilon(y_i(a)u) \quad 2T\right)\sqrt{\frac{a}{t\varepsilon(y_i(a)u)^3}} \geq \quad _i(T).$$

Proof. In the following, we will use $\sim_i^T(N_{i,t\ 1})$ to bound the bias introduced by $\ _i^T(N_{i,t\ 1})$ and, subsequently, to find a number of pulls such that the algorithm cannot suggest pulling a suboptimal arm. Using Lemma C.4, we have that $\forall i \in [K], \forall t \in [T]$ and when $1 \leq h_{i,t} \leq t1\{2 \quad N_{i,t\ 1}\}$ with $N_{i,t\ 1} \geq 2$, it holds that:

$$\sim_i^T(N_{i,t\ 1}) \quad _i(T) \leq \frac{1}{2}\left(2T \quad 2N_{i,t\ 1} \quad h_{i,t} \quad 1\right) \quad _i(N_{i,t\ 1} \quad 2h_{i,t} \quad 1). \tag{20}$$

Let us assume that, at round t, the R-UCBE algorithm pulls the arm $i \in [K]$ such that $i \quad i\ (T)$. From now on, to avoid weighing down the notation, we will omit the dependence of the optimal arm $i\ (T)$ on the budget T, simply denoting it as i , and the window size will be denoted ad $h_{i,t} \quad h(N_{i,t\ 1})$. By construction, the algorithm chooses the arm with the largest upper confidence bound $B_i^T(N_{i,t\ 1})$. Thus, we have that $B_i^T(N_{i,t\ 1}) \geq B_i^T(N_{i\ ,t\ 1})$. Now, we want to identify the minimum number of pulls such that this event no longer occurs, assuming that the good event holds. We have that, if we pull such an arm $i \quad i$, it holds:

$$\begin{aligned} B_i^T(N_{i,t\ 1}) &\geq B_i^T(N_{i\ ,t\ 1}) \\ B_i^T(N_{i,t\ 1}) \quad B_i^T(N_{i\ ,t\ 1}) &\geq 0 \\ _i(T) \quad B_i^T(N_{i,t\ 1}) \quad B_i^T(N_{i\ ,t\ 1}) &\geq \quad _i(T). \end{aligned}$$

Using the definition of $\ _i(T)$ and the definition of the upper confidence bound $B_i^T(N_{i,t\ 1})$ in Equation (3) for i and i , we have:

$$_i\ (T) \quad _i(T) \quad _i^T(N_{i,t\ 1}) \quad _i^T(N_{i,t\ 1}) \quad _i^T(N_{i\ ,t\ 1}) \quad _i^T(N_{i\ ,t\ 1}) \geq \quad _i(T).$$

Given Assumption 2.1 we have that $_i\ (T) \leq \sim_i^T(N_{i\ ,t\ 1})$, and, therefore, we have:

$$\sim_i^T(N_{i\ ,t\ 1}) \quad _i(T) \quad _i^T(N_{i,t\ 1}) \quad _i^T(N_{i,t\ 1}) \quad _i^T(N_{i\ ,t\ 1}) \quad _i^T(N_{i\ ,t\ 1}) \geq \quad _i(T);$$

and, since under the good event , it holds that $\sim_i^T(N_{i\ ,t\ 1}) \quad _i^T(N_{i\ ,t\ 1}) \quad _i^T(N_{i\ ,t\ 1}) \quad 0$, we have:

$$\begin{aligned} _i(T) \quad _i^T(N_{i,t\ 1}) \quad _i^T(N_{i,t\ 1}) &\geq \quad _i(T) \\ _i(T) \quad _i^T(N_{i,t\ 1}) \quad \sim_i^T(N_{i,t\ 1}) \quad \underbrace{_i^T(N_{i,t\ 1}) \quad \sim_i^T(N_{i,t\ 1})}_{|(D)|} &\geq \quad _i(T); \end{aligned}$$

where we added and subtracted $\sim_i^T(N_{i,t\ 1})$ in the last equation. Under the good event , we can upper bound $|(D)| \leq |\sim_i^T(N_{i,t\ 1}) \quad _i^T(N_{i,t\ 1})| \leq \ _i^T(N_{i,t\ 1})$:

$$\sim_i^T(N_{i,t\ 1}) \quad _i(T) \quad 2\ _i^T(N_{i,t\ 1}) \geq \quad _i(T).$$

Using Equation (20), and substituting the definition of $\ _i^T(N_{i,t\ 1})$ provided in Equation (4), we have:

$$\underbrace{\frac{1}{2}\left(2T \quad 2N_{i,t\ 1} \quad h_{i,t} \quad 1\right)}_{\leq 2T} \quad _i(N_{i,t\ 1} \quad 2h_{i,t} \quad 1)$$

$$2 \quad \underbrace{(T \quad N_{i,t\ 1} \quad h_{i,t} \quad 1)}_{\leq T}\sqrt{\frac{a}{h_{i,t}^3}} \geq \quad _i(T)$$

$$\underbrace{T\,\mu_i(1-2\varepsilon)N_{i,u}}_{(A)} \;-\; 2\,\underbrace{T\sqrt{\frac{a}{t-N_{i,u}}}}_{(B)} \;\geq\; \mu_i(T); \tag{21}$$

where we used the definition of $h_{i,t} := t-N_{i,t}u$ and the fact that $N_{i,t+1} \geq N_{i,t} + 1$ since at time t the algorithm pulls the i-th arm.

This concludes the proof. □

Theorem 4.1. *Under Assumption 2.1, let a^+ be the largest positive value of a satisfying:*

$$T \sum_{i \neq i^*(T)} y_i(a) \leq 1, \tag{5}$$

where for every $i \in [K]$, $y_i(a)$ is the largest integer for which it holds:

$$\underbrace{T\,\mu_i(1-2\varepsilon)N_{i,u}}_{(A)} \;-\; 2T\,\underbrace{\sqrt{\frac{a}{t-N_{i,u}}}}_{(B)} \;\geq\; \mu_i(T). \tag{6}$$

If a^+ exists, then for every $a \in (0, a^+]$ the error probability of R-UCBE is bounded by:

$$e_T(\textit{R-UCBE}) \leq 2TK\exp\left(-\frac{a}{10}\right). \tag{7}$$

Proof. From the definition of the error probability, we have:

$$e_T(\text{R-UCBE}) = \mathbb{P}\left(\hat{I}(T) \neq i^*(T)\right) = \mathbb{P}(I_{T+1} \neq i^*(T)).$$

Therefore, we need to evaluate the probability that the R-UCBE algorithm would pull a suboptimal arm in the $T+1$ round. Given that Assumption 2.1 and that each suboptimal arms have been pulled a number of times $N_{i,T}$ (and cannot be pulled more) at the end of the time budget T, under the good event $\mathcal{E}$, we are guaranteed to recommend the optimal arm if:

$$T \sum_{i \neq i^*(T)} N_{i,T} \geq 1. \tag{22}$$

If Equation (22) holds, a suboptimal arm can be selected by R-UCBE for the next round $T+1$ only if the good event $\mathcal{E}$ does not hold $e_T(\text{R-UCBE}) = \mathbb{P}(\mathcal{E}^c)$, where we denote with ^{c} the complementary of event $\mathcal{E}$. This probability is upper bounded by Lemma C.5 as:

$$e_T(\text{R-UCBE}) = \mathbb{P}(\mathcal{E}^c) \leq 2TK\exp\left(-\frac{a}{10}\right).$$

We now derive a condition for a in order to make Equation (22) hold. Thanks to Lemma C.6 we know that $N_{i,T} \geq y_i(a)$ where $y_i(a)$ is the maximum integer such that:

$$T\,\mu_i(1-2\varepsilon)y_i(a) - 2T\sqrt{\frac{a}{t-y_i(a)}} \geq \mu_i(T).$$

From this condition, we observe that $y_i(a)$ is an increasing function of a. Therefore, we can select a in the interval $(0, a^+]$, where a^+ is the maximum value of a such that:

$$T \sum_{i \neq i^*(T)} y_i(a) \geq 1. \tag{23}$$

Note that we are not guaranteed that such a value of $a^+ > 0$ exists. In such a case, we cannot provide meaningful guarantees on the error probability of R-UCBE. □

Corollary 4.2. *Under Assumptions 2.1 and 2.2, if the time budget T satisfies:*

$$T \geq \begin{cases} c^{-1}(1) \left(2\ln(q) - 1\right) \left(H_{1,1}(\ln(Tq)) - \ln(K) - 1\right)^{-1} & \text{if } P(1,3/2) \\ c^{\frac{2}{3}}(1) \left(2\ln(q) - \frac{2}{3}\right) \left(H_{1,2/3}(\ln(Tq)) - \ln(K) - 1\right)^{\frac{3}{2}} & \text{if } P(3/2,\infty) \end{cases}; \tag{8}$$

there exists $a > 0$ defined as:

$$a \in \begin{cases} \frac{3}{4e^2} \left(\frac{T^{-1}}{H_{1,1}(\ln(Tq))} \left(\ln(K) - 1\right)\right)^2 c(1) \left(2\ln(q) - 1\right)^2 & \text{if } P(1,3/2) \\ \frac{3}{4e^2} \left(\frac{T^{1/3}}{H_{1,2/3}(\ln(Tq))} \left(\ln(K) - 1\right)\right)^{3/2} c(1) \left(2\ln(q) - 1\right)^2 & \text{if } P(3/2,\infty) \end{cases};$$

where $H_{1,\alpha}(\ln(Tq)) := \sum_{i=1}^{\ln(Tq)} \frac{1}{i^{\alpha}}$ for $\alpha > 0$. Then, for every $a \in (0, a_s)$, the error probability of R-UCBE is bounded by:

$$e_T(\text{R-UCBE}) \leq 2TK \exp\left(-\frac{a}{10}\right).$$

Proof. We recall that Assumption 2.2 states that all the increment functions $\gamma_i(n)$ are such that $\gamma_i(n) \leq c\ln(n)$. We use such a fact to provide an explicit solution for the optimal value of a. From Theorem 4.1 and using the fact that $\gamma_i(n) \leq c\ln(n)$, we have that Equation (6) becomes:

$$\frac{Tc}{\ln(1)\left(2\ln(q) - 1\right)} \left(\frac{2tT}{\ln(y)\, u^{\frac{3}{2}}}\right)^{\frac{1}{2}} \geq \gamma_i(Tq). \tag{24}$$

Or, more restrictively:

$$\frac{Tc\ln(1)\left(2\ln(q) - 1\right)}{\ln(y)\left(y - 1\right)} \left(\frac{2T^{-\frac{3}{2}}}{\ln(y)\left(y-1\right)^{\frac{3}{2}}}\right)^{\frac{1}{2}} \geq \gamma_i(Tq).$$

Let us solve Equation (24) by analyzing separately the two cases in which one of the two terms in the l.h.s. of such equation become prevalent.

Case 1: $P(3/2, \infty)$ In this branch, we can upper bound the left-side part of the inequality in Equation (24) by:

$$\frac{Tc\ln(1)\left(2\ln(q) - 1\right)}{\ln(y)\left(y-1\right)^{\frac{3}{2}}} \left(\frac{2T^{-\frac{3}{2}}}{\ln(y)\left(y-1\right)^{\frac{3}{2}}}\right)^{\frac{1}{2}} \geq \gamma_i(Tq).$$

Thus, we can derive:

$$\gamma_i(a\ln(q)) \leq 1 \left(\frac{Tc\ln(1)\left(2\ln(q) - 1\right)}{\gamma_i(Tq)} \left(2T^{-\frac{3}{2}}\right)^{\frac{1}{2}}\right)^{\frac{2}{3}}. \tag{25}$$

Using the above value in Equation (23), provides:

$$T - \sum_{i=1}^{\ln(Tq)} \gamma_i(a\ln(q)) > 0$$

$$T - \left(\ln(K) - 1\right) \left(Tc\ln(1)\left(2\ln(q) - 1\right)\right)^{\frac{2}{3}} \left(2T^{-\frac{3}{2}}\right)^{\frac{1}{3}} \sum_{i=1}^{\ln(Tq)} \frac{1}{i^{\frac{2}{3}}} > 0$$

$$T - \left(\ln(K) - 1\right) \left(Tc\ln(1)\left(2\ln(q) - 1\right)\right)^{\frac{2}{3}} \left(2T^{-\frac{3}{2}}\right)^{\frac{1}{3}} H_{1,2/3}(\ln(Tq)) > 0$$

$$a < \frac{\left(\frac{T^{1/3} - T^{2/3}\left(\ln(K) - 1\right)}{H_{1,2/3}(\ln(Tq))}\right)^{\frac{3}{2}} c(1)\left(2\ln(q) - 1\right)^2}{4e^2 - 3}$$

$$a \quad \frac{\frac{\sqrt{T^{1\{3}\ \sqrt{K\ \ 1}\sqrt{\frac{3}{2}}}{\sqrt{H_{1;2\{3}\sqrt{T}\sqrt{\frac{3}{2}}}\ \ c1\ \ 2''}}{4^{\ 2''\ 3}};$$

where the last expression is obtained by observing that $T \geq 1$ and for obtaining a more manageable expression, under the assumption that $\frac{\sqrt{T^{1\{3}\ \sqrt{K\ \ 1}\sqrt{\frac{3}{2}}}{\sqrt{H_{1;2\{3}\sqrt{T}\sqrt{\frac{3}{2}}}\ \ c1\ \ 2''} \geq 0.$

This implies a constraint on the minimum time budget T, which explicit form for the case $\in (\frac{3}{2}; 8]$ is provided in Lemma C.7

Case 2: $\in (1; \frac{3}{2}]$ In this case, we enforce the more restrictive condition:

$$\frac{Tc1\ \ 2''}{y\ \ 1} \quad \frac{2T\ ''\ \frac{3}{2}a^{\frac{1}{2}}}{y\ \ 1} \geq \ _i(T);$$

the value for the number of pulls is:

$$y_i(a) = 1 \quad \frac{Tc1\ \ 2''\ \ 2T\ ''\ \frac{3}{2}a^{\frac{1}{2}}}{\ _i(T)}^{\frac{1}{}}: \tag{26}$$

and the value for a becomes:

$$T \quad \sum_{i\ i\ (T)} N_{i;T} > 0$$

$$T \quad K\ \ 1) \quad Tc1\ \ 2'') \quad 2T\ ''\ \frac{3}{2}a^{\frac{1}{2}}\Big)^{\frac{1}{}} \quad \sum_{i\ i\ (T)} \frac{1}{\ _i(T)^{\frac{1}{}}} > 0$$

$$T \quad K\ \ 1) \quad Tc1\ \ 2'') \quad 2T\ ''\ \frac{3}{2}a^{\frac{1}{2}}\Big)^{\frac{1}{}} \quad H_{1;1\{}(T) > 0$$

$$a \quad \frac{\frac{\sqrt{T^{1\ 1\{}\ \sqrt{T}^{\ 1\{}\sqrt{K\ \ 1}\sqrt{}}{\sqrt{H_{1;1\{}\sqrt{T}\sqrt{}}\ \ c1\ \ 2''}{4^{\ 2''\ 3}}$$

$$a \quad \frac{\frac{\sqrt{T^{1\ 1\{}\ \sqrt{K\ \ 1}\sqrt{}}{\sqrt{H_{1;1\{}\sqrt{T}\sqrt{}}\ \ c1\ \ 2''}{4^{\ 2''\ 3}};$$

where the last expression is obtained by observing that $T \geq 1$ and for obtaining a more convenient expression, under the assumption that $\frac{\sqrt{T^{1\ 1\{}\ \sqrt{K\ \ 1}\sqrt{}}{\sqrt{H_{1;1\{}\sqrt{T}\sqrt{}}\ \ c1\ \ 2''} \geq 0.$

Also here, this implies a constraint on the minimum time budget T for the case $\in (1; \frac{3}{2}]$, which explicit form is provided in Lemma C.7 □

Lemma C.7. *Under Assumptions 2.1 and 2.2, the minimum time budget T for which the theoretical guarantees of R-UCBE hold is:*

$$T \geq \begin{cases} c^{\frac{1}{}}1\ \ 2'')^{\ 1}H_{1;1\{}(T)\ \ K\ \ 1)^{\overline{\ \ 1}} & \text{if } \in (1; 3\{2] \\ c^{\frac{2}{3}}1\ \ 2'')^{\ \frac{2}{3}}\ H_{1;2\{3}(T)\ \ K\ \ 1)^{\ 3} & \text{if } \in [3\{2; 8) \end{cases}$$

and $H_{1;}(T) := \sum_{i\ i\ (T)} \frac{1}{\ _i(T)^{\frac{1}{}}}$ *for* > 1.

Proof. Given Corollary 4.2, we want to find the values of T such that a value of $a \in (0; a\ s]$ should exist. This implies having $a > 0$. Given the value of , we can derive a lower bound for the time budget T.

Case 1: $\alpha \in (\frac{3}{2}, 8)$:

$$\frac{(T^{1-\alpha/3} - K^{1-\alpha/3})^{\frac{3}{2}}}{H_{1:2-\alpha/3}(T-K)^{\frac{3}{2}}} \geq c(1-\alpha)\varepsilon \geq 0.$$

From this, it follows:

$$T \geq \left(c^{2-\alpha/3}(1-\alpha)\varepsilon^{2-\alpha/3} H_{1:2-\alpha/3}(T-K) + K^{1-\alpha/3}\right)^{3}.$$

Case 2: $\alpha \in (1, \frac{3}{2})$:

$$\frac{(T^{1-\alpha} - K^{1-\alpha})}{H_{1:1-\alpha}(T-K)} \geq c(1-\alpha)\varepsilon \geq 0.$$

From this, it follows:

$$T \geq \left(c^{\frac{1}{1-\alpha}}(1-\alpha)\varepsilon^{1} H_{1:1-\alpha}(T-K) + K^{1-\alpha}\right)^{\frac{1}{1-\alpha}}.$$

□

C.3 Proofs of Section 5

In this appendix, we provide the proofs we have omitted in the main paper for what concerns the theoretical results about R-SR. We recall that with a slight abuse of notation, as done in Section 5, we denote with $\Delta_{(i)}(T)$ the i^{th} gap rearranged in increasing order, i.e., we have $\Delta_{(i)}(T) \leq \Delta_{(j)}(T)$ for $i \leq j$.

Lemma C.8. *For every arm $i \in [K]$ and every round $t \in [T]$, let us define:*

$$\bar{\Delta}_i(t) \coloneqq \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}+1}^{h_{i,t-1}} \Delta_i(l),$$

if $N_{i,t} \geq 2$, then $\Delta_i(t) \geq \bar{\Delta}_i(t)$, and if $h_{i,t} > N_{i,t}-2$, it holds that:

$$\Delta_i(T) - \bar{\Delta}_i(N_{i,t}) \leq T - \Delta_i(t)N_{i,t}-2. \tag{27}$$

Proof. The proof follows trivially from Lemma C.2. □

Lemma C.9 (Lower Bound for the Time Budget for R-SR)**.** *Under Assumptions 2.1 and 2.2, the R-SR algorithm is s.t. the minimum value for the horizon T ensuring that $\forall j \in [K-1]$ and $\forall i \in [K]$:*

$$T - \Delta_{(j)}(N_j - 1) \leq \frac{K - 1 - j}{2}\Delta_{(j)}(T),$$

is:

$$T \geq c^{\frac{1}{1-\alpha}} 2^{\frac{1}{1-\alpha}} \overline{\log}(K)^{\frac{1}{1-\alpha}} \max_{i \in [2,K]} \left(i^{\frac{1}{1-\alpha}} \Delta_{(i)}(T)^{\frac{1}{1-\alpha}}\right).$$

Proof. First, using Assumption 2.2, we derive an upper bound on the bias between $\Delta_{(i)}(T)$ and $\bar{\Delta}_{(i)}(N_j)$ (r.h.s. of Equation 27), where N_j is a generic time corresponding to the end of a phase of the R-SR algorithm:

$$T - \Delta_{(j)}(N_j - 2) \leq cT(N_j - 2)^{-\alpha}.$$

Substituting the definition of N_j into the above equation, we get:

$$\begin{aligned} T(N_j - 2)^{-\alpha} &\geq T - \frac{1}{\overline{\log}(K)}\frac{T-K}{K-1-j} - 1 \\ &\geq T - \frac{1}{\overline{\log}(K)}\frac{T}{K-1-j} - 1 \end{aligned} \tag{28}$$

$$\Delta_T \geq \frac{T}{\overline{\log}(K)\,(K-1-j)}. \tag{29}$$

Requiring that, for a generic N_j, the maximum possible bias is lower than a fraction of the suboptimality gap of arm $K-1-j$:

$$\begin{aligned}
cT\sqrt{N_j\{2} &\leq \frac{(K-1-j)\mu(T)}{2} \\
cT\,\frac{T}{2\overline{\log}(K)\,(K-1-j)} &\leq \frac{(K-1-j)\mu(T)}{2} \\
T^{1-\alpha} &\leq \frac{(K-1-j)\mu(T)}{c2^{1-\alpha}\,\overline{\log}(K)\,(K-1-j)} \\
T &\leq \left(\frac{(K-1-j)\mu(T)}{c^{\frac{1}{1-\alpha}}2^{\frac{1}{1-\alpha}}\,\overline{\log}(K)\,(K-1-j)^{\frac{1}{1-\alpha}}}\right)^{\frac{1}{1-\alpha}}.
\end{aligned}$$

Requiring that the above condition holds for all the phases $j \in \{0,\dots,K-2\}$ we have:

$$\begin{aligned}
T &\leq \max_{j\in\{0,\dots,K-2\}} \left(\frac{(K-1-j)\mu(T)}{c^{\frac{1}{1-\alpha}}2^{\frac{1}{1-\alpha}}\,\overline{\log}(K)\,(K-1-j)^{\frac{1}{1-\alpha}}}\right)^{\frac{1}{1-\alpha}} \\
&\leq c^{\frac{1}{1-\alpha}}2^{\frac{1}{1-\alpha}}\overline{\log}(K)^{\frac{1}{1-\alpha}} \max_{j\in\{0,\dots,K-2\}} \left(\frac{(K-1-j)\mu(T)}{K-1-j}\right)^{\frac{1}{1-\alpha}} \\
&\leq c^{\frac{1}{1-\alpha}}2^{\frac{1}{1-\alpha}}\overline{\log}(K)^{\frac{1}{1-\alpha}} \max_{j\in\{0,\dots,K-2\}} \left((K-1-j)^{\frac{1}{1-\alpha}}\mu(T)^{\frac{1}{1-\alpha}}\right) \\
&\leq c^{\frac{1}{1-\alpha}}2^{\frac{1}{1-\alpha}}\overline{\log}(K)^{\frac{1}{1-\alpha}} \max_{i\in\{2,\dots,K\}} i^{\frac{1}{1-\alpha}}\,\mu(T)^{\frac{1}{1-\alpha}}.
\end{aligned}$$

□

Theorem 5.1. *Under Assumptions 2.1 and 2.2, if the time budget T satisfies:*

$$T \geq 2^{\frac{1}{1-\alpha}}c^{\frac{1}{1-\alpha}}\overline{\log}(K)^{\frac{1}{1-\alpha}} \max_{i\in\{2,\dots,K\}} \left(i^{\frac{1}{1-\alpha}}\,\mu(T)^{\frac{1}{1-\alpha}}\right); \tag{10}$$

then, the error probability of R-SR is bounded by:

$$e_T(\text{R-SR}) \leq \frac{K(K-1)}{2}\exp\left(-\frac{\varepsilon^2}{8}\,\frac{T-K}{\overline{\log}(K)H_2}\right);$$

where $H_2 = \max_{i\in\{1,\dots,K\}} i\,(\Delta_{(i)})^{-2}$ and $\overline{\log}(K) = \frac{1}{2}+\sum_{i=2}^{K}\frac{1}{i}$.

Proof. The R-SR algorithm makes an error when at the end of a phase j the optimal arm has a pessimistic estimator $\hat{\Delta}_1(N_j)$ is smallest among the arms, formally:

$$\begin{aligned}
e_T(\text{R-SR}) &\leq \sum_{j=0}^{K-2}\sum_{i=1}^{K-j-1}\mathbb{P}\left\{\hat{\Delta}_{(1)}(N_j) \geq \hat{\Delta}_{(i)}(N_j)\right\} \\
&\leq \sum_{j=0}^{K-2}(K-1-j)\,\mathbb{P}\left\{\hat{\Delta}_{(1)}(N_j) \geq \hat{\Delta}_{(i)}(N_j)\right\} \\
&\leq \sum_{j=0}^{K-2}\sum_{i=K-j}^{K}\mathbb{P}\left\{\hat{\Delta}_{(1)}(N_j) \geq \hat{\Delta}_{(i)}(N_j)\right\};
\end{aligned}$$

where we use a union bound over the phases and over the arms still in the available arm set X_{j-1} in each phase. Let us focus on $\mathbb{P}\left[\hat{\mu}_{1,j} \geq \hat{\mu}_{i,j}\right]$. We have that the optimal arm has a smaller pessimistic estimator than the i^{th} one when:

$$\hat{\mu}_{i,j} - \hat{\mu}_{1,j} \leq 0$$

$$\hat{\mu}_{i,j} - \hat{\mu}_{1,j} \leq 0$$

$$\hat{\mu}_{1,j} - \sqrt{\frac{\alpha T}{2N_{1,j}}} \leq \hat{\mu}_{i,j} + \sqrt{\frac{\alpha T}{2N_{i,j}}} \tag{30}$$

$$\underbrace{\hat{\mu}_{1,j} - \sqrt{\frac{\alpha T}{2N_{1,j}}}}_{\leq T_{1,j}} \leq \hat{\mu}_{i,j} + \underbrace{\sqrt{\frac{\alpha T}{2N_{i,j}}}}_{\leq \mu_{i,j}} \tag{31}$$

$$\hat{\mu}_{1,j} - \mu_{1,j} \leq \hat{\mu}_{i,j} - \mu_{i,j} + \sqrt{\frac{\alpha T}{2N_{i,j}}} \tag{32}$$

where we added $\mu_{i,j}$ to derive Equation (30), and added $\mu_{1,j}$ to derive Equation (31), we used the results in Lemma C.8 and from the fact that the reward function is increasing. Since we are with a time budget T satisfying Theorem C.9, we have that:

$$T_{1,j} \geq \frac{\mu_{i,j}}{2}. \tag{33}$$

Substituting into Equation (32) the above, we have:

$$\hat{\mu}_{1,j} - \mu_{1,j} \leq \hat{\mu}_{i,j} - \mu_{i,j} \geq \frac{\mu_{i,j}}{2},$$

and the error probability becomes:

$$e_T(\text{R-SR}) \leq \sum_{j=1}^{K-1} \sum_{i=j+1}^{K} \mathbb{P}\left[\hat{\mu}_{1,j} - \mu_{1,j} \leq \hat{\mu}_{i,j} - \mu_{i,j} \geq \frac{\mu_{i,j}}{2}\right].$$

For the previous argumentation, we apply the Azuma-Hoëffding's inequality to the latter probability:

$$e_T(\text{R-SR}) \leq \sum_{j=1}^{K-1} \sum_{i=j+1}^{K} \exp\left(-\frac{\varepsilon N_j \left(\frac{\mu_{i,j}}{2}\right)^2}{2\sigma^2}\right)$$

$$\leq \sum_{j=1}^{K-1} j \exp\left(-\frac{\varepsilon N_j}{8\sigma^2} \Delta_{K-j}^2\right).$$

Now, given that:

$$\frac{\varepsilon N_j}{8\sigma^2} \Delta_{K-j}^2 \geq \frac{\varepsilon}{8\sigma^2} \frac{T-K}{\overline{\log}(K)(K-j)} \Delta_{K-j}^2$$

$$\geq \frac{\varepsilon}{8\sigma^2} \frac{T-K}{\overline{\log}(K) H_2(T)},$$

we finally derive the following:

$$e_T(\text{R-SR}) \leq \frac{K(K-1)}{2} \exp\left(-\frac{\varepsilon}{8\sigma^2} \frac{T-K}{\overline{\log}(K) H_2(T)}\right),$$

which concludes the proof. □

C.4 Proofs of Section 6

In this appendix, we provide the proofs of the lower bound on the error probability presented in Section 6.

Theorem 6.1. *For every algorithm* A*, there exists a deterministic SRB satisfying Assumptions 2.1 and 2.2 with* $K \geq 8^{1/(\rho-1)}$ *such that the optimal arm* $i^*(T)$ *cannot be identified for some time budgets* T *unless:*

$$T \geq H_{1;1/(\rho-1)}(T) \sum_{i \neq i^*(T)} \frac{1}{\Delta_i^{\frac{1}{1-\rho}}}. \tag{11}$$

Proof. We define for every suboptimal arm $i \in [2, K]$ the suboptimality gap reached at $T \to \infty$ as $\Delta_i \in (0, 1/2]$. We consider the base instance ν (see Figure 5) in which define the (deterministic) reward functions are defined for $\lambda > 1$ and $n \in \mathbb{N}$ as:

$$\nu_1(n) = \frac{1}{2} + \left(\frac{1}{2}\right)^{\frac{1}{n^{\rho-1}}},$$

$$\nu_i(n) = \min\left\{\underbrace{\frac{1}{2} + \left(\frac{1}{2}\right)^{\frac{1}{n^{\rho-1}}}}_{:=\nu_1(n)}, \frac{1}{2} - \Delta_i\right\} \qquad i \in [2, K].$$

Clearly, ν fulfills Assumption 2.1 and it is simple to show that also Assumption 2.2 is satisfied. Indeed, by first-order Taylor expansion:

$$\nu_1(n) - \nu_1(n+1) = \nu_1(n) - \nu_1(n+1) \leq \sup_{x \in [n, n+1]} \left|\frac{\partial}{\partial x} \nu_1(x)\right| \tag{34}$$

$$\frac{1}{2} \sup_{x \in [n, n+1]} x^{-\rho} \leq \frac{1}{2} n^{-\rho};$$

$$\nu_i(n) - \nu_i(n+1) = \nu_i(n) - \nu_i(n+1) \leq \sup_{x \in [n, n+1]} \left|\frac{\partial}{\partial x} \nu_i(x)\right|$$

$$(1-\rho) \frac{1}{2} \Delta_i \sup_{x \in [n, n+1]} x^{-\rho} \leq (1-\rho) n^{-\rho}$$

Thus, we simply take $c = 1$ in Assumption 2.2. Let us define n_i the number of pulls in which arm $i \in [2, K]$ reaches the stationary behavior:

$$\frac{1}{2} + \Delta_i = \frac{1}{2} + \left(\frac{1}{2}\right)^{\frac{1}{n_i^{\rho-1}}} \iff n_i = \left(\frac{1}{2\Delta_i}\right)^{\frac{1}{1-\rho}}.$$

A sufficient condition on the time budget so that the optimal arm is 1 (i.e., $i^*(T) = 1$) is given by $T \geq T_0$, where T_0 is the point in which the curve of the optimal arm intersects that of any of the suboptimal arms $i \in [2, K]$:

$$\frac{1}{2}(1 - T_0^{-1}) + \left(\frac{1}{2}\right)^{\frac{1}{T_0^{\rho-1}}} = \max_{i \in [2,K]} \left(\frac{1}{2} - \Delta_i\right)^{\frac{1}{1-\rho}}.$$

Consider now the regime in which $T \geq T_0$. We proceed by contradiction. Suppose that there exists an algorithm A that identifies the optimal arm such that on the bandit ν and that the suboptimal arm $i \in [2, K]$ has an expected number of pulls satisfying:

$$\mathbb{E}[N_i(T)] \leq n_i. \tag{35}$$

Consider now the alternative bandit ν^1 constructed from ν by keeping all the arms unaltered, except for arm i that is made optimal:

$$\nu^1_i(n) = \frac{1}{2} + \left(\frac{1}{2}\right)^{\frac{1}{n^{\rho-1}}};$$

$$\nu_j^1(n) = \nu_j(n), \quad \forall j \in [K] \setminus \{i\}.$$

Clearly the bandit ν^1 fulfills Assumption 2.1 and, with calculations similar to those in Equation (34), we conclude that it satisfies Assumption 2.2 with $c = 1$. A sufficient condition on T for which arm i is optimal in bandit ν^1 is that $T \geq T_2$ in which the curve of arm i intersects that of the arms j such that $\Delta_j \geq \Delta_i$:

$$\frac{1}{2}\Delta_i(1 - T^{-1}) \leq \frac{1}{2}\Delta_j \iff T \geq \max_{j \in [K]: \Delta_j \geq \Delta_i} \left(\frac{1/2}{\Delta_i - \Delta_j}\right)^{\frac{1}{1-\alpha}}.$$

Thus, we take:

$$T_2 := \left(\frac{1/2}{\Delta_i}\right)^{\frac{1}{1-\alpha}}.$$

Clearly, for $T \geq T_2$ since all the suboptimality gaps are at most $1/2$. Thus, we continue in the regime $T \geq T_2$. Since $\nu_i^1(n) = \nu_i(n)$ if $n \leq n_i$, it follows that under condition (35), algorithm $\mathcal{A}$ cannot distinguish between the two bandits and, consequently, cannot identify the optimal arm on bandit ν^1. Thus, it must follow, from the contradiction, that:

$$\mathbb{E}_{\nu}[N_i(T)] \geq n_i.$$

By summing over $i \in [2, K]$, we obtain:

$$T \geq \sum_{i \in [2,K]} \mathbb{E}_{\nu}[N_i(T)] \geq \sum_{i \in [2,K]} n_i \geq \sum_{i \in [2,K]} \left(\frac{1/2}{\Delta_i}\right)^{\frac{1}{1-\alpha}} \geq \sum_{i \in [2,K]} \left(\frac{1}{4\Delta_i}\right)^{\frac{1}{1-\alpha}} := T^{\star}. \tag{36}$$

Thus, we have found an interval $T \in [T_2, T^{\star})$ in which identification cannot be performed. Notice that it is simple to enforce that $T^{\star} \geq T_2$ with a sufficiently large number of arms $K \geq 2^{\frac{1}{1-\alpha}}$.

To conclude, we need to relate Δ_i with $\Delta_i(T)$ and $\Delta_i^1(T)$. We perform the computation for both the instances and ν^1, in the regime $T \geq 2^{\frac{1}{1-\alpha}} T_2$. Let us start with :

$$\Delta_i(T) \geq \frac{1}{2}(1 - T^{-1}) - \Delta_i \geq \frac{1}{2} - \Delta_i \geq \frac{1}{2}T^{-1} \geq \frac{\Delta_i}{2}, \quad \forall i \in [2, K]$$

We move to ν^1:

$$\Delta_1^1(T) \geq \frac{1}{2} - \Delta_i(1 - T^{-1}) - \frac{1}{2}(1 - T^{-1}) \geq \Delta_1(1 - T^{-1}) \geq \frac{\Delta_i}{2};$$

$$\Delta_i^1(T) \geq \frac{1}{2} - \Delta_i(1 - T^{-1}) - \frac{1}{2} + \Delta_i \geq \Delta_i T^{-1} \geq \frac{\Delta_i}{2}, \quad \forall i \in [2, K].$$

Thus, a necessary condition for the correct identification of the optimal arm is:

$$T \geq \sum_{i \in [2,K]} \left(\frac{1/2 - 2\Delta_i(T)}{4\Delta_i(T)}\right)^{\frac{1}{1-\alpha}} \geq \sum_{i \in [2,K]} \left(\frac{1}{8\Delta_i(T)}\right)^{\frac{1}{1-\alpha}} \geq 2^{\frac{1}{1-\alpha}} T^{\star}.$$

Similarly, with $K \geq 8^{\frac{1}{1-\alpha}}$, we can enforce $2^{\frac{1}{1-\alpha}} T^{\star} \geq 2^{\frac{1}{1-\alpha}} T_2$. □

Theorem 6.2. *For every algorithm $\mathcal{A}$ run with a time budget T fulfilling Equation* (11)*, there exists a SRB satisfying Assumptions 2.1 and 2.2 with $K \geq 8^{1/(1-\alpha)}$ such that the error probability is lower bounded by:*

$$e_T(\mathcal{A}) \geq \frac{1}{4}\exp\left(\frac{-8T}{\sigma^2 H_{1:2}(T)}\right), \quad \textit{where} \quad H_{1:2}(T) := \sum_{i \neq i_1(T)} \frac{1}{\Delta_i^2(T)}.$$

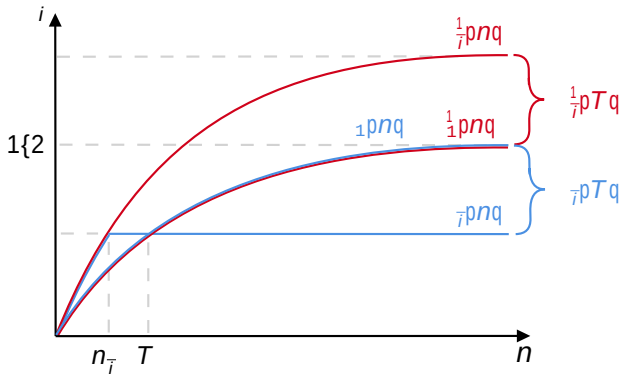


Figure 5: Instances and ^{1} of SRB used in Theorem 6.1 and Theorem 6.2.

Proof. The proof imports the technique from (Kaufmann et al., 2016, Theorem 16 and 17). We consider the Gaussian bandit with variance ^{2} equal for all the arms and the expected reward $_{i}\mathrm{p}n\mathrm{q}$ as in the base instance of proof of Theorem 6.1 (see Figure 5). Let us define by convention $_{1}$ $_{2}$. Let A be an algorithm, it is simple to show that there exists an arm i P JKK, such that:

$$\mathrm{E}\mathrm{r}N_{i}\mathrm{p}T\mathrm{q}\mathrm{s} \text{ ¤ } \frac{T}{H_{2}\ ^{2}_{i}};$$

where H_{2} $^{\circ}_{i}$ $^{K}_{1}$ $_{i}$ ^{2}. We consider two cases. Suppose that i 1 and we construct the alternative Gaussian bandit ^{1} with the same variance ^{2} and the expected rewards defined as follows:

$$\begin{aligned} {}^{1}_{1}\mathrm{p}T\mathrm{q} &\quad \min\ \frac{1}{2}\ 1\quad n^{1}\quad ;\frac{1}{2}\quad 2\ _{1}\ ;^{*}_{"} \\ {}^{1}_{i}\mathrm{p}T\mathrm{q} &\quad {}_{i}\mathrm{p}T\mathrm{q}; \quad i \text{ P J2; } K\mathrm{K}: \end{aligned}$$

For T sufficiently large as in Theorem 6.1, while in bandit the optimal arm is 1, in bandit ^{1} the optimal arm is 2. Instead, suppose that i 1 and we construct the alternative Gaussian bandit ^{1} with the same variance ^{2} and the expected rewards defined as follows:

$$\begin{aligned} {}^{1}_{i}\mathrm{p}T\mathrm{q} &\quad \frac{1}{2}\quad {}_{i}\ \mathrm{p}1\quad n^{1}\quad \mathrm{q}; \\ {}^{1}_{i}\mathrm{p}T\mathrm{q} &\quad {}_{i}\mathrm{p}T\mathrm{q}; \quad i \text{ P } J K\mathrm{K}\mathrm{z}\mathrm{t}i\mathrm{u}: \end{aligned}$$

For T sufficiently large as in Theorem 6.1, while in bandit the optimal arm is 1, in bandit ^{1} the optimal arm is i. Let us denote with $_{i}\mathrm{p}t\mathrm{q}$ the distribution of the reward at time t for arm i. By the Bretagnolle-Huber's inequality, we obtain:

$$\begin{aligned} \max\mathrm{t}e_{T}\mathrm{p}\ \mathrm{q}; e_{T}\mathrm{p}\ ^{1}\mathrm{q}\mathrm{u} &¥ \frac{1}{4}\exp\quad \mathrm{E}\ \sum_{t\ 1}^{T}\ 1\mathrm{t}I_{t}\quad i\mathrm{u}D_{K\,L}\mathrm{p}\ {}_{i}\mathrm{p}t\mathrm{q};\ {}^{1}_{i}\mathrm{p}t\mathrm{q}\mathrm{q} \\ &\quad \frac{1}{4}\exp\quad \mathrm{E}\ \sum_{t\ 1}^{T}\ 1\mathrm{t}I_{t}\quad i\mathrm{u}\frac{\mathrm{p}\ {}_{i}\mathrm{p}N_{i;t}\mathrm{q}\quad {}^{1}_{i}\mathrm{p}N_{i;t}\mathrm{q}\mathrm{q}^{2}}{2\ ^{2}} \\ &¥ \frac{1}{4}\exp\quad \mathrm{E}\,\mathrm{r}N_{i}\mathrm{p}T\mathrm{q}\mathrm{s}\frac{\mathrm{p}2\ {}_{i}\mathrm{q}^{2}}{2\ ^{2}}\ :\quad \frac{1}{4}\exp\quad \frac{2T}{^{2}H_{2}} \\ &¥ \frac{1}{4}\exp\quad \frac{2T}{^{2}H_{2}}\quad ; \end{aligned}$$

where we observed that for every n P JTK, we have $|\ {}_{i}\mathrm{p}n\mathrm{q}\quad {}^{1}_{i}\mathrm{p}n\mathrm{q}|$ ¤ 2 $_{i}$. To conclude, we relate H_{2} with $H_{1;2}\mathrm{p}T\mathrm{q}$. Using an argument analogous to that of the last part of the proof Theorem 6.1 it is simple to observe that, for sufficiently large T, we have $_{i}$ ¤ 2 $_{i}\mathrm{p}T\mathrm{q}$ (condition already satisfied

by requesting the bound on T of Theorem 6.1), from which we have:

$$H_2 \sum_{i=1}^{K} {}_i^2 \quad {}_1^2 \sum_{i=2}^{K} {}_i^2 \leq \sum_{i=2}^{K} {}_i^2 \leq \frac{1}{4} \sum_{i=2}^{K} {}_i(T) \quad {}^2 \quad \frac{1}{4} H_{1;2}(T).$$

□

C.5 Auxiliary Lemmas

Lemma C.10 (Höeffding-Azuma's inequality for weighted martingales)**.** *Let $F_1 \subseteq \cdots \subseteq F_n$ be a filtration and $X_1, \ldots, X_n$ be real random variables such that X_t is F_t-measurable, $\mathbb{E}[X_t | F_{t-1}] = 0$ (i.e., a martingale difference sequence), and $\mathbb{E}[\exp(\lambda X_t) | F_{t-1}] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$ for any $\lambda > 0$ (i.e., σ^2-subgaussian). Let $\alpha_1, \ldots, \alpha_n$ be non-negative real numbers. Then, for every $\varepsilon > 0$ it holds that:*

$$\mathbb{P}\left[\sum_{t=1}^{n} \alpha_t X_t > \varepsilon\right] \leq 2 \exp\left(\frac{-2\varepsilon^2}{\sigma^2 \sum_{t=1}^{n} \alpha_t^2}\right).$$

Proof. For a complete demonstration of this statement, we refer to Lemma C.5 of Metelli et al. (2022). □

Lemma C.11. *Let $\alpha \geq 1$, then it holds that:*

$$H_{1;1/\alpha}(T)^{\alpha-1} + H_{1;1/(\alpha-1)}(T) \leq H_{1;1/\alpha}(T).$$

Proof. We prove the equivalent statement, being $\alpha \geq 1$:

$$H_{1;1/(\alpha-1)}(T)^{\alpha} + H_{1;1/\alpha}(T)^{\alpha-1} \geq H_{1;1/\alpha}(T).$$

Recalling that the function $(x+y)^{\alpha-1}$ is subadditive, being $\alpha - 1 \leq 1$, we have:

$$H_{1;1/(\alpha-1)}(T)^{\alpha} + H_{1;1/\alpha}(T)^{\alpha-1} \geq \sum_{i=1}^{K} \frac{1}{{}_i(T)^{1/(\alpha-1)}}^{\alpha-1} + \sum_{i=1}^{K} \frac{1}{{}_i(T)^{1/\alpha}}^{\alpha-1} \geq H_{1;1/\alpha}(T).$$

□

D Theoretical Analysis of a baseline: RR-SW

In this appendix, we provide the theoretical analysis for the algorithm `Round Robin Sliding Window` (`RR-SW`), as it represents the most intuitive baseline for this setting. First, we need to formalize the algorithm, whose pseudo-code is provided in Algorithm 3.

Algorithm 3: `RR-SW`.

Input : Time budget T, Number of arms K, Window size τ

```
Initialize t ← 1
Estimate N ← T/K
for i ∈ [K] do
    for l ∈ [N] do
        Pull arm i and observe x_t
        t ← t + 1
    end
    Update μ̂_i(N)
end
Recommend i(T) ∈ arg max_{i∈[K]} μ̂_i(N)
```

Algorithm The algorithm takes as input the time budget T and the number of arms K. Then, it computes the number of pulls $N = \frac{T}{K}$ we need to perform for each arm. After having computed the number of pulls, RR-SW plays all the arms N times in a round-robin fashion. After the N pulls, it estimates $\hat{\Delta}_i(N)$ using the last εN samples (i.e., the ones from $(1-\varepsilon)N$ to N). Finally, it recommends $I(T)$, corresponding to the one which the highest estimated $\hat{\Delta}_i(N)$.

Error probability bound Before presenting the error probability bound for RR-SW, we need to introduce $\Delta_{(2)}(T)$, which represents the minimum suboptimality gap at a given time budget T. It is actually the gap between the optimal arm and the first sub-optimal one. Formally: $\Delta_{(2)}(T) := \min_{i \neq i^*(T)} \{\Delta_i(T)\}$. Given this quantity, the error probability for the RR-SW algorithm can be bounded as follows.

Theorem D.1. *Under Assumptions 2.1 and 2.2, considering a time budget T satisfying:*

$$T \geq 2^{\frac{1}{\alpha-1}} c^{\frac{1}{\alpha-1}} (1-\varepsilon)^{\frac{1}{\alpha-1}} K^{\frac{\alpha}{\alpha-1}} \Delta_{(2)}(T)^{-\frac{1}{\alpha-1}}; \tag{37}$$

the error probability of RR-SW is bounded by:

$$e_T(\text{RR-SW}) \leq K \exp\left(-\frac{\varepsilon T}{8 K^2} \Delta_{(2)}(T)^2\right).$$

Some comments are in order. First, it is worth noting how, as expected, by increasing the number of samples considered in the estimator, we reduce the error probability $e_T(\cdot)$ at the cost of a more strict constraint on the time budget T. This is due to the request that the arms must be already separated at the beginning of the window we use to estimate the $\hat{\Delta}_i(N)$. Second, the error probability scales as an (inverse) function only of the smallest suboptimality gap $\Delta_{(2)}(T)$.

D.1 Proofs

Before demonstrating Theorem D.1, we need to introduce the following technical lemma.

Lemma D.2 (Lower Bound for the Time Budget)**.** *Under Assumptions 2.1 and 2.2, the RR-SW algorithm is s.t. the minimum value for the horizon T ensuring that $\forall i \in [K]$:*

$$T \Delta_i((1-\varepsilon)N) \geq \frac{\Delta_{(2)}(T)}{2};$$

where $N = \frac{T}{K}$ and $\varepsilon \in (0, 1)$ is:

$$T \geq 2^{\frac{1}{\alpha-1}} c^{\frac{1}{\alpha-1}} (1-\varepsilon)^{\frac{1}{\alpha-1}} K^{\frac{\alpha}{\alpha-1}} \Delta_{(2)}(T)^{-\frac{1}{\alpha-1}}.$$

Proof. First of all, we recall that $N = \frac{T}{K}$ is the number of times each arm has been pulled, considering K arms by running a round-robin procedure until we reach a time budget T. We consider the pessimistic estimator described in Section 3. Considering such an estimator and the RR-SW algorithm, which runs a round-robin procedure, what we get at the end of the time budget is a sliding-window estimator for the value of $\Delta_i(T)$, which will include the lasts $(1-\varepsilon)\frac{T}{K}$ samples. In this lemma, we want to find the minimum value of the time budget T for which, at the first samples we consider, the real process of the arms are separated by at least $\frac{\Delta_i(T)}{2}$. In this estimator, we consider samples in the range of $[(1-\varepsilon)\frac{T}{K}; \frac{T}{K}]$, so we need to ensure, given Assumption 2.1, that:

$$T \Delta_i((1-\varepsilon)N) \geq \frac{\Delta_{(2)}(T)}{2}. \tag{38}$$

Given that, for Assumptions 2.1 and 2.2, it holds:

$$T \Delta_i((1-\varepsilon)N) \geq T \Delta_i((1-\varepsilon)N)$$
$$T \Delta_i\left((1-\varepsilon)\frac{T}{K}\right)$$

$$\Delta_i \geq c\left(1, \sqrt{\frac{T}{K}}\right) \quad . \tag{39}$$

By introducing the term derived in Equation (39) into Equation (38) we obtain:

$$T \geq c\left(1, \sqrt{\frac{T}{K}}\right) \Delta \geq \frac{\Delta_{2}(T)}{2}.$$

This implies that the minimum time budget T which guarantees the initial condition of Equation (39) is:

$$T \geq 2^{\frac{1}{\alpha}} c^{\frac{1}{\alpha}}\left(1, \sqrt{\frac{1}{K}}\right)^{\frac{1}{\alpha}} \Delta_{2}(T)^{-\frac{1}{\alpha}},$$

where $\Delta_{2}(T)$ is the minimum suboptimality gap ($\Delta_{2}(T) \triangleq \min_{i} \{\Delta_i(T) > 0\}$). □

Now, we can find the error probability $e_T(\texttt{RR-SW})$, which will hold for all the time budgets which satisfy the condition of Lemma D.2.

Theorem D.1. *Under Assumptions 2.1 and 2.2, considering a time budget T satisfying:*

$$T \geq 2^{\frac{1}{\alpha}} c^{\frac{1}{\alpha}}\left(1, \sqrt{\frac{1}{K}}\right)^{\frac{1}{\alpha}} \Delta_{2}(T)^{-\frac{1}{\alpha}}, \tag{37}$$

the error probability of RR-SW is bounded by:

$$e_T(\textit{RR-SW}) \leq K \exp\left(-\frac{\varepsilon T}{8K^{2}}\Delta_{2}(T)^{2}\right).$$

Proof. The RR-SW algorithm makes an error in predicting the best arm when, at the end of the process (at T total pulls), the optimal arm has a pessimistic estimator $\hat{\mu}_1(N)$ that is not the highest among the arms (we consider w.l.o.g. that the best arm is the arm 1). Formally:

$$\begin{aligned} e_T(\texttt{RR-SW}) &\triangleq \mathbb{P}\left(\exists i \in \{2,\dots,K\} : \hat{\mu}_1(N) \leq \hat{\mu}_i(N)\right) \\ &\leq \sum_{i=2}^{K} \mathbb{P}\left(\hat{\mu}_1(N) \leq \hat{\mu}_i(N)\right). \end{aligned}$$

Let us focus on a single arm i, where we want to upper bound the probability that $\mathbb{P}(\hat{\mu}_1(N) \leq \hat{\mu}_i(N))$. Let us focus on the term inside the probability:

$$\hat{\mu}_i(N) \geq \hat{\mu}_1(N)$$

$$\hat{\mu}_i(N) \geq \hat{\mu}_1(N) \geq 0$$

$$\hat{\mu}_1(T) \geq \hat{\mu}_1(N) \geq \hat{\mu}_i(N) \geq \hat{\mu}_i(T) \tag{40}$$

$$\underbrace{\hat{\mu}_1(T) \geq \hat{\mu}_1(N)}_{\leq T - \Delta_1(N)} \geq \hat{\mu}_i(N) \geq \underbrace{\hat{\mu}_i(T)}_{\leq \mu_i} \tag{41}$$

$$\hat{\mu}_1(N) \geq \hat{\mu}_1(N) - \hat{\mu}_i(N) \geq \hat{\mu}_i(T) - \mu_i \geq T - \Delta_1(N) - \mu_i \tag{42}$$

where we added $\mu_i(T)$ to derive Equation (40), and added $-\hat{\mu}_1(N)$ to derive Equation (41), we used the results in Lemma C.8 and from the fact that the reward function is increasing. Considering a time budget T satisfying Theorem D.2, and $\Delta_i(T) \geq \Delta_{2}(T)$, $\forall i \in \{2,\dots,K\}$, we have that:

$$T - \Delta_1(N) \leq \frac{\Delta_i(T)}{2}. \tag{43}$$

Equation (43) holds since we are considering a time budget T which satisfies a more restrictive condition (we are considering a time budget at which this separation already holds for $1 - \varepsilon N$, so it also holds now).

Substituting Equation (43) into Equation (42) the above, we have:

$$\hat{\mu}_1(N) - \bar{\mu}_1(N) \geq \hat{\mu}_i(N) - \bar{\mu}_i(N) \geq \frac{\Delta_i(T)}{2},$$

and the error probability becomes:

$$e_T(\text{RR-SW}) \leq \sum_{i=1}^{K} \mathbb{P}\left[\hat{\mu}_1(N) - \bar{\mu}_1(N) \geq \hat{\mu}_i(N) - \bar{\mu}_i(N) \geq \frac{\Delta_i(T)}{2}\right].$$

For the previous argumentation, we apply the Azuma-Hoëffding's inequality and the union bound:

$$e_T(\text{RR-SW}) \leq \sum_{i=1}^{K} \exp\left(\frac{-N\left(\frac{\Delta_i(T)}{2}\right)^2}{2\sigma^2}\right)$$

$$\leq K \exp\left(\frac{-T}{8K\sigma^2}\right) \leq \frac{2}{(2T)}.$$

□

E On the necessity of Concavity Assumption

In Assumption 2.1, we introduce two conditions on the behavior of the expected payoffs over time: they must be non-decreasing and concave. If even one of these two lacks, the error probability cannot be guaranteed to be decreasing as a function of the budget T. From an intuitive perspective, this is similar to what happens for regret minimization in (Metelli et al., 2022, Theorem 4.2), in which the authors demonstrate the non-learnability (i.e., an $\Omega(T)$ cumulative regret lower bound) when these two assumptions do not hold.

From a technical perspective, we can demonstrate that the error probability no longer depends on T when we remove the *concavity* assumption, as stated in the Proposition below.

Proposition E.1. *For every algorithm $\mathcal{A}$, there exists a SRB satisfying the non-decreasing assumption ($\Delta_i(n) \geq 0, \forall i \in [K], n \in [T]$), such that the error probability is lower bounded by:*

$$e_T(\mathcal{A}) \geq \frac{1}{4}\exp\left(-\frac{1}{8}\right).$$

Proof. Consider two Gaussian bandits with unit variance. Let ν be a 2-armed bandit with expected rewards $\mu_1(t) = 1/2$ and $\mu_2(t) = 3/4$, both $\forall t \in [T]$, thus its optimal arm is $i^*(T) = 2$ (we recall that the optimal arm in our setting is the one having the highest expected reward at T). Let ν^1 be a 2-armed bandit with expected values $\mu_1(t) = 1/2$ $\forall t \in [T-1]$, $\mu_1(T) = 1$ and $\mu_2(t) = 3/4$ $\forall t \in [T]$, thus the optimal arm is $i^*_1(T) = 1$. Notice that bandit ν^1 violates the concavity assumption. Now, applying the Bretagnolle-Huber inequality, we have that:

$$\max\left(\mathbb{P}[\hat{I}(T) = 2], \mathbb{P}_1[\hat{I}(T) = 1]\right) \geq \frac{1}{4}\exp\left(-\mathbb{E}\left[\sum_{t=1}^{T} D_{\text{KL}}(\mu_{I_t}(N_{I_t,t}), \mu^1_{I_t}(N_{I_t,t}))\right]\right) \tag{44}$$

$$\geq \frac{1}{4}\exp\left(-D_{\text{KL}}(\mu_1(T), \mu^1_1(T))\right) \tag{45}$$

$$\geq \frac{1}{4}\exp\left(-\frac{1}{8}\right); \tag{46}$$

where $\hat{I}(T)$ is the arm recommended at time T and having observed that $D_{\text{KL}}(\mu_{I_t}(N_{I_t,t}), \mu^1_{I_t}(N_{I_t,t})) = 0$ if $t \neq T$ regardless the arm $I_t \in \{1, 2\}$ and that $D_{\text{KL}}(\mu_{I_T}(N_{I_T,T}), \mu^1_{I_T}(N_{I_T,T})) \leq D_{\text{KL}}(\mu_1(T), \mu^1_1(T)) = 1/8$. □

	b	c	
Arm 1	37	1	1
Arm 2	10	0.88	1
Arm 3	1	0.78	1
Arm 4	10	0.7	1
Arm 5	20	0.5	1

Table 2: Numerical values of the parameters characterizing the functions for the synthetically generated setting.

F Experimental Details

In this section, we provide all the details about the presented experiments.

The payoff functions characterizing the arms shown in Figure 2 belong to the family:

$$F \quad \left\{ f(x) \quad c \quad 1 \quad \frac{b}{(b^{1\{} \quad x)} \quad ; \right.$$

where $c;$ $\in (0; 1]$ and $b \in [0;$ $8)$. Note that, by construction, all the functions laying in F satisfy the Assumptions 2.1 and 2.2. In particular, the largest value of satisfying Assumption 2.2, for the setting presented in Section 8, is 1.3. In Table 2, we report the value of the parameters characterizing the function employed in the synthetically generated setting presented in the main paper.

F.1 Parameters Values for the Algorithms

This section provides a detailed view of the parameter values we employed in the presented experiments. More specifically, the parameters, which may still depend on the time budget T and on the number of arms K, are set as follows:

- UCB-E: for the exploration parameter a, we used the optimal value, i.e., the one that minimizes the upper bound of the error probability, as prescribed in Audibert et al. (2010), formally:
$$a \quad \frac{25(T \quad K)}{36 H_1};$$
where $H_1 \quad \sum_{i \ i} (T)^{-1}_{i}$;
- R-UCBE: we used the value prescribed by Corollary 4.2 where we set the value 1.3;
- ETC and Rest-Sure: we set 0.8 and U 1 as suggested by Cella et al. (2021).

F.2 Running Time

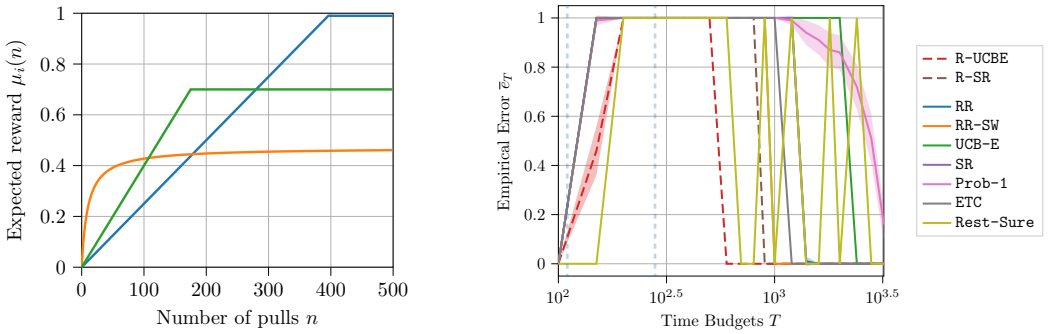
The code used for the results provided in this section has been run on an Intel(R) I7 9750H @ 2.6GHz CPU with 16 GB of *LPDDR4* system memory. The operating system was *MacOS* 13.1, and the experiments were run on *Python* 3.10. A run of R-UCBE over a time budget of T 3200 takes 0.07 seconds (on average), while a run of R-SR takes 0.06 seconds (on average).

G Additional Experimental Results

In this section, we present additional results in terms of empirical error $\bar{e}_T$ of R-UCBE, R-SR, and the other baselines presented in Section 8.

G.1 Challenging scenario

Here we test the algorithms on a challenging scenario in which we consider K 3 arms whose increment changes *abruptly*. The setting is presented in Figure 6a. The results corresponding to



(a) Synthetic setting.

(b) Results on synthetic setting.

Figure 6: Challenging scenario in which the arm reward increment rate changes abruptly.

such a setting are presented in Figure 6b. In this case, the last time the optimal arm does not change anymore is T 280. Similarly to the synthetic setting presented in the main paper, we have two different behaviors for time budgets T 280 and T ¡ 280. For short time budgets, the algorithm providing the best performance is `Rest-Sure`, and the second best is `R-UCBE`. Conversely, for time budgets T ¡ 550 `R-UCBE` provides a correct suggestion in most of the cases providing an error e_T(`R-UCBE`) 0:01. Instead, `Rest-Sure` is not consistently providing reliable suggestions. This is allegedly due to the fact that such an algorithm has been designed to work in less general settings than the one we are tackling. Even in this case, the `R-SR` starts providing a small value for the error probability after `R-UCBE` does, at T 1000. However, it is still better behaving than the other baseline algorithms. Note that `Rest-Sure` has a peculiar behavior. Indeed, it seems that even for large values of the time budget, it does not consistently suggest the optimal arm (i.e., the error probability does not go to zero). This is likely due to the nature of the parametric shape enforced by the algorithm, which may result in unpredictable behaviors when it does not reflect the nature of the real reward functions.

G.2 Sensitivity Analysis on the Noise Variance

In what follows, we report the analysis of the robustness of the analyzed algorithms as noise standard deviation changes in the collected samples. The setting we considered is the one described in Section 8. The results are provided in Figure 7. Let us focus on the performances of the `R-UCBE` algorithm. For small values of the standard deviation (0:01), we have the same behavior in terms of error probability, i.e., a progressive degradation of the performances for time budget T 150. Indeed, at this time budget, the expected rewards of 3 arms are close to each other, and determining the optimal arm is a challenging problem. However, the performances are better or equal to all the other algorithms even at this point. Conversely, for values of the standard deviation ¥ 0:05, the performance of `R-UCBE` starts to degrade, with behavior for 0:5 which is constant w.r.t. the chosen time budget with a value of e_T(`R-UCBE`) 0:8. This suggests that such an algorithm suffers in the case the stochasticity of the problem is significant. Let us focus on `R-SR`. This algorithm does not change its performances w.r.t. changes in terms of . Indeed, only for 0:5, we have that it does not provide an error probability close to zero for time budget T ¡ 1000. However, excluding `R-UCBE`, we have that the `R-SR` algorithm is the best/close to the best performing algorithm. This is also true in the case of 0:5, in which the `R-UCBE` fails in providing a reliable recommendation for the optimal arm with a large probability.

G.3 Real-world Experiment on IMDB dataset

Description We validate our algorithms and the baselines on an AutoML task, namely an *online best model selection* problem with a real-world dataset. We employ the IMDB dataset, made of 50; 000 reviews of movies (scores from 0 to 10). We preprocessed the data as done by Metelli et al. (2022), and run the algorithms for time budgets T P T : t500; 1000; : : : ; 15000; 20000; 30000u. A graphical representation of the reward (in this case, represented by the accuracy) of the different models is presented in Figure 8. Since, in this case, we only had a single realization to estimate the error probability e_T(A), we report the success rate R(A) instead, i.e., the ratio between the number

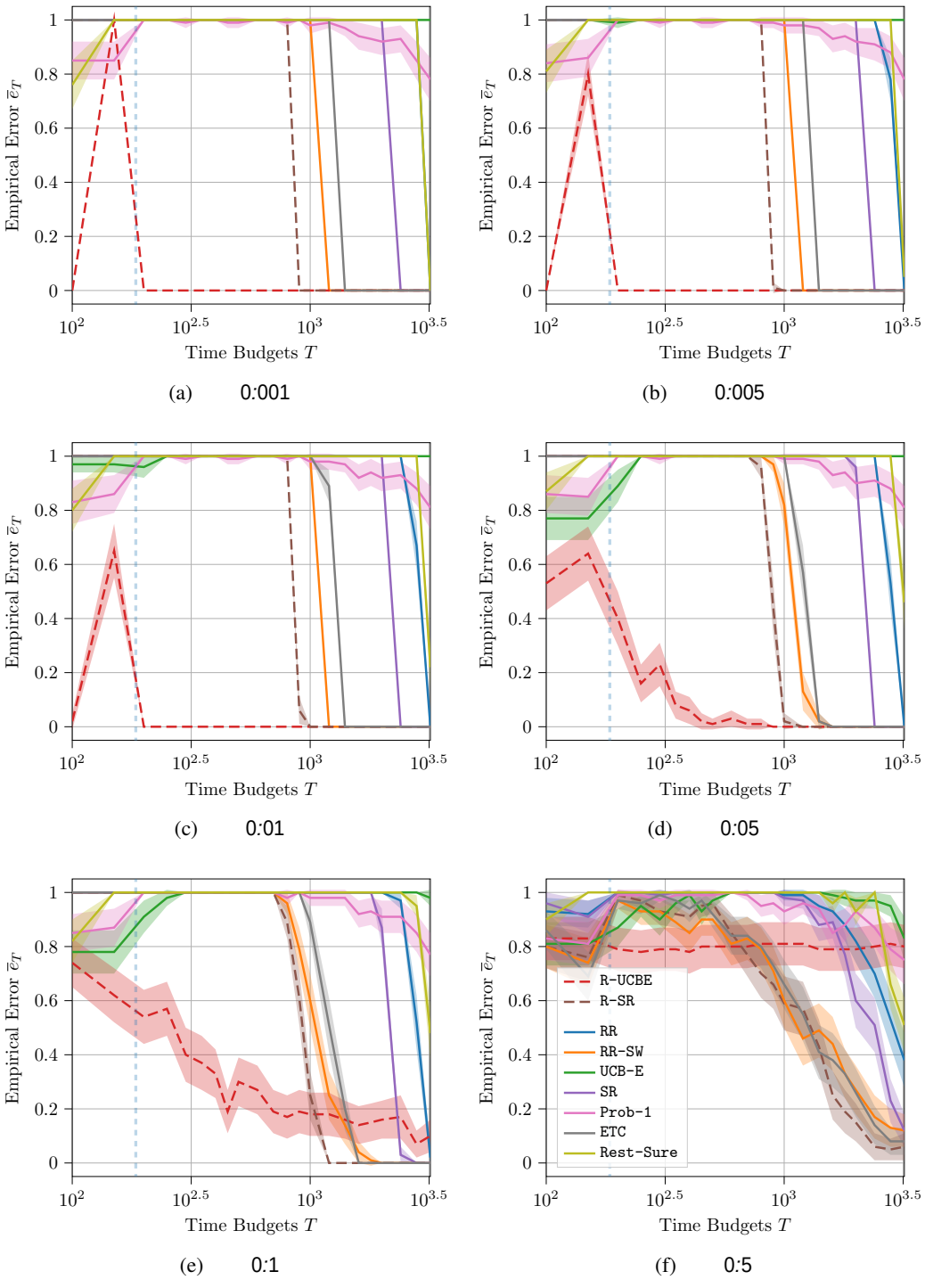


Figure 7: Empirical error probability for the synthetically generated setting, with different values of the noise standard deviation .

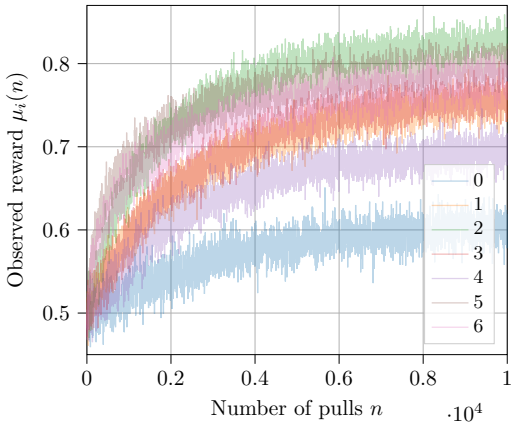


Figure 8: Rewards for the arms of the IMDB experiments.

		T											
		500	1000	2000	3000	4000	5000	7000	10000	15000	20000	30000	RpUq
	Optimal Arm	5	5	2	2	2	2	2	2	2	2	2	
Algorithms	R-UCBE (ours)	6	5	2	5	2	2	2	2	2	2	2	9{11
	R-SR (ours)	5	5	5	5	5	5	5	5	2	2	2	5{11
	RR	5	5	5	5	5	5	5	5	5	5	5	2{11
	RR-SW	5	5	5	5	5	5	5	5	5	2	2	4{11
	SR	5	5	5	5	5	5	5	5	5	5	2	3{11
	UCB-E	5	5	5	5	5	5	5	5	5	5	5	2{11
	Prob-1	1	5	2	5	5	5	5	1	5	6	2	3{11
	ETC	5	5	5	5	5	5	5	5	5	5	2	3{11
	Rest-Sure	6	5	2	2	2	1	0	2	5	0	2	6{11

Table 3: Optimal arm for different time budgets on the IMDB dataset (first row) and corresponding recommendations provided by the algorithms (second to last row). In the last column, we compute the corresponding success rate.

of times an algorithm provides a correct suggestion and the number of budget values we considered, formally defined as $RpAq: \frac{1}{|T|} \sum_{T \in T} \mathbb{1}\{\hat{I} = i\}$ (the larger, the better).

Results The results are reported in Table 3. The algorithm with the largest success rate $RpUq$ is the R-UCBE, while R-SR provides the third best success rate. Moreover, Rest-Sure, the only algorithm providing a success rate larger than R-SR, has issues with large time budgets since for $T \geq 5000$ is able to provide only 2 correct guesses of the optimal arm over 6 attempts. Conversely, our algorithms progressively provide more and more correct guesses as the time budget T increases. The above results on a real-world dataset corroborate the evidence presented above that the proposed algorithms outperform state-of-the-art ones for the BAI problem in SRB.