

## A Related Works

In this appendix, 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) for the first time. Their work and subsequently the one by Metelli et al. (2022) analyzed the problem from a regret minimization point of view.

**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 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.

## B 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.

## C 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  $\mathcal{O}(1)$  computational complexity at each step for the update of the estimates for the pessimistic estimator  $\hat{\mu}_i(t)$  and optimistic estimator  $\check{\mu}_i^T(t)$ .

With a slight abuse of notation, only in this appendix, for the sake of simplicity, we denote with  $\bar{x}_{i,n}$  the reward collected at the  $n^{\text{th}}$  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  $\hat{\mu}_i$  can be written as:

$$\hat{\mu}_i(t) = \frac{\bar{a}_i}{h_{i,t}},$$

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

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

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

Instead, the *optimistic* estimator  $\check{\mu}_i^T(t)$ , is updated using:

$$\check{\mu}_i^T(t) = \frac{1}{h_{i,t}} \left( \bar{a}_i + \frac{T(\bar{a}_i - \bar{b}_i)}{h_{i,t}} - \frac{\bar{c}_i - \bar{d}_i}{h_{i,t}} \right).$$

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:

$$\begin{aligned} \bar{b}_i &\leftarrow \begin{cases} \bar{b}_i + \bar{x}_{i,N_{i,t}-h_{i,t}} - \bar{x}_{i,N_{i,t}-2h_{i,t}} & \text{if } h_{i,t} = h_{i,t-1} \\ \bar{b}_i + \bar{x}_{i,N_{i,t}-2h_{i,t}+1} & \text{otherwise} \end{cases}, \\ \bar{c}_i &\leftarrow \begin{cases} \bar{c}_i + N_{i,t} \cdot \bar{x}_{i,N_{i,t}} - (N_{i,t} - h_{i,t}) \cdot \bar{x}_{i,N_{i,t}-h_{i,t}} & \text{if } h_{i,t} = h_{i,t-1} \\ \bar{c}_i + N_{i,t} \cdot \bar{x}_{i,N_{i,t}} & \text{otherwise} \end{cases}, \\ \bar{d}_i &\leftarrow \begin{cases} \bar{d}_i + N_{i,t} \cdot \bar{x}_{i,N_{i,t}-h_{i,t}} - (N_{i,t} - h_{i,t}) \cdot \bar{x}_{i,N_{i,t}-2h_{i,t}} & \text{if } h_{i,t} = h_{i,t-1} \\ \bar{d}_i + (N_{i,t} - h_{i,t}) \cdot \bar{x}_{i,N_{i,t}-2h_{i,t}+1} + \bar{b}_i & \text{otherwise} \end{cases}. \end{aligned}$$

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

## D 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  $i \in \llbracket K \rrbracket$  and on the round  $t \in \llbracket T \rrbracket$ . When needed, we will particularize the choice for the case in which the window size depends on the number of pulls only  $h_{i,t} = h(N_{i,t-1})$ .

### D.1 Proofs of Section 3

**Lemma D.1.** *Under Assumption 2.1, for every  $i \in \llbracket K \rrbracket$ ,  $j, k \in \mathbb{N}$  with  $k < j$ , it holds that:*

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

475 *Proof.* Using Assumption 2.1, we get:

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

476 where the first inequality comes from the concavity of the expected reward function (Assumption 2.1),  
477 and the second equality comes from the definition of increment.  $\square$

478 **Lemma D.2.** For every arm  $i \in \llbracket K \rrbracket$ , every round  $t \in \llbracket T \rrbracket$ , and window width  $1 \leq h_{i,t} \leq \lfloor N_{i,t-1}/2 \rfloor$ ,  
479 let us define:

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

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

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

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

$$\begin{aligned} \mu_i(T) &= \mu_i(l) + \sum_{j=l}^{T-1} \gamma_i(j) \\ &\leq \mu_i(l) + (T-l) \gamma_i(l-1) \end{aligned} \tag{12}$$

$$\leq \mu_i(l) + (T-l) \frac{\mu_i(l) - \mu_i(l-h_{i,t})}{h_{i,t}}, \tag{13}$$

483 where Equation (12) follows from Assumption 2.1, and Equation (13) is obtained from Lemma D.1.  
484 By averaging over the most recent  $1 \leq h_{i,t} \leq \lfloor N_{i,t-1}/2 \rfloor$  pulls, we get:

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

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

$$\tilde{\mu}_i^T(N_{i,t}) - \mu_i(T) = \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \left( \mu_i(l) + (T-l) \frac{\mu_i(l) - \mu_i(l-h_{i,t})}{h_{i,t}} \right) - \mu_i(T) \tag{14}$$

$$\begin{aligned} &\leq \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} (T-l) \frac{\mu_i(l) - \mu_i(l-h_{i,t})}{h_{i,t}} \\ &= \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} (T-l) \frac{1}{h_{i,t}} \sum_{j=l-h_{i,t}}^{l-1} \gamma_j(l) \\ &\leq \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} (T-l) \gamma_i(l-h_{i,t}) \end{aligned} \tag{15}$$

$$\leq \frac{1}{2} (2T - 2N_{i,t-1} + h_{i,t} - 1) \gamma_i(N_{i,t-1} - 2h_{i,t} + 1), \tag{16}$$

486 where Equation (14) follows from Assumption 2.1 applied as  $\mu_i(l) \leq \mu_i(N_{i,t})$ , Equation (15) follows  
487 from Assumption 2.1 and bounding  $\frac{1}{h_{i,t}} \sum_{j=l-h_{i,t}}^{l-1} \gamma_j(l) \leq \gamma_i(l-h_{i,t})$ , and Equation (16) is derived  
488 still from Assumption 2.1,  $\gamma_i(l-h_{i,t}) \leq \gamma_i(N_{i,t-1} - 2h_{i,t} + 1)$  and computing the summation.  $\square$

489 **Lemma D.3.** For every arm  $i \in \llbracket K \rrbracket$ , every round  $t \in \llbracket T \rrbracket$ , and window width  $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$ ,  
 490 let us define:

$$\begin{aligned}\tilde{\mu}_i^T(N_{i,t}) &:= \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t}-1} \left( X_i(l) + (T-l) \frac{X_i(l) - X_i(l-h_{i,t})}{h_{i,t}} \right), \\ \check{\beta}_i^T(N_{i,t}) &= \sigma(T - N_{i,t-1} + h_{i,t} - 1) \sqrt{\frac{a}{h_{i,t}^3}},\end{aligned}$$

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

$$\mathbb{P}(\forall t \in \llbracket T \rrbracket : |\tilde{\mu}_i^T(N_{i,t}) - \check{\mu}_i^T(N_{i,t})| > \check{\beta}_i^T(N_{i,t})) \leq 2T \exp\left(-\frac{a}{10}\right).$$

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

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

$$\begin{aligned}\mathbb{P}(\forall t \in \llbracket T \rrbracket : |\tilde{\mu}_i^T(N_{i,t}) - \check{\mu}_i^T(N_{i,t})| > \check{\beta}_i^T(N_{i,t})) \\ \leq \mathbb{P}(\exists n \in \llbracket 0, T \rrbracket \text{ s.t. } h_{i,n} \geq 1 : |\tilde{\mu}_i^T(n) - \check{\mu}_i^T(n)| > \check{\beta}_i^T(n)) \\ \leq \sum_{n \in \llbracket 0, T \rrbracket : h_{i,n} \geq 1} \mathbb{P}(|\tilde{\mu}_i^T(n) - \check{\mu}_i^T(n)| > \check{\beta}_i^T(n)),\end{aligned}\tag{17}$$

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

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

$$\begin{aligned}h_{i,n} (\tilde{\mu}_i^T(n) - \check{\mu}_i^T(n)) \\ = \sum_{l=n-h_{i,n}+1}^n \left( \eta_i(l) - (T-l) \cdot \frac{\eta_i(l) - \eta_i(l-h_{i,n})}{h_{i,n}} \right) \\ = \sum_{l=n-h_{i,n}+1}^n \left( 1 - \frac{T-l}{h_{i,n}} \right) \cdot \eta_i(l) - \sum_{l=n-h_{i,n}+1}^n \left( \frac{T-l}{h_{i,n}} \right) \cdot \eta_i(l-h_{i,n}).\end{aligned}$$

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

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

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

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

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

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

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

512

□

513 **Lemma D.4.** For every arm  $i \in \llbracket K \rrbracket$ , every round  $t \in \llbracket T \rrbracket$ , and window width  $1 \leq h_{i,t} \leq \lfloor N_{i,t-1}/2 \rfloor$ ,  
514 let us define:

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

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

$$\mu_i(T) - \bar{\mu}_i(N_{i,t}) \leq \frac{1}{2}(2T - 2N_{i,t-1} + h_{i,t} - 1)\gamma_i(N_{i,t-1} - h_{i,t} + 1).$$

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

$$\mu_i(T) = \mu_i(l) + \sum_{j=l}^{T-1} \gamma_i(j) \geq \mu_i(l). \tag{18}$$

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

$$\begin{aligned}
\mu_i(T) &= \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \left( \mu_i(l) + \sum_{j=l}^{T-1} \gamma_i(j) \right) \\
&= \bar{\mu}_i(N_{i,t}) + \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \sum_{j=l}^{T-1} \gamma_i(j) \\
&\leq \bar{\mu}_i(N_{i,t}) + \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} \sum_{j=l}^{T-1} \gamma_i(j) \\
&\leq \bar{\mu}_i(N_{i,t}) + \frac{1}{2}(2T - 2N_{i,t-1} + h_{i,t} - 1)\gamma_i(N_{i,t-1} - h_{i,t} + 1),
\end{aligned}$$

519 where we used Assumption 2.1.

□

520 **Lemma D.5.** For every arm  $i \in \llbracket K \rrbracket$ , every round  $t \in \llbracket T \rrbracket$ , and window width  $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$ ,  
 521 let us define:

$$\hat{\mu}_i^T(N_{i,t}) := \frac{1}{h_{i,t}} \sum_{l=N_{i,t-1}-h_{i,t}+1}^{N_{i,t-1}} X_i(l),$$

$$\hat{\beta}_i^T(N_{i,t}) = \sigma \sqrt{\frac{a}{h_{i,t}}},$$

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

$$\mathbb{P}\left(\forall t \in \llbracket T \rrbracket : |\hat{\mu}_i(N_{i,t}) - \bar{\mu}_i(N_{i,t})| > \hat{\beta}_i(N_{i,t})\right) \leq 2T \exp\left(-\frac{a}{2}\right).$$

525 *Proof.* Before starting the proof, it is worth noting that under the event  $\{h_{i,t} = 0\}$ , it holds that  
 526  $\hat{\mu}_i^T(t) = \bar{\mu}_i^T(t) = \hat{\beta}_i^T(t) = +\infty$ . Thus, under the convention that  $\infty - \infty = 0$ , then  $0 > \hat{\beta}_i^T(t)$  is not  
 527 satisfied. For this reason, we need to perform our analysis under the event  $\{h_{i,t} \geq 1\}$ .

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

$$\begin{aligned} & \mathbb{P}\left(\forall t \in \llbracket T \rrbracket : |\hat{\mu}_i(N_{i,t}) - \bar{\mu}_i(N_{i,t})| > \hat{\beta}_i(N_{i,t})\right) \\ & \leq \mathbb{P}\left(\exists n \in \llbracket 0, T \rrbracket \text{ s.t. } h_{i,n} \geq 1 : |\hat{\mu}_i(n) - \bar{\mu}_i(n)| > \hat{\beta}_i(n)\right) \\ & \leq \sum_{n \in \llbracket 0, T \rrbracket : h_{i,n} \geq 1} \mathbb{P}\left(|\hat{\mu}_i(n) - \bar{\mu}_i(n)| > \hat{\beta}_i(n)\right), \end{aligned} \quad (19)$$

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

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

$$h_{i,n} (\hat{\mu}_i(n) - \bar{\mu}_i(n)) = \sum_{l=n-h_{i,n}+1}^n \eta_i(l).$$

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

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

$$\begin{aligned} \mathbb{P}\left(|\hat{\mu}_i(n) - \bar{\mu}_i(n)| \geq \hat{\beta}_i(n)\right) & \leq \mathbb{P}\left(\left|\sum_{l=n-h_{i,n}+1}^n \eta_i(l)\right| \geq h_{i,n} \cdot \hat{\beta}_i(n)\right) \\ & \leq 2 \exp\left(-\frac{h_{i,n} \hat{\beta}_i(n)^2}{2\sigma^2}\right) \\ & = 2 \exp\left(-\frac{a}{2}\right). \end{aligned}$$

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

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

541 □

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

$$\hat{\beta}_i(n) - \hat{\zeta}_i(n) \leq \hat{\mu}_i(n) - \mu_i(n) \leq \hat{\beta}_i(n),$$

544 where  $\hat{\beta}_i(n) := \sigma \sqrt{\frac{a}{h(n)}}$  and  $\hat{\zeta}_i(n) := \frac{1}{2}(2T - n + h(n) - 1) \gamma_i(n - h(n) + 1)$ .

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

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

$$\check{\beta}_i^T(n) \leq \check{\mu}_i^T(n) - \mu_i(n) \leq \check{\beta}_i^T(n) + \check{\zeta}_i^T(n),$$

548 where  $\check{\beta}_i^T(n) := \sigma \cdot (T - n + h(n) - 1) \sqrt{\frac{a}{h(n)^3}}$  and  $\check{\zeta}_i^T(n) := \frac{1}{2}(2T - n + h(n) - 1) \gamma_i(n - 2h(n) + 1)$ .

549 *Proof.* The proof simply combines Lemmas D.4 and D.3 and a union bound over the arms. □

## 550 D.2 Proofs of Section 4

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

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

$$\Psi := \left\{ \forall i \in \llbracket K \rrbracket, \forall t \in \llbracket T \rrbracket : |\check{\mu}_i^T(t) - \tilde{\mu}_i^T(t)| < \check{\beta}_i^T(t) \right\},$$

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

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

559 **Lemma D.6.** *Under Assumption 2.1 and assuming that the good event  $\Psi$  holds, the maximum number*  
 560 *of pulls  $N_{i,T}$  of a sub-optimal arm ( $i \neq i^*(T)$ ) performed by the R-UCBE is upper bounded by the*  
 561 *maximum integer  $y_i(a)$  which satisfies the following condition:*

$$T\gamma_i(\lfloor (1 - 2\varepsilon)y_i(a) \rfloor) + 2T\sigma \cdot \sqrt{\frac{a}{[\varepsilon y_i(a)]^3}} \geq \Delta_i(T).$$

562 *Proof.* In the following, we will use  $\tilde{\mu}_i^T(N_{i,t-1})$  to bound the bias introduced by  $\check{\mu}_i^T(N_{i,t-1})$  and,  
 563 subsequently, to find a number of pulls such that the algorithm cannot suggest pulling a suboptimal  
 564 arm. Using Lemma D.4, we have that  $\forall i \in \llbracket K \rrbracket, \forall t \in \llbracket T \rrbracket$  and when  $1 \leq h_{i,t} \leq \lfloor 1/2 \cdot N_{i,t-1} \rfloor$  with  
 565  $N_{i,t-1} \geq 2$ , it holds that:

$$\tilde{\mu}_i^T(N_{i,t-1}) - \mu_i(T) \leq \frac{1}{2} \cdot (2T - 2N_{i,t-1} + h_{i,t} - 1) \cdot \gamma_i(N_{i,t-1} - 2h_{i,t} + 1). \quad (20)$$

Let us assume that, at round  $t$ , the R-UCBE algorithm pulls the arm  $i \in [K]$  such that  $i \neq i^*(T)$ . 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 as  $h_{i,t} = h(N_{i,t-1})$ . 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  $\Psi$  holds. We have that, if we pull such an arm  $i \neq i^*$ , it holds:

$$\begin{aligned} B_i^T(N_{i,t-1}) &\geq B_{i^*}^T(N_{i^*,t-1}) \\ B_i^T(N_{i,t-1}) - B_{i^*}^T(N_{i^*,t-1}) &\geq 0 \\ \Delta_i(T) + B_i^T(N_{i,t-1}) - B_{i^*}^T(N_{i^*,t-1}) &\geq \Delta_i(T). \end{aligned}$$

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

$$\mu_{i^*}(T) - \mu_i(T) + \tilde{\mu}_i^T(N_{i,t-1}) + \check{\beta}_i^T(N_{i,t-1}) - \tilde{\mu}_{i^*}^T(N_{i^*,t-1}) - \check{\beta}_{i^*}^T(N_{i^*,t-1}) \geq \Delta_i(T).$$

Given Assumption 2.1 we have that  $\mu_{i^*}(T) \leq \tilde{\mu}_{i^*}^T(N_{i^*,t-1})$ , and, therefore, we have:

$$\tilde{\mu}_{i^*}^T(N_{i^*,t-1}) - \mu_i(T) + \tilde{\mu}_i^T(N_{i,t-1}) + \check{\beta}_i^T(N_{i,t-1}) - \tilde{\mu}_{i^*}^T(N_{i^*,t-1}) - \check{\beta}_{i^*}^T(N_{i^*,t-1}) \geq \Delta_i(T),$$

and, since under the good event  $\Psi$ , it holds that  $\tilde{\mu}_{i^*}^T(N_{i^*,t-1}) - \tilde{\mu}_i^T(N_{i,t-1}) - \check{\beta}_{i^*}^T(N_{i^*,t-1}) < 0$ , we have:

$$\begin{aligned} -\mu_i(T) + \tilde{\mu}_i^T(N_{i,t-1}) + \check{\beta}_i^T(N_{i,t-1}) &\geq \Delta_i(T) \\ -\mu_i(T) + \check{\beta}_i^T(N_{i,t-1}) + \tilde{\mu}_i^T(N_{i,t-1}) + \underbrace{\tilde{\mu}_i^T(N_{i,t-1}) - \tilde{\mu}_{i^*}^T(N_{i^*,t-1})}_{(D)} &\geq \Delta_i(T), \end{aligned}$$

where we added and subtracted  $\tilde{\mu}_i^T(N_{i,t-1})$  in the last equation. Under the good event  $\Psi$ , we can upper bound  $|(D)| = |\tilde{\mu}_i^T(N_{i,t-1}) - \tilde{\mu}_{i^*}^T(N_{i^*,t-1})| < \check{\beta}_i^T(N_{i,t-1})$ :

$$\tilde{\mu}_i^T(N_{i,t-1}) - \mu_i(T) + 2\check{\beta}_i^T(N_{i,t-1}) \geq \Delta_i(T).$$

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

$$\begin{aligned} \frac{1}{2} \underbrace{(2T - 2N_{i,t-1} + h_{i,t} - 1)}_{\leq 2T} \cdot \gamma_i(N_{i,t-1} - 2h_{i,t} + 1) + \\ + 2\sigma \cdot \underbrace{(T - N_{i,t-1} + h_{i,t} - 1)}_{\leq T} \cdot \sqrt{\frac{a}{h_{i,t}^3}} &\geq \Delta_i(T) \\ \underbrace{T \cdot \gamma_i(\lfloor (1 - 2\varepsilon)N_{i,t} \rfloor)}_{(A)} + \underbrace{2\sigma T \sqrt{\frac{a}{[\varepsilon N_{i,t}]^3}}}_{(B)} &\geq \Delta_i(T), \end{aligned} \quad (21)$$

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

This concludes the proof.  $\square$

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

$$T - \sum_{i \neq i^*(T)} y_i(a) \geq 1, \quad (5)$$



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

$$\underbrace{T\gamma_i(\lfloor (1-2\varepsilon)y \rfloor)}_{(A)} + \underbrace{2T\sigma\sqrt{\frac{a}{[\varepsilon y]^3}}}_{(B)} \geq \Delta_i(T). \quad (6)$$

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

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

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

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

589 Therefore, we need to evaluate the probability that the R-UCBE algorithm would pull a suboptimal  
590 arm in the  $T+1$  round. Given that Assumption 2.1 and that each suboptimal arms have been pulled  
591 a number of times  $N_{i,T}$  at the end of the time budget  $T$ , under the good event  $\Psi$ , we are guaranteed  
592 to recommend the optimal arm if:

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

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

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

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

$$T\gamma_i(\lfloor (1-2\varepsilon)y_i(a) \rfloor) + 2T\sigma\sqrt{\frac{a}{[\varepsilon y_i(a)]^3}} \geq \Delta_i(T).$$

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

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

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

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

$$T \geq \begin{cases} \left( c^{\frac{1}{\beta}} (1-2\varepsilon)^{-1} (H_{1,1/\beta}(T)) + (K-1) \right)^{\frac{\beta}{\beta-1}} & \text{if } \beta \in (1, 3/2) \\ \left( c^{\frac{2}{3}} (1-2\varepsilon)^{-\frac{2}{3}\beta} (H_{1,2/3}(T)) + (K-1) \right)^3 & \text{if } \beta \in [3/2, +\infty) \end{cases}, \quad (8)$$

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

$$a^* = \begin{cases} \frac{\epsilon^3}{4\sigma^2} \left( \left( \frac{T^{1-1/\beta} - (K-1)}{H_{1,1/\beta}(T)} \right)^\beta - c(1-2\varepsilon)^{-\beta} \right)^2 & \text{if } \beta \in (1, 3/2) \\ \frac{\epsilon^3}{4\sigma^2} \left( \left( \frac{T^{1/3} - (K-1)}{H_{1,2/3}(T)} \right)^{3/2} - c(1-2\varepsilon)^{-\beta} \right)^2 & \text{if } \beta \in [3/2, +\infty) \end{cases},$$

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

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

606 *Proof.* We recall that Assumption 2.2 states that all the increment functions  $\gamma_i(n)$  are such that  
 607  $\gamma_i(n) \leq cn^{-\beta}$ . We use such a fact to provide an explicit solution for the optimal value of  $a^*$ . From  
 608 Theorem 4.1 and using the fact that  $\gamma_i(n) \leq cn^{-\beta}$ , we have that Equation (6) becomes:

$$\frac{Tc}{[(1-2\varepsilon)y]^\beta} + \frac{2tT\sigma a^{\frac{1}{2}}}{[\varepsilon y]^{\frac{3}{2}}} \geq \Delta_i(T). \quad (24)$$

609 Or, more restrictively:

$$\frac{Tc(1-2\varepsilon)^{-\beta}}{(y-1)^\beta} + \frac{2T\sigma\varepsilon^{-\frac{3}{2}}a^{\frac{1}{2}}}{(y-1)^{\frac{3}{2}}} \geq \Delta_i(T).$$

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

**Case 1:**  $\beta \in [\frac{3}{2}, \infty)$  In this branch, we can upper bound the left-side part of the inequality in Equation (24) by:

$$\frac{Tc(1-2\varepsilon)^{-\beta}}{(y-1)^{\frac{3}{2}}} + \frac{2T\sigma\varepsilon^{-\frac{3}{2}}a^{\frac{1}{2}}}{(y-1)^{\frac{3}{2}}} \geq \Delta_i(T).$$

612 Thus, we can derive:

$$y_i(a) \leq 1 + \left( \frac{Tc(1-2\varepsilon)^{-\beta} + 2T\sigma\varepsilon^{-\frac{3}{2}}a^{\frac{1}{2}}}{\Delta_i(T)} \right)^{\frac{2}{3}}. \quad (25)$$

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

$$\begin{aligned} T - \sum_{i \neq i^*(T)} y_i(a) &> 0 \\ T - (K-1) - \left( Tc(1-2\varepsilon)^{-\beta} + 2T\sigma\varepsilon^{-\frac{3}{2}}a^{\frac{1}{2}} \right)^{\frac{2}{3}} \sum_{i \neq i^*(T)} \frac{1}{\Delta_i^{\frac{2}{3}}(T)} &> 0 \\ T - (K-1) - \left( Tc(1-2\varepsilon)^{-\beta} + 2T\sigma\varepsilon^{-\frac{3}{2}}a^{\frac{1}{2}} \right)^{\frac{2}{3}} H_{1,2/3}(T) &> 0 \\ a &< \frac{\left( \frac{(T^{1/3} - T^{-2/3}(K-1))^{\frac{3}{2}}}{(H_{1,2/3}(T))^{\frac{3}{2}}} - c(1-2\varepsilon)^{-\beta} \right)^2}{4\sigma^2\varepsilon^{-3}} \\ a &< \frac{\left( \frac{(T^{1/3} - (K-1))^{\frac{3}{2}}}{(H_{1,2/3}(T))^{\frac{3}{2}}} - c(1-2\varepsilon)^{-\beta} \right)^2}{4\sigma^2\varepsilon^{-3}}, \end{aligned}$$

614 where the last expression is obtained by observing that  $T \geq 1$  and for obtaining a more manageable  
 615 expression, under the assumption that  $\frac{(T^{1/3} - (K-1))^{\frac{3}{2}}}{(H_{1,2/3}(T))^{\frac{3}{2}}} - c(1-2\varepsilon)^{-\beta} \geq 0$ .

616 This implies a constraint on the minimum time budget  $T$ , which explicit form for the case  $\beta \in [\frac{3}{2}, \infty)$   
 617 is provided in Lemma D.7

618 **Case 2:**  $\beta \in (1, \frac{3}{2})$  In this case, we enforce the more restrictive condition:

$$\frac{Tc(1-2\varepsilon)^{-\beta}}{(y-1)^\beta} + \frac{2T\sigma\varepsilon^{-\frac{3}{2}}a^{\frac{1}{2}}}{(y-1)^\beta} \geq \Delta_i(T),$$

619 the value for the number of pulls is:

$$y_i(a) \leq 1 + \left( \frac{Tc(1-2\varepsilon)^{-\beta} + 2T\sigma\varepsilon^{-\frac{3}{2}}a^{\frac{1}{2}}}{\Delta_i(T)} \right)^{\frac{1}{\beta}}. \quad (26)$$

620 and the value for  $a^*$  becomes:

$$\begin{aligned} T - \sum_{i \neq i^*(T)} N_{i,T} &> 0 \\ T - (K-1) - \left( Tc(1-2\varepsilon)^{-\beta} + 2T\sigma\varepsilon^{-\frac{3}{2}}a^{\frac{1}{2}} \right)^{\frac{1}{\beta}} \sum_{i \neq i^*(T)} \frac{1}{\Delta_i^{\frac{1}{\beta}}(T)} &> 0 \\ T - (K-1) - \left( Tc(1-2\varepsilon)^{-\beta} + 2T\sigma\varepsilon^{-\frac{3}{2}}a^{\frac{1}{2}} \right)^{\frac{1}{\beta}} H_{1,1/\beta}(T) &> 0 \\ a &< \frac{\left( \frac{(T^{1-1/\beta} - T^{-1/\beta}(K-1))^\beta}{(H_{1,1/\beta}(T))^\beta} - c(1-2\varepsilon)^{-\beta} \right)^2}{4\sigma^2\varepsilon^{-3}} \\ a &< \frac{\left( \frac{(T^{1-1/\beta} - (K-1))^\beta}{(H_{1,1/\beta}(T))^\beta} - c(1-2\varepsilon)^{-\beta} \right)^2}{4\sigma^2\varepsilon^{-3}}, \end{aligned}$$

621 where the last expression is obtained by observing that  $T \geq 1$  and for obtaining a more convenient  
622 expression, under the assumption that  $\frac{(T^{1-1/\beta} - (K-1))^\beta}{(H_{1,1/\beta}(T))^\beta} - c(1-2\varepsilon)^{-\beta} \geq 0$ .

623 Also here, this implies a constraint on the minimum time budget  $T$  for the case  $\beta \in (1, \frac{3}{2})$ , which  
624 explicit form is provided in Lemma D.7  $\square$

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

$$T \geq \begin{cases} \left( c^{\frac{1}{\beta}}(1-2\varepsilon)^{-1} (H_{1,1/\beta}(T)) + (K-1) \right)^{\frac{\beta}{\beta-1}} & \text{if } \beta \in (1, 3/2) \\ \left( c^{\frac{2}{3}}(1-2\varepsilon)^{-\frac{2}{3}\beta} (H_{1,2/3}(T)) + (K-1) \right)^3 & \text{if } \beta \in [3/2, +\infty) \end{cases}$$

627 and  $H_{1,\eta}(T) := \sum_{i \neq i^*(T)} \frac{1}{\Delta_i^\eta(T)}$  for  $\eta \leq 1$ .

628 *Proof.* Given Corollary 4.2, we want to find the values of  $T$  such that a value of  $a \in [0, a^*]$  should  
629 exist. This implies having  $a^* \geq 0$ . Given the value of  $\beta$ , we can derive a lower bound for the time  
630 budget  $T$ .

**Case 1:**  $\beta \in [\frac{3}{2}, \infty)$  :

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

From this, it follows:

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

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

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

From this, it follows:

$$T \geq \left( c^{\frac{1}{\beta}}(1-2\varepsilon)^{-1} (H_{1,1/\beta}(T)) + (K-1) \right)^{\frac{\beta}{\beta-1}}.$$

631

$\square$

632 **D.3 Proofs of Section 5**

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

**Lemma D.8.** *For every arm  $i \in \llbracket K \rrbracket$  and every round  $t \in \llbracket T \rrbracket$ , let us define:*

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

637 if  $N_{i,t} \geq 2$ , then  $\mu_i(t) \geq \bar{\mu}_i(t)$ , and if  $h_{i,t} \leq N_{i,t}/2$ , it holds that:

$$\mu_i(T) - \bar{\mu}_i(N_{i,t}) \leq T \gamma_i(\lfloor N_{i,t}/2 \rfloor). \quad (27)$$

638 *Proof.* The proof follows trivially from Lemma D.2.  $\square$

**Lemma D.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  $\forall j \in \llbracket K-1 \rrbracket$  and  $\forall i \in \llbracket K \rrbracket$ :*

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

639 is:

$$T \geq c^{\frac{1}{1-\beta}} 2^{\frac{1+\beta}{\beta-1}} \overline{\log}(K)^{\frac{\beta}{\beta-1}} \max_{i \in \llbracket 2, K \rrbracket} \left\{ i^{\frac{\beta}{\beta-1}} \Delta_{(i)}^{-\frac{1}{\beta-1}}(T) \right\}.$$

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

$$T \gamma_i(\lfloor N_j/2 \rfloor) \leq cT [N_j/2]^{-\beta}.$$

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

$$T [N_j/2]^{-\beta} \leq T \cdot \left( \frac{1}{\overline{\log}(K)} \cdot \frac{T-K}{K+1-j} - 1 \right)^{-\beta} \quad (28)$$

$$\begin{aligned} &\leq T \cdot \left( \frac{1}{\overline{\log}(K)} \cdot \frac{T}{K+1-j} - 1 \right)^{-\beta} \\ &\leq T \cdot \left( \frac{T}{\overline{\log}(K) \cdot (K+1-j)} \right)^{-\beta}. \end{aligned} \quad (29)$$

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

$$\begin{aligned} cT [N_j/2]^{-\beta} &\leq \frac{\Delta_{(K+1-j)}(T)}{2} \\ cT \left( \frac{T}{2\overline{\log}(K) \cdot (K+1-j)} \right)^{-\beta} &\leq \frac{\Delta_{(K+1-j)}(T)}{2} \\ T^{1-\beta} &\leq \frac{\Delta_{(K+1-j)}(T)}{c2^{1+\beta} \cdot (\overline{\log}(K) \cdot (K+1-j))^{\beta}} \\ T &\geq \frac{\Delta_{(K+1-j)}^{\frac{1}{1-\beta}}(T)}{c^{\frac{1}{1-\beta}} 2^{\frac{1+\beta}{1-\beta}} \cdot (\overline{\log}(K) \cdot (K+1-j))^{\frac{\beta}{1-\beta}}}. \end{aligned}$$

646 Requiring that the above condition holds for all the phases  $j \in \llbracket K-1 \rrbracket$  we have:

$$\begin{aligned}
T &\geq \max_{j \in \llbracket K-1 \rrbracket} \left\{ \frac{\Delta_{(K+1-j)}^{\frac{1}{1-\beta}}(T)}{c^{\frac{1}{1-\beta}} 2^{\frac{1+\beta}{1-\beta}} \cdot (\overline{\log}(K) \cdot (K+1-j))^{\frac{\beta}{1-\beta}}} \right\} \\
&\geq c^{\frac{1}{1-\beta}} 2^{-\frac{1+\beta}{1-\beta}} \overline{\log}(K)^{-\frac{\beta}{1-\beta}} \max_{j \in \llbracket K-1 \rrbracket} \left\{ \left( \frac{\Delta_{(K+1-j)}(T)}{(K+1-j)^\beta} \right)^{\frac{1}{1-\beta}} \right\} \\
&\geq c^{\frac{1}{1-\beta}} 2^{-\frac{1+\beta}{1-\beta}} \overline{\log}(K)^{-\frac{\beta}{1-\beta}} \cdot \max_{j \in \llbracket K-1 \rrbracket} \left\{ \left( (K+1-j)^\beta \Delta_{(K+1-j)}^{-1}(T) \right)^{\frac{1}{\beta-1}} \right\} \\
&\geq c^{\frac{1}{1-\beta}} 2^{-\frac{1+\beta}{1-\beta}} \overline{\log}(K)^{-\frac{\beta}{1-\beta}} \max_{i \in \llbracket 2, K \rrbracket} \left\{ i^{\frac{\beta}{\beta-1}} \Delta_{(i)}^{-\frac{1}{\beta-1}}(T) \right\}.
\end{aligned}$$

647

□

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

$$T \geq 2^{\frac{\beta+1}{\beta-1}} c^{\frac{1}{\beta-1}} \overline{\log}(K)^{\frac{\beta}{\beta-1}} \max_{i \in \llbracket 2, K \rrbracket} \left\{ i^{\frac{\beta}{\beta-1}} \Delta_{(i)}(T)^{-\frac{1}{\beta-1}} \right\}, \quad (10)$$

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

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

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

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

$$\begin{aligned}
e_T(\text{R-SR}) &\leq \mathbb{P}(\exists j \in \llbracket K-1 \rrbracket \exists i \in \llbracket K+1-j, K \rrbracket : \hat{\mu}_{(1)}(N_j) < \hat{\mu}_{(i)}(N_j)) \\
&\leq \sum_{j=1}^{K-1} \mathbb{P}(\exists i \in \llbracket K+1-j, K \rrbracket : \hat{\mu}_{(1)}(N_j) < \hat{\mu}_{(i)}(N_j)) \\
&\leq \sum_{j=1}^{K-1} \sum_{i=K+1-j}^K \mathbb{P}(\hat{\mu}_{(1)}(N_j) \leq \hat{\mu}_{(i)}(N_j)),
\end{aligned}$$

652 where we use a union bound over the phases and over the arms still in the available arm set  $\mathcal{X}_{j-1}$  in  
653 each phase. Let us focus on  $\mathbb{P}(\hat{\mu}_{(1)}(N_j) \leq \hat{\mu}_{(i)}(N_j))$ . We have that the optimal arm has a smaller  
654 pessimistic estimator than the  $i^{\text{th}}$  one when:

$$\hat{\mu}_{(i)}(N_j) \geq \hat{\mu}_{(1)}(N_j)$$

$$\hat{\mu}_{(i)}(N_j) - \hat{\mu}_{(1)}(N_j) \geq 0$$

$$\mu_{(1)}(T) - \hat{\mu}_{(1)}(N_j) + \hat{\mu}_{(i)}(N_j) - \mu_{(i)}(T) \geq \Delta_{(i)}(T) \quad (30)$$

$$\underbrace{\mu_{(1)}(T) - \bar{\mu}_{(1)}(N_j)}_{\leq T \cdot \gamma_{(1)}(N_j-1)} - \hat{\mu}_{(1)}(N_j) + \bar{\mu}_{(1)}(N_j) + \hat{\mu}_{(i)}(N_j) - \underbrace{\mu_{(i)}(T)}_{\leq -\bar{\mu}_{(i)}(N_j)} \geq \Delta_{(i)}(T) \quad (31)$$

$$-\hat{\mu}_{(1)}(N_j) + \bar{\mu}_{(1)}(N_j) + \hat{\mu}_{(i)}(N_j) - \bar{\mu}_{(i)}(N_j) \geq \Delta_{(i)}(T) - T \cdot \gamma_{(1)}(N_j-1) \quad (32)$$

655 where we added  $\pm \Delta_{(i)}(T)$  to derive Equation (30), and added  $\pm \bar{\mu}_{(1)}(N_j)$  to derive Equation (31),  
656 we used the results in Lemma D.8 and from the fact that the reward function is increasing. Since we  
657 are with a time budget  $T$  satisfying Theorem D.9, we have that:

$$T \gamma_{(1)}(N_j-1) \leq \frac{\Delta_{(i)}(T)}{2}. \quad (33)$$

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

$$-\hat{\mu}_{(1)}(N_j) + \bar{\mu}_{(1)}(N_j) + \hat{\mu}_{(i)}(N_j) - \bar{\mu}_{(i)}(N_j) \geq \frac{\Delta_{(i)}(T)}{2},$$

and the error probability becomes:

$$e_T(\mathbf{R-SR}) \leq \sum_{j=1}^{K-1} \sum_{i=K+1-j}^K \mathbb{P} \left( -\hat{\mu}_{(1)}(N_j) + \bar{\mu}_{(1)}(N_j) + \hat{\mu}_{(i)}(N_j) - \bar{\mu}_{(i)}(N_j) \geq \frac{\Delta_{(i)}(T)}{2} \right).$$

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

$$\begin{aligned} e_T(\mathbf{R-SR}) &\leq \sum_{j=1}^{K-1} \sum_{i=K+1-j}^K \exp \left( -\frac{\varepsilon N_j \left( \frac{\Delta_{(i)}(T)}{2} \right)^2}{2\sigma^2} \right) \\ &\leq \sum_{j=1}^{K-1} j \exp \left( -\frac{\varepsilon N_j \Delta_{(K+1-j)}^2}{8\sigma^2} \right). \end{aligned}$$

659 Now, given that:

$$\begin{aligned} \frac{\varepsilon N_j}{8\sigma^2} \Delta_{(K+1-j)}^2 &\geq \frac{\varepsilon}{8\sigma^2} \frac{T-K}{\log(K)(K+1-j) \Delta_{(K+1-j)}^{-2}} \\ &\geq \frac{\varepsilon}{8\sigma^2} \frac{T-K}{\log(K)H_2(T)}, \end{aligned}$$

660 we finally derive the following:

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

661 which concludes the proof.  $\square$

#### 662 D.4 Proofs of Section 6

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

665 **Theorem 6.1.** *For every algorithm  $\mathfrak{A}$ , there exists a deterministic SRB satisfying Assumptions 2.1  
666 and 2.2 such that the optimal arm  $i^*(T)$  cannot be identified for some time budgets  $T$  unless:*

$$T \geq H_{1,1/(\beta-1)}(T) = \sum_{i \neq i^*(T)} \frac{1}{\Delta_i(T)^{\frac{1}{\beta-1}}}. \quad (11)$$

667 *Proof.* We define for every suboptimal arm  $i \in \llbracket 2, K \rrbracket$  the suboptimality gap reached at  $T \rightarrow +\infty$   
668 as  $\Delta_i \in (0, 1/2]$ . We consider the base instance  $\nu$  (see Figure 5) in which define the (deterministic)  
669 reward functions are defined for  $\beta > 1$  and  $n \in \mathbb{N}$  as:

$$\begin{aligned} \mu_1(n) &= \frac{1}{2} (1 - n^{1-\beta}), \\ \mu_i(n) &= \min \left\{ \underbrace{\left( \frac{1}{2} + \Delta_i \right) (1 - n^{1-\beta})}_{=:\mu'_i(n)}, \frac{1}{2} - \Delta_i \right\} \quad i \in \llbracket 2, K \rrbracket. \end{aligned}$$

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

$$\begin{aligned}\gamma_1(n) &= \mu_1(n+1) - \mu_1(n) \leq \sup_{x \in [n, n+1]} \frac{\partial}{\partial x} \mu_1(x) \\ &= \frac{\beta-1}{2} \sup_{x \in [n, n+1]} x^{-\beta} = \frac{\beta-1}{2} n^{-\beta},\end{aligned}\tag{34}$$

672

$$\begin{aligned}\gamma_i(n) &= \mu_i(n+1) - \mu_i(n) \leq \sup_{x \in [n, n+1]} \frac{\partial}{\partial x} \mu'_i(x) \\ &= (\beta-1) \left( \frac{1}{2} + \Delta_i \right) \sup_{x \in [n, n+1]} x^{-\beta} = (\beta-1) n^{-\beta}\end{aligned}$$

Thus, we simply take  $c = \beta - 1$  in Assumption 2.1. Let us define  $n_i^*$  the number of pulls in which arm  $i \in \llbracket 2, K \rrbracket$  reaches the stationary behavior:

$$\left( \frac{1}{2} + \Delta_i \right) (1 - n^{1-\beta}) = \frac{1}{2} - \Delta_i \implies n_i^* = \left( \frac{1/2 + \Delta_i}{2\Delta_i} \right)^{\frac{1}{\beta-1}}.$$

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

$$\frac{1}{2} (1 - T^{1-\beta}) = \frac{1}{2} - \Delta_i \implies T^* := \max_{i \in \llbracket 2, K \rrbracket} \left( \frac{1}{2\Delta_i} \right)^{\frac{1}{\beta-1}}.$$

Consider now the regime in which  $T \geq T^*$ . We proceed by contradiction. Suppose that there exists an algorithm  $\mathfrak{A}$  that identifies the optimal arm such that on the bandit  $\nu$  and that the suboptimal arm  $\bar{i} \in \llbracket 2, K \rrbracket$  has an expected number of pulls satisfying:

$$\mathbb{E}_{\mu}[N_{\bar{i}}(T)] < n_{\bar{i}}^*.\tag{35}$$

Consider now the alternative bandit  $\nu'$  constructed from  $\nu$  by keeping all the arms unaltered, except for arm  $\bar{i}$  that is made optimal:

$$\begin{aligned}\mu'_{\bar{i}}(n) &= \left( \frac{1}{2} + \Delta_{\bar{i}} \right) (1 - n^{1-\beta}), \\ \mu'_j(n) &= \mu_j(n), \quad j \in \llbracket K \rrbracket \setminus \{\bar{i}\}.\end{aligned}$$

Clearly the bandit  $\nu'$  fulfills Assumption 2.1 and, with calculations similar to those in Equation (34), we conclude that it satisfies Assumption 2.2 with  $c = \beta - 1$ . A sufficient condition on  $T$  for which arm  $\bar{i}$  is optimal in bandit  $\nu'$  is that  $T \geq T_2$  in which the curve of arm  $\bar{i}$  intersects that of the arms  $j$  such that  $\Delta_j \geq \Delta_{\bar{i}}$ :

$$\left( \frac{1}{2} + \Delta_{\bar{i}} \right) (1 - T^{1-\beta}) = \frac{1}{2} - \Delta_j \implies T \geq \max_{j \in \llbracket K \rrbracket : \Delta_j \geq \Delta_{\bar{i}}} \left( \frac{1/2 + \Delta_{\bar{i}}}{\Delta_{\bar{i}} + \Delta_j} \right)^{\frac{1}{\beta-1}}.$$

Thus, we take:

$$T_2 := \left( \frac{1/2 + \Delta_{\bar{i}}}{2\Delta_{\bar{i}}} \right)^{\frac{1}{\beta-1}}.$$

Clearly, for  $T^* \geq T_2$  since all the suboptimality gaps are at most  $1/2$ . Thus, we continue in the regime  $T \geq T^*$ . Since  $\mu'_{\bar{i}}(n) = \mu_{\bar{i}}(n)$  if  $n < n_{\bar{i}}^*$ , it follows that under condition (35), algorithm  $\mathfrak{A}$  cannot distinguish between the two bandits and, consequently, cannot identify the optimal arm on

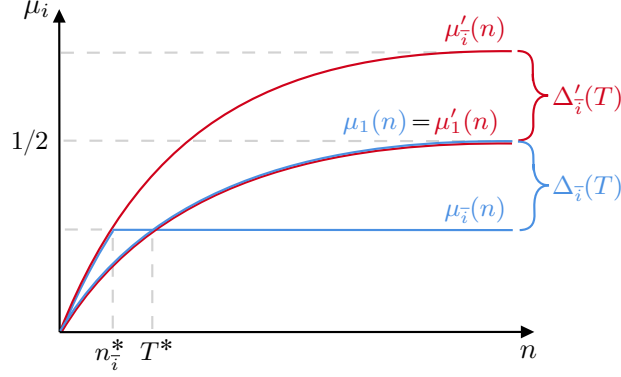


Figure 5: Instances  $\nu$  and  $\nu'$  of SRB used in Theorem 6.1 and Theorem 6.2.

bandit  $\nu'$ . Thus, it must follow, from the contradiction, that:

$$\mathbb{E}_{\nu}[N_{\bar{i}}(T)] \geq n_{\bar{i}}^*.$$

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

$$T \geq \sum_{\bar{i} \in [2, K]} \mathbb{E}_{\nu}[N_{\bar{i}}(T)] \geq \sum_{\bar{i} \in [2, K]} n_{\bar{i}}^* = \sum_{\bar{i} \in [2, K]} \left( \frac{1/2 + \Delta_{\bar{i}}}{2\Delta_{\bar{i}}} \right)^{\frac{1}{\beta-1}} \geq \sum_{\bar{i} \in [2, K]} \left( \frac{1}{4\Delta_{\bar{i}}} \right)^{\frac{1}{\beta-1}} =: T^{\dagger}.$$

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

To conclude, we need to relate  $\Delta_i$  with  $\Delta_i(T)$  and  $\Delta'_i(T)$ . We perform the computation for both the instances  $\nu$  and  $\nu'$ , in the regime  $T \geq 2^{\frac{1}{\beta-1}} T^*$ . Let us start with  $\nu$ :

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

We move to  $\nu'$ :

$$\begin{aligned} \Delta'_1(T) &= \left( \frac{1}{2} + \Delta_{\bar{i}} \right) (1 - T^{1-\beta}) - \frac{1}{2}(1 - T^{1-\beta}) = \Delta_{\bar{i}}(1 - T^{1-\beta}) \geq \frac{\Delta_{\bar{i}}}{2}, \\ \Delta'_i(T) &= \left( \frac{1}{2} + \Delta_{\bar{i}} \right) (1 - T^{1-\beta}) - \left( \frac{1}{2} - \Delta_i \right) \geq \Delta_i - \frac{1}{2}T^{1-\beta} \geq \frac{\Delta_i}{2}, \quad i \in [2, K] \setminus \{\bar{i}\}. \end{aligned}$$

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

$$T \geq \sum_{\bar{i} \in [2, K]} \left( \frac{1/2 + 2\Delta_{\bar{i}}(T)}{4\Delta_{\bar{i}}(T)} \right)^{\frac{1}{\beta-1}} \geq \sum_{\bar{i} \in [2, K]} \left( \frac{1}{8\Delta_{\bar{i}}(T)} \right)^{\frac{1}{\beta-1}} = 2^{-\frac{1}{\beta-1}} T^{\dagger}.$$

Similarly, with  $K \geq 8^{\frac{1}{\beta-1}}$ , we can enforce  $2^{-\frac{1}{\beta-1}} T^{\dagger} \geq 2^{\frac{1}{\beta-1}} T^*$ .  $\square$

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

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

*Proof.* The proof imports the technique from (Kaufmann et al., 2016, Theorem 16 and 17). We consider the Gaussian bandit  $\mu$  with variance  $\sigma^2$  equal for all the arms and the expected reward  $\mu_i(n)$  as in the base instance of proof of Theorem 6.1 (see Figure 5). Let us define by convention  $\Delta_1 = \Delta_2$ .



705 Let  $\mathfrak{A}$  be an algorithm, it is simple to show that there exists an arm  $\bar{i} \in \llbracket K \rrbracket$ , such that:

$$\mathbb{E}_{\mu}[N_{\bar{i}}(T)] \leq \frac{T}{H_2^+ \Delta_{\bar{i}}^2},$$

706 where  $H_2^+ = \sum_{i=1}^K \Delta_i^{-2}$ . We consider two cases. Suppose that  $\bar{i} = 1$  and we construct the alternative  
 707 Gaussian bandit  $\nu'$  with the same variance  $\sigma^2$  and the expected rewards defined as follows:

$$\begin{aligned} \mu'_1(T) &= \min \left\{ \frac{1}{2} (1 - n^{1-\beta}), \frac{1}{2} - 2\Delta_1 \right\}, \\ \mu'_i(T) &= \mu_i(T), \quad i \in \llbracket 2, K \rrbracket. \end{aligned}$$

708 For  $T$  sufficiently large as in Theorem 6.1, while in bandit  $\nu$  the optimal arm is 1, in bandit  $\nu'$  the  
 709 optimal arm is 2. Instead, suppose that  $\bar{i} \neq 1$  and we construct the alternative Gaussian bandit  $\nu'$   
 710 with the same variance  $\sigma^2$  and the expected rewards defined as follows:

$$\begin{aligned} \mu'_{\bar{i}}(T) &= \left( \frac{1}{2} + \Delta_{\bar{i}} \right) (1 - n^{1-\beta}), \\ \mu'_i(T) &= \mu_i(T), \quad i \in \llbracket K \rrbracket \setminus \{\bar{i}\}. \end{aligned}$$

711 For  $T$  sufficiently large as in Theorem 6.1, while in bandit  $\nu$  the optimal arm is 1, in bandit  $\nu'$  the  
 712 optimal arm is  $\bar{i}$ . Let us denote with  $\nu_i(t)$  the distribution of the reward at time  $t$  for arm  $i$ . By the  
 713 Bretagnolle-Huber's inequality, we obtain:

$$\begin{aligned} \max\{e_T(\nu), e_T(\nu')\} &\geq \frac{1}{4} \exp \left( -\mathbb{E}_{\nu} \left[ \sum_{t=1}^T \mathbb{1}\{I_t = \bar{i}\} D_{KL}(\nu_i(t), \nu'_i(t)) \right] \right) \\ &= \frac{1}{4} \exp \left( -\mathbb{E}_{\nu} \left[ \sum_{t=1}^T \mathbb{1}\{I_t = \bar{i}\} \frac{(\mu_{\bar{i}}(N_{\bar{i},t}) - \mu'_{\bar{i}}(N_{\bar{i},t}))^2}{2\sigma^2} \right] \right) \\ &\geq \frac{1}{4} \exp \left( -\mathbb{E}_{\nu} [N_{\bar{i}}(T)] \frac{(2\Delta_{\bar{i}})^2}{2\sigma^2} \right) = \frac{1}{4} \exp \left( -\frac{2T}{\sigma^2 H_2^+} \right) \\ &\geq \frac{1}{4} \exp \left( -\frac{2T}{\sigma^2 H_2^+} \right), \end{aligned}$$

714 where we observed that for every  $n \in \llbracket T \rrbracket$ , we have  $|\mu_{\bar{i}}(n) - \mu'_{\bar{i}}(n)| \leq 2\Delta_{\bar{i}}$ . To conclude, we relate  
 715  $H_2^+$  with  $H_{1,2}(T)$ . Using an argument analogous to that of the last part of the proof Theorem 6.1 it  
 716 is simple to observe that, for sufficiently large  $T$ , we have  $\Delta_{\bar{i}} \leq 2\Delta_i(T)$ , from which we have:

$$H_2^+ = \sum_{i=1}^K \Delta_i^{-2} = \Delta_1^{-2} + \sum_{i=2}^K \Delta_i^{-2} \geq \sum_{i=2}^K \Delta_i^{-2} \geq \frac{1}{4} \sum_{i=2}^K \Delta_i(T)^{-2} = \frac{1}{4} H_{1,2}(T).$$

717 □

## 718 D.5 Auxiliary Lemmas

719 **Lemma D.10** (Hoeffding-Azuma's inequality for weighted martingales). *Let  $\mathcal{F}_1 \subset \dots \subset \mathcal{F}_n$  be a*  
 720 *filtration and  $X_1, \dots, X_n$  be real random variables such that  $X_t$  is  $\mathcal{F}_t$ -measurable,  $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = 0$*   
 721 *(i.e., a martingale difference sequence), and  $\mathbb{E}[\exp(\lambda X_t) | \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$  for any  $\lambda > 0$  (i.e.,*  
 722  *$\sigma^2$ -subgaussian). Let  $\alpha_1, \dots, \alpha_n$  be non-negative real numbers. Then, for every  $\kappa \geq 0$  it holds that:*

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

723 *Proof.* For a complete demonstration of this statement, we refer to Lemma C.5 of Metelli et al.  
 724 (2022).  $\square$

725 **Lemma D.12.** *Let  $\beta > 1$ , then it holds that:*

$$H_{1,1/\beta}(T)^{\beta/(\beta-1)} \geq H_{1,1/(\beta-1)}(T).$$

726 *Proof.* We prove the equivalent statement, being  $\beta > 1$ :

$$H_{1,1/(\beta-1)}(T)^{(\beta-1)/\beta} \leq H_{1,1/\beta}(T).$$

727 Recalling that the function  $(\cdot)^{(\beta-1)/\beta}$  is subadditive, being  $(\beta-1)/\beta < 1$ , we have:

$$H_{1,1/(\beta-1)}(T)^{(\beta-1)/\beta} = \left( \sum_{i \neq i^*(T)} \frac{1}{\Delta_i(T)^{1/(\beta-1)}} \right)^{(\beta-1)/\beta} \leq \sum_{i \neq i^*(T)} \frac{1}{\Delta_i(T)^{1/\beta}} = H_{1,1/\beta}(T).$$

728  $\square$

## 729 E Theoretical Analysis of a baseline: RR-SW

730 In this appendix, we provide the theoretical analysis for the algorithm Round Robin Sliding  
 731 Window (RR-SW), as it represents the most intuitive baseline for this setting. First, we need to  
 732 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  $\varepsilon$

```

1 Initialize  $t \leftarrow 1$ 
2 Estimate  $N \leftarrow \frac{T}{K}$ 
3 for  $i \in \llbracket K \rrbracket$  do
4   for  $l \in \llbracket N \rrbracket$  do
5     Pull arm  $i$  and observe  $x_t$ 
6      $t \leftarrow t + 1$ 
7   end
8   Update  $\hat{\mu}_i(N)$ 
9 end
10 Recommend  $\hat{I}^*(T) \in \arg \max_{i \in \llbracket K \rrbracket} \hat{\mu}_i(N)$ 

```

---

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

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

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

$$T \geq 2^{\frac{1}{\beta-1}} c^{\frac{1}{\beta-1}} (1 - \varepsilon)^{-\frac{\beta}{\beta-1}} K^{\frac{\beta}{\beta-1}} \Delta_{(2)}^{-\frac{1}{\beta-1}}(T), \quad (36)$$

*the error probability of RR-SW is bounded by:*

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

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  $\hat{\mu}_i(N)$ . Second, the error probability scales as an (inverse) function only of the smallest suboptimality gap  $\Delta_{(2)}(T)$ .

## E.1 Proofs

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

**Lemma E.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  $\forall i \in \llbracket K \rrbracket$ :*

$$T \gamma_i((1 - \varepsilon)(N - 1)) \leq \frac{\Delta_{(2)}(T)}{2},$$

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

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

*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  $\mu_i(T)$ , 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:

$$T \gamma_i((1 - \varepsilon)(N - 1)) \leq \frac{\Delta_{(2)}(T)}{2}. \quad (37)$$

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

$$\begin{aligned} T \gamma_i((1 - \varepsilon)(N - 1)) &\leq T \gamma_i((1 - \varepsilon)N) \\ &= T \gamma_i\left((1 - \varepsilon)\frac{T}{K}\right) \\ &\leq T c \left((1 - \varepsilon)\frac{T}{K}\right)^{-\beta}. \end{aligned} \quad (38)$$

By introducing the term derived in Equation (38) into Equation (37) we obtain:

$$T c \left((1 - \varepsilon)\frac{T}{K}\right)^{-\beta} \leq \frac{\Delta_{(2)}(T)}{2}.$$

This implies that the minimum time budget  $T$  which guarantees the initial condition of Equation (38) is:

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

where  $\Delta_{(2)}(T)$  is the minimum suboptimality gap ( $\Delta_{(2)}(T) = \min_{i \neq i^*(T)} \{\Delta_i(T)\}$ ).  $\square$

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

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

$$T \geq 2^{\frac{1}{\beta-1}} c^{\frac{1}{\beta-1}} (1 - \varepsilon)^{-\frac{\beta}{\beta-1}} K^{\frac{\beta}{\beta-1}} \Delta_{(2)}^{-\frac{1}{\beta-1}}(T), \quad (36)$$

the error probability of RR-SW is bounded by:

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

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

$$\begin{aligned} e_T(\text{RR-SW}) &= \mathbb{P}(\exists i \in \llbracket K \rrbracket : \hat{\mu}_1(N) < \hat{\mu}_i(N)) \\ &\leq \sum_{i \in \llbracket K \rrbracket} \mathbb{P}(\hat{\mu}_1(N) < \hat{\mu}_i(N)). \end{aligned}$$

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

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

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

$$\mu_1(T) - \hat{\mu}_1(N) + \hat{\mu}_i(N) - \mu_i(T) \geq \Delta_i(T) \quad (39)$$

$$\underbrace{\mu_1(T) - \bar{\mu}_1(N)}_{\leq T \cdot \gamma_1(N-1)} - \hat{\mu}_1(N) + \bar{\mu}_1(N) + \hat{\mu}_i(N) - \underbrace{\mu_i(T)}_{\leq -\bar{\mu}_i(N)} \geq \Delta_i(T) \quad (40)$$

$$-\hat{\mu}_1(N) + \bar{\mu}_1(N) + \hat{\mu}_i(N) - \bar{\mu}_i(N) \geq \Delta_i(T) - T \cdot \gamma_1(N-1) \quad (41)$$

774 where we added  $\pm \Delta_i(T)$  to derive Equation (39), and added  $\pm \bar{\mu}_1(N)$  to derive Equation (40), we  
 775 used the results in Lemma D.8 and from the fact that the reward function is increasing. Considering a  
 776 time budget  $T$  satisfying Theorem E.2, and  $\Delta_i(T) \geq \Delta_{(2)}(T)$ ,  $\forall i \in \llbracket K \rrbracket$ , we have that:

$$T \gamma_1(N-1) \leq \frac{\Delta_i(T)}{2}. \quad (42)$$

777 Equation (42) holds since we are considering a time budget  $T$  which satisfies a more restrictive  
 778 condition (we are considering a time budget at which this separation already holds for  $(1 - \varepsilon)N$ , so  
 779 it also holds now).

Substituting Equation (42) into Equation (41) the above, we have:

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

and the error probability becomes:

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

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

$$\begin{aligned} e_T(\text{RR-SW}) &\leq \sum_{i=1}^K \exp\left(-\frac{\varepsilon N \left(\frac{\Delta_i(T)}{2}\right)^2}{2\sigma^2}\right) \\ &\leq K \exp\left(-\frac{\varepsilon T}{8K\sigma^2} \Delta_{(2)}^2(T)\right). \end{aligned}$$

781

□

## 782 G Experimental Details

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

	$b$	$c$	$\psi$
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.

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

$$F = \left\{ f(x) = c \left( 1 - \frac{b}{(b^{1/\psi} + x)^\psi} \right) \right\},$$

where  $c, \psi \in (0, 1]$  and  $b \in [0, +\infty)$ . Note that, by construction, all the functions laying in  $F$  satisfy the Assumptions 2.1 and 2.2. In particular, the largest value of  $\beta$  satisfying Assumption 2.2, for the setting presented in Section 7, is  $\beta = 1.3$ . In Table 2, we report the value of the parameters characterizing the function employed in the the synthetically generated setting presented in the main paper.

## G.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:

$$a = \frac{25(T - K)}{36H_1},$$

where  $H_1 = \sum_{i \neq i^*(T)} \frac{1}{\Delta_i^2}$ ;

- R-UCBE: we used the value prescribed by Corollary 4.2 where we set the value  $\beta = 1.3$ ;
- ETC and Rest-Sure: we set  $\rho = 0.8$  and  $U = 1$  as suggested by Cella et al. (2021).

## G.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  $\approx 0.07$  seconds (on average), while a run of R-SR takes  $\approx 0.06$  seconds (on average).

## H 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 7.

### H.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 such a setting are presented in Figure 6b. In this case, the last time the optimal arm does not change anymore is  $T = 400$ . Similarly to the synthetic setting presented in the main paper, we have two different behaviors for time budgets  $T < 400$  and  $T > 400$ . For short time budgets, the algorithm

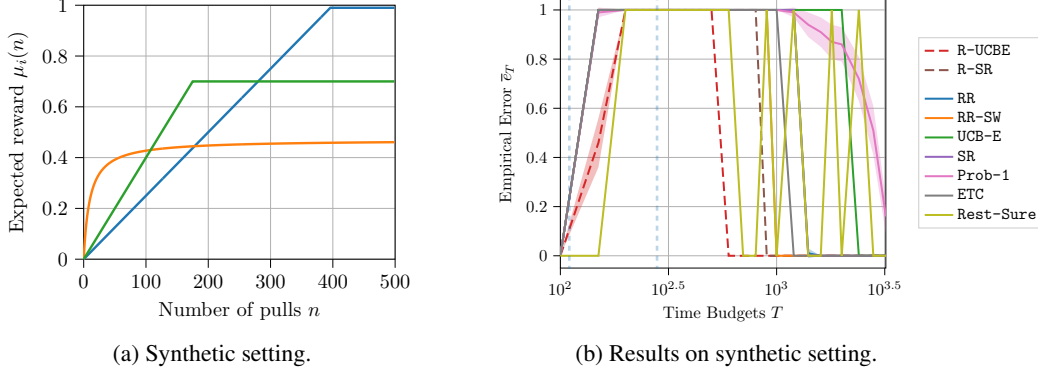


Figure 6: Challenging scenario in which the arm reward increment rate changes abruptly.

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  $\bar{e}_T(\text{R-UCBE}) < 0.01$ . 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  $T \approx 1000$ . 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.

## H.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  $\sigma$  changes in the collected samples. The setting we considered is the one described in Section 7. 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 ( $\sigma < 0.01$ ), 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  $\sigma \geq 0.05$ , the performance of R-UCBE starts to degrade, with behavior for  $\sigma = 0.5$  which is constant w.r.t. the chosen time budget with a value of  $\bar{e}_T(\text{R-UCBE}) = 0.8$ . 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  $\sigma$ . Indeed, only for  $\sigma = 0.5$ , 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  $\sigma = 0.5$ , in which the R-UCBE fails in providing a reliable recommendation for the optimal arm with a large probability.

## H.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  $T \in \mathcal{T} := \{500, 1000, \dots, 15000, 20000, 30000\}$ . 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  $\bar{e}_T(\mathcal{A})$ , we report the success rate  $R(\mathcal{A})$  instead, i.e., the ratio between the number of times an algorithm provides a correct suggestion and the number of budget values we considered, formally defined as  $R(\mathcal{A}) := \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \mathbb{1}\{\hat{I}^* = i^*\}$  (the larger, the better).

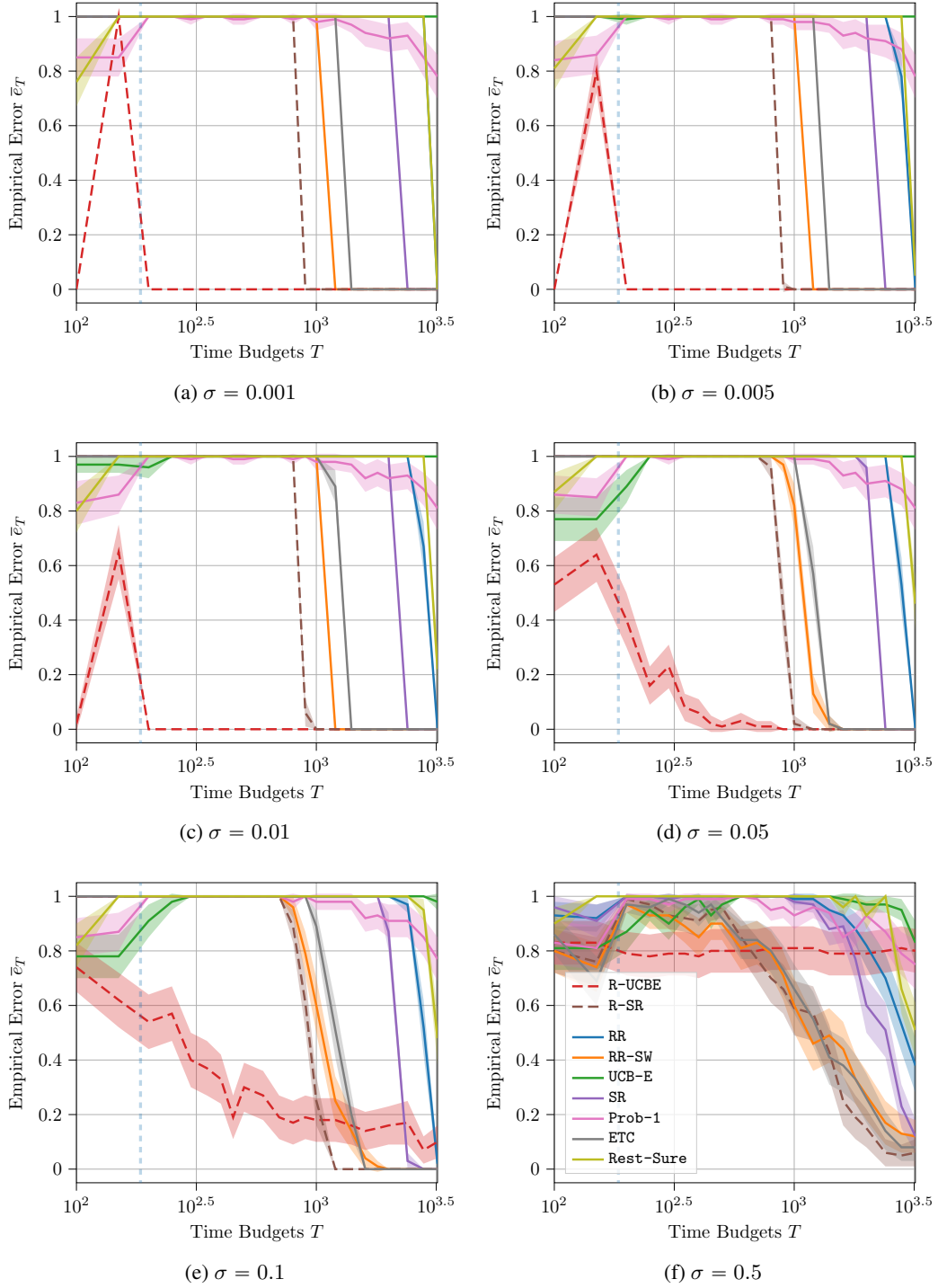


Figure 7: Empirical error probability for the synthetically generated setting, with different values of the noise standard deviation  $\sigma$ .

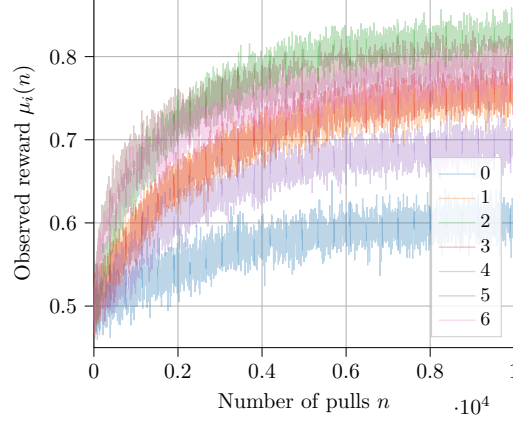


Figure 8: Rewards for the arms of the IMDB experiments.

		$T$												$R(\mathfrak{U})$
		500	1000	2000	3000	4000	5000	7000	10000	15000	20000	30000		
Algorithms	Optimal Arm	5	5	2	2	2	2	2	2	2	2	2		
	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.

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