Budgeted Improving Bandits

Matilde Tullii
Fairplay Team
Crest, Ensae, IP Paris

Nadav Merlis
Technion, Israel

Vianney Perchet
Fairplay Team
Crest, Ensae, IP Paris
AI Lab Criteo

Abstract

Motivated by the idea that large algorithmic infrastructures, such as neural networks, must carefully plan retraining procedures due to budget constraints, this work proposes a new multi-armed bandit framework in which the agent can modify the distribution of the arms up to M times throughout interaction. Each modification, referred to as a retraining step, leads to an improvement in the distributions that either increases the reward obtained or makes the optimal arm easier to identify. Specifically, we analyze two settings: in the first one (Improvable Arms), we assume that each retraining step increases the mean of all arms and reduces their variance. In the second setting (Decreasing Biases), we assume that the reward observations are obfuscated by some bias, which each retraining step helps to eliminate. We propose algorithms for both models and prove that they exhibit optimal rates with respect to the time horizon T.

1 Introduction

In recent years, the use of machine learning systems has seen unprecedented growth. Such innovation is accompanied by an exponential increase in the amount of data accessible to the algorithms; data that is often gradually expanded due to continued interaction with users, as in recommendation systems or AI chatbots. This poses new challenges for service providers: while having well-calibrated algorithms is fundamental for their widespread application and their reliability, executing a retraining process is often exceedingly costly and comes with substantial environmental impacts. One way to account for these costs is to limit the number of permitted retrainings (determined, for instance, by budget constraints). This raises the question of how to optimally schedule retrainings after the initial deployment.

We formulate this problem as a multi-armed bandit instance where an agent, while interacting with the environment, has the possibility to retrain the model up to M times. Each retraining step improves the arm distribution by either increasing the average reward or correcting internal biases, rendering the optimal arm easier to identify. As data keeps being collected over time, and retraining using a bigger dataset is always preferable, we assume that arm improvements become more significant the longer the agent delays the retraining. Thus, the agent faces a new dilemma: whether to benefit sooner from an earlier upgrade with a smaller impact or wait longer for a more substantial improvement. We study two different improvement models: in the first, each retraining yields a decrease in the reward variance, alongside a positive shift of the average of the distributions of the arms. This models the enhanced performance exhibited by an algorithm after a retraining process, which enables it to better recognize the optimal action and thus receive an increased reward. Such behavior could be observed, for instance, when retraining a recommendation system, such that the modified algorithm offers better suggestions (increased average) and leads to more consistent users' reactions (reduced variance). In the second setting, we assume the observations collected by the agent to be corrupted by some bias, as it is the case for models that are trained over limited amount of data, or for which the training dataset is not completely representative of the broader population Thus, each retraining step

18th European Workshop on Reinforcement Learning (EWRL 2025).

corresponds to a possibility to correct the algorithm, hence identifying and mitigating the effect of the bias. For both models, we design algorithms to select both a sequence of actions and retraining times that achieve optimal regret rates with respect to the time horizon T.

1.1 Outline and contributions

We introduce a novel multi-armed bandit setting, called Budgeted Improving Bandits (BIB), in which the agent can positively modify the distribution of the arms M times across the interaction. We particularly assume that the extent to which arms improve increases when these changes occur later in time. We analyze two different settings:

- BIB with Improvable Arms (BIB-IA): in this model, each retraining increases the mean of all arms by a constant factor $c_j \leq \overline{c}$ and reduces their variance by a factor bounded by $t^{-\alpha}$, with $\alpha \leq 1, \overline{c} \geq 0$ known parameters. We provide an algorithm that exhibits an optimal regret of $\mathcal{R}_T = \tilde{\mathcal{O}}(\max\{T^{1-\alpha/1-\alpha^{M+1}}, \overline{c}T^{1/2+\alpha}\})$ in high probability.
- BIB with Decreasing Biases (BIB-DB): in this setting, we assume the observations obtained to be biased. Each retraining of the model scales down the bias, from an initial variance of σ, by a factor again bounded by $t^{-\alpha}$. We propose an algorithm that achieves a regret rate of $\tilde{\mathcal{O}}(\sigma T^{1-\alpha/2/1-(\alpha/2)^{M+1}} + \phi\sqrt{T})$, where ϕ is the variance of the rewards.

The rest of the paper is organized as follows. In Section 2, we present the two different models, the sequential interaction protocol between the agent and the environment, and the notations used. Sections 3 and 4 respectively include the analysis of BIB-IA and BIB-DB. In both cases, we define the estimator for the average, the algorithm used, and provide bounds on the regret obtained by the algorithms and their optimality with respect to the time horizon T. The technical proofs for the results presented are deferred to Appendix A and B.

1.2 Related Work

The multi-armed bandit framework [Lattimore and Szepesvári, 2020] offers a powerful formalism to model sequential decision-making in an unknown environment. From the initial seminal works [Thompson, 1933, Robbins, 1952], a vast literature has originated, which examines countless variants of the original problem. Our work bears some resemblance to settings in which the distribution of the arms might change during the interaction, as in non-stationary bandits [Besbes et al., 2014, Wei and Luo, 2021, Auer et al., 2019], restless bandits [Whittle, 1988, Guha et al., 2010, Ortner et al., 2012], and rising bandits Heidari et al. [2016], Montenegro et al. [2023], Metelli et al. [2022]. However, the fact that in our case the changes in the distributions are directly controlled by the agent marks a significant difference with these works, whose main results aim at outlining algorithms able to face unknown and possibly abrupt changes in the distribution, or for which the change in the rewards of an arm across rounds is a function of the number of times the arm has been played.

Some comparison can be drawn as well with the field of batched bandits [Perchet et al., 2015, Gao et al., 2019, Esfandiari et al., 2021], where the time horizon T is considered to be divided into batches by a grid of time steps, and the learner can only rely on the information gathered during previously concluded batches. While there is no change of distribution between different batches, this line of research also focuses on understanding how to select the optimal grid of times in which we split the total run, in order to minimize the excess regret.

Another related line of research is multi-fidelity multi-armed bandits [Kandasamy et al., 2016b,a, Wang et al., 2023, Poiani et al., 2024]. Inspired by multi-fidelity in other fields (*e.g.* Bayesian optimization), this framework studies a setting in which, for every arm, the agent has access to m different approximations of the distribution of its reward. These approximations are increasing in quality, meaning that their average concentrates tighter around the true one, but also increasing in the cost that choosing each of them yields. The agent can choose which fidelity to require at every time step, aiming to simultaneously minimize the regret and the overall costs of playing. Despite some similarities with our second model, as in the existence of several degrees in the quality of the observations, our setting is again quite different, since the agent is granted only a limited number of switches (M) among different distributions and cannot choose to downgrade to a worse quality observation after having retrained the model. Finally, the extent to which the distributions of the arms

in our problem improve depends on the retraining time; this is a key aspect that does not exist in the multi-fidelity case.

Lastly, one way to interpret our setting is through the lens of scaling laws of large language models [Kaplan et al., 2020, Hestness et al., 2017, Cherti et al., 2023]. This term refers to power laws that link the performance of a model to its parameters, such as the number of input parameters, the size of the training dataset, or the computing power. Several empirical works have highlighted how the performance of the predictors with respect to the size of the training dataset m scales as $1/m^{\alpha}$ with $\alpha \leq 1$. This motivated our choice of the retaining function as further detailed in Section 2.

2 Setting

Consider a multi-armed bandit model where, across T rounds, an agent sequentially chooses an action among a set $\mathcal{K}$ (with $|\mathcal{K}| = K$) and receives a reward drawn from an associated unknown distribution. Specifically, at each round t, the agent plays an action i_t and earns a reward $r_t^{\tau}(i_t)$. It then either observes the real reward or a biased observation $o_t^{\tau}(i)$ (both rewards and observations are to be specified shortly). Before each round starts, the agent can decide to *retrain* the model. This retraining step corresponds to an improvement in the distribution of all the arms, which either increases the average reward and reduces their variance, or makes the optimal arm easier to identify by correcting its bias. We denote the hard limit on the number of permitted retrainings by M and by $\tau = \{\widehat{t}_1, \dots, \widehat{t}_M\}$, the set of time-steps at which retraining happens, with the convention $\widehat{t}_0 = 0, \widehat{t}_{M+1} = T$. We also use the notation $\lfloor t \rfloor_{\tau}$ ($\iota(\lfloor t \rfloor_{\tau})$) to denote the latest step (retraining index) at which the model was retrained before time t, that is,

$$\lfloor t \rfloor_{\tau} = \max_{j \in [M]} \left\{ \widehat{t}_j : t \geq \widehat{t}_j \right\} \quad \text{and} \quad \iota(\lfloor t \rfloor_{\tau}) = \operatorname*{argmax}_{j \in [M]} \left\{ \widehat{t}_j : t \geq \widehat{t}_j \right\}.$$

We now explicitly specify that two improvement models, determining both $r_t^{\tau}(i_t)$ and $o_t^{\tau}(i)$, both depend on a decreasing function $\mathbb{s}$ that encapsulates the retraining gain.

Budgeted Improving Bandits with Improvable Arms (BIB-IA) In this case, each retraining step j has the effect of increasing the average of the distribution by a positive quantity c_j, and reducing the variance by a factor depending on the time when such retraining occured. This models the intuitive idea that an improvement in the infrastructure leads to better and less fluctuating rewards. Formally, we assume the reward of arm i to be a random variable of the form

$$r_t^{\tau}(i) = \mu_t^{\tau}(i) + \mathbb{s}(\lfloor t \rfloor_{\tau}) \xi_t \coloneqq \mu(i) + \sum_{j=0}^{\iota(\lfloor t \rfloor_{\tau})} c_j + \mathbb{s}(\lfloor t \rfloor_{\tau}) \xi_t \,, \tag{1}$$

where ξ_t is a σ-conditionally subgaussian centered random variable. After playing an arm i_t, the agent gets to observe the true reward $r_t^{\tau}(i_t)$. In the equation above, $\mu_t^{\tau}(i)$ indicates the average reward at time t, composed of the initial average $\mu(i)$ and the sum of the increments c_j, accumulated thanks to the previous retraining processes. The individual values c_j are unknown to the agent; yet, we assume $c_j \leq \overline{c},\ \forall j = 1, \dots, M$ where $\overline{c}$ is a known quantity, and denote $c_0 = 0$.

Budgeted Improving Bandits with Decreasing Biases (BIB-DB) In the second model, we instead assume that the agent does not have access to the true reward resulted from choosing a certain action, but only receives an observation corrupted by some bias. The effect of a retraining process is to reduce this bias. Formally, the reward and observation obtained by playing arm i at time t with retraining steps τ are given by

$$\begin{aligned} r_t^{\tau}(i) &= \mu(i) + \varepsilon_t \\ o_t^{\tau}(i) &= r_t^{\tau}(i) + \mathbb{s}(\lfloor t \rfloor_{\tau}) \xi_{\lfloor t \rfloor_{\tau}}(i) \,. \end{aligned} \tag{2}$$

Thus, the observation aggregates the reward at time t, $r_t^{\tau}(i)$, depending only on the quality of the arm i, and a bias term *that remains fixed between retraining steps* (i.e., $\mu_t^{\tau}(i) = \mu(i)$). We assume $\xi_{\lfloor t \rfloor_{\tau}}$ and ε_t to be mutually independent and, similarly to the previous model, respectively σ and ϕ-conditionally subgaussian centered random variables.

Algorithm 1 Sequential interaction between the agent and the environment

Input: Time horizon T, arms $\mathcal{K}$, parameter α.

while $t \leq T$ **do**
  Agent picks action i_t
  if $t = \widehat{t}_j \in \tau$ **then**
    Set $\lfloor t \rfloor_\tau = \widehat{t}_j$
  end if
  BIB-IA Agent receives reward $r_t^\tau(i_t)$
  BIB-DB Agent receives reward $r_t^\tau(i_t)$, while observing $o_t^\tau(i_t)$
  $t \leftarrow t + 1$
end while

In both models, we assume that for all the arms, the associated reward and observation distributions have support bounded by R regardless of the retraining times τ.

In order to derive meaningful results, we make the following assumption on the function $\mathbb{s}$ that governs the decrease in variance/bias between retraining epochs.

Assumption 1. *There exists $\alpha \in [0, 1]$ such that $\mathbb{s}(x) \leq x^{-\alpha/2}$ if $x > 0$ and $\mathbb{s}(0) < 1$.*

Remark 1. *The proposed algorithms and the concentration bounds derived for the estimators also work for more general functions $\mathbb{s}$. Still, the assumed bounds allow us to recover precise expressions for the optimal retraining times and sharper regret bounds. In particular, this assumption is inspired by the scaling laws of large language models. Such laws relate the performance of neural networks to the size of their input parameters, as the dimension of the training set. Specifically, they have been demonstrated [Kaplan et al., 2020] to exhibit a power-like behavior with exponents smaller than 1.*

The sequential interaction for both models is detailed in Algorithm 1.

As standard in the multi-armed bandit literature, the performance of an algorithm is measured in terms of (pseudo)-regret. The goal is to design an algorithm that chooses the retraining times τ and a sequence of actions that minimize the quantity

$$\mathcal{R}_T = \sum_{t=1}^{T} \mu_t^{\tau^*}(i^*) - \mu_t^\tau(i_t)\,. \tag{3}$$

Our benchmark, in this case, is an omniscient adversary who knows the distributions of the rewards, hence the optimal arm i^*, and chooses the optimal retraining times τ^*. Note that if i^* is known, then the optimal retraining times are also deterministic and problem-independent. Indeed, in the first model, the optimal strategy is to promptly exploit all the retraining budget, to benefit from the maximal increased average for as many rounds as possible, from which $\tau^* = \{1, \dots, M\}$. In the other case, since the bias term impacts only the observations, the retraining times have no effect on the true rewards. Their sole role is to make the best arm more easily identifiable, hence the optimal strategy is independent of them and any choice of τ^* would yield to the same cumulative reward.

Remark 2. *In the problem we are considering, these changes in the distribution are meant to model a costly retraining process that must be carefully scheduled; therefore, we limit the analysis to the case $M \ll T$. Note that the case $M \simeq T$ is of independent interest, since it corresponds to a framework in which the distribution of the arms continuously varies with time, hence equivalent to a nonstationary bandit setting. Specifically Roychowdhury et al. [2025] studies a bandit problem in which the distribution of every arm at time t is Gaussian with variance σ^2/t. Such setting could be traced back to our BIB-IA model, for $M = T$ and a specific choice of the parameters involved.*

Notations In what follows, we denote by $[[a, b]]$ the set of all integers between a and b and $\wedge$, $\vee$ represent respectively the minimum and the maximum between two quantities. $R_T = \tilde{\mathcal{O}}(B_T)$ means that there exists a (possibly problem-dependent) constant C such that $R_T = \mathcal{O}(\log(T)^C B_T)$.

3 Budgeted Improving Bandits with Improvable Arms (BIB-IA)

In this section, we analyze the BIB-IA model introduced in Section 2, where the rewards follow the distribution detailed in Equation (1). An example of such a setting can be found in recommendation

or advertising systems. In these cases, a retraining process is an improvement of the underlying algorithm, responsible for suggesting content, which better identifies users' preferences. Therefore, after each of these modifications, when proposing an ad, which corresponds to playing a specific arm in the bandit setting, it will be possible to observe more consistent clients' reactions and an increased click-through rate, which can be respectively translated into decreased variance and improved average of the distributions of the rewards. In this case, the agent faces two different types of exploration-exploitation dilemmas. On one side, there is the intrinsic one of the bandit framework, which forces the agent to choose between exploring the actions, hence gathering more information about their distributions, or exploiting what seems to be the optimal action, with the risk of suffering considerable regret if this is not the case. On the other side, the agent also has to decide when to retrain the model, knowing that a longer wait yields a more significant variance improvement. Simultaneously, since each retraining comes with an increase in the average for every arm, the agent might decide to sacrifice some possible improvement in the variance to favor an earlier increase in the average reward of each arm. In contrast, the optimal strategy of the benchmark, already knowing the best arm, should exhaust all the retraining budget in the first rounds to benefit the most from the improved reward. Thus, the total regret is caused by both the delay in the retraining and the choice of suboptimal arms. In the rest of this section, we show how to balance both these tradeoffs in order to obtain order-optimal regret bounds.

Let τ be a set of retraining times. We consider an estimator for the average $\mu(i)$ of each arm, defined as a convex combination of the rewards obtained at each round. Formally, let

$$\Sigma_t^\tau(i) = \sum_{p \leq t} \frac{1}{\sigma^2(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}} \mathbb{1}\{i_p = i\} \vee 1$$

then, the estimator is defined as

$$\widehat{r}_t^\tau(i) = \frac{1}{\Sigma_t^\tau(i)} \sum_{p \leq t} \frac{r_p^\tau(i)}{\sigma^2(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}} \mathbb{1}\{i_p = i\}\,. \tag{4}$$

This estimator allows us to define an upper and lower confidence bound for each arm's initial average $\mu(i)$ since it satisfies the following concentration result, adapted to our case from the standard Hoeffding-Azuma inequality.

Lemma 1. *Fix an arm i, consider $(r_t^\tau(i_t))_{t\geq 1}$ the sequence of random rewards, which is adapted with respect to the filtration $\mathcal{F}_t = \sigma(r_1^\tau(i_1), \dots, r_{t-1}^\tau(i_{t-1}))$. Then for any $t \in \mathbb{N}_*$ and $\delta_t \in (0,1)$, the estimator defined in Equation* (4) *satisfies*

$$\Sigma_t^\tau(i) = 0 \text{ or } \left| \widehat{r}_t^\tau(i) - \left(\mu(i) + \sum_{j=0}^{\lfloor t \rfloor_\tau} c_j \right) \right| \leq 2\sqrt{\frac{\log(1/\delta_t)}{\Sigma_t^\tau(i)}}$$

with probability at least $1 - 2T^\alpha t\delta_t$.

The algorithm we consider is an instance of Successive-Elimination, whose details are described at Algorithm 2. The upper and lower confidence bounds, thanks to the lemma above, are defined as

$$\begin{aligned} UCB_t^\tau(i,\delta) &= \widehat{r}_t^\tau(i) + 2\sqrt{\frac{\log(1/\delta)}{\Sigma_t^\tau(i)}} \\ LCB_t^\tau(i,\delta) &= \widehat{r}_t^\tau(i) - 2\sqrt{\frac{\log(1/\delta)}{\Sigma_t^\tau(i)}}\,. \end{aligned} \tag{5}$$

In the definition of Algorithm 2, we require both the elimination of the arms and the retraining to happen only at the end of loops inside which all the still active arms have been played in a round-robin fashion. In what follows, we define an auxiliary set of switching times $\tilde{\tau} = \{\tilde{t}_1, \dots, \tilde{t}_M\}$, which are given in input to the algorithm. During the run, they will be approximated such that each instant $\widehat{t}_j \in \tau$ corresponds to the rounding of $\tilde{t}_j$ to the first instant of time, after $\tilde{t}_j$, where a loop over all the

Algorithm 2 Successive Elimination for BIB

Input: Time horizon T, arms $\mathcal{K}$, parameter α, set of set of active arms $\mathcal{A}_0 = \mathcal{K}$, estimates $UCB_0^\tau(i,\delta) = LCB_0^\tau(i,\delta) = 0,\ \forall i \in \mathcal{K}$, confidence parameter δ, designated retratining times $\tilde{\tau} = \{\tilde{t}_1, \dots, \tilde{t}_M\}$, adaptive retraining times $\tau = \varnothing$.

while $t \leq T$ **do**
  for $i \in \mathcal{A}_{t-1}$ **do**
    Agent picks action $i_t = i$
    BIB-IA Agent receives $r_t^\tau(i_t)$ and updates $UCB_t^\tau(i,\delta)$ and $LCB_t^\tau(i,\delta)$
    BIB-DB Agent receives $r_t^\tau(i_t)$, observes $o_t^\tau(i_t)$ and updates $UCB_t^\tau(i,\delta)$ and $LCB_t^\tau(i,\delta)$
    $\mathcal{A}_t \leftarrow \mathcal{A}_{t-1}$
    $t \leftarrow t+1$
  end for
  $z := \max_{k \in \mathcal{A}_{t-2}} LCB_{t-1}^\tau(k,\delta)$
  $\mathcal{A}_{t-1} \leftarrow \{i \in \mathcal{A}_{t-2},\ UCB_{t-1}^\tau(i,\delta) \geq z\}$
  if $\lfloor t \rfloor_\tau < \lfloor t \rfloor_{\tilde{\tau}}$ **then**
    Set Retrain the model and add $\tau \leftarrow \tau \cup \{t\}$
  end if
end while

still active arms has ended. Note that this creates adaptivity in the definition of τ, since the precise value of the retraining times is a random quantity, as it depends on the number of active arms which is, in its turn, a random quantity depending on the interaction between the agent and the environment.

The algorithm enjoys the following regret bound:

Theorem 1. *Let $\beta(j)$ be the function $\beta(j) = \frac{1-\alpha^j}{1-\alpha}$, and c^* and λ the quantities*

$$c^* = A\left(\frac{T}{M^2}\right)^{\frac{1/2-\beta(M)}{1+\alpha\beta(M)}} \tag{6}$$

and

$$\lambda = \left[A^2 \frac{T}{M^2} \min\left\{\frac{1}{c^{*2}}, \frac{1}{\bar{c}^2}\right\}\right]^{\frac{1}{(2+\alpha)\beta(M)}}, \tag{7}$$

where $A = 8\sigma\sqrt{K\log(4KT^{\alpha+5})}$. Then, the regret of Algorithm 2, with estimate as in Equation (4)*, confidence intervals as in Equation* (5) *and confidence parameter $\delta = 1/4KT^{5+\alpha}$, satisfies*

$$\mathcal{R}_T \leq \tilde{\mathcal{O}}\left(K^{\frac{3}{2}+\frac{1}{2\beta(M)(2+\alpha)}} M^{1+\frac{1}{\beta(M+1)}} \cdot \max\left\{T^{\frac{1}{2\beta(M+1)}},\ \bar{c}T^{\frac{1}{2+\alpha}}\right\}\right)$$

with probability at least $1 - 1/T$, when the retraining times are chosen as $\tilde{t}_j = \lambda^{\beta(j)}$.

Sketch of proof. (See Appendix A.2.1 for the full proof) The proof works by initially decomposing the regret as

$$\mathcal{R}_T = \sum_{t=1}^{T}\left(\mu_t^{\tau^*}(i^*) - \mu_t^\tau(i^*)\right) + \sum_{t=1}^{T}(\mu_t^\tau(i^*) - \mu_t^\tau(i_t)).$$

The first term corresponds to the regret suffered due to an incorrect choice of the retraining times with respect to the optimal strategy and is bounded by $\bar{c}M\widehat{t}_M$. The second term, instead, is caused by the choice of suboptimal arms. We further decompose it into the regret accumulated in each of the $M+1$ epochs. Using the regret decomposition lemma [Lattimore and Szepesvári, 2020, Lemma 4.5] and the concentration bound for the estimator, we find that for each epoch j, the regret is of the order of $N_j^\tau(i)/\sqrt{\Sigma_{\widehat{t}_{j+1}}^\tau(i) - \Sigma_{\widehat{t}_j}^\tau(i)}$, where $N_j^\tau(i)$ is the number of times arm i has been played during epoch j. Since during the j-th epoch, the variance of arm is constant and equal to $\widehat{t}_j^{-\alpha}$, the difference $\Sigma_{\widehat{t}_{j+1}}^\tau(i) - \Sigma_{\widehat{t}_j}^\tau(i)$ is of the order $\widehat{t}_j^\alpha N_j(i)$. Furthermore, using the fact that $N_j(i) \leq \widehat{t}_{j+1}$, we prove that the regret accumulated in each epoch is proportional to $\sqrt{\widehat{t}_{j+1}/\widehat{t}_j^\alpha}$. From which,

$$\mathcal{R}_T \leq \bar{c}M\widehat{t}_M + \sum_{j=0}^{M}\sqrt{\widehat{t}_{j+1}/\widehat{t}_j^\alpha},$$

and substituting the choice of the retraining times as specified in the statement yields the result. □

The behavior of the switching times depending on the parameter $\overline{c}$ The switching times specified in Theorem 1, illustrate the optimal strategy to face this bandit problem depending on the parameter $\overline{c}$. Indeed, when $\overline{c}$ is small, and by consequence so are the individual increases c_j, the agent should prioritize a choice of retraining times that allows her to gather the biggest amount of information, aiming at reducing the variance, and consequently minimizing the component of the regret caused by an incorrect choice of the arm. Conversely, when $\overline{c}$ grows, the priority should be given to improving the average, since the increase in the reward brought by each retrain step is substantial. This is reflected in the definition of the switching times through to the quantity λ which is inversely proportional to $\overline{c}$, such that for large values of $\overline{c}$ the retraining times converge to the optimal ones $\tau^* = \{1, \dots, M\}$. A particular case is when $\overline{c} = 0$, corresponding to the one in which there is no improvement in the average generated by a retraining step, which instead has the only effect of reducing the variance. Similarly as before it is possible to decompose the regret as $\mathcal{R}_T = \sum_{j=0}^{M} \sqrt{\widehat{t}_{j+1}/\widehat{t}_j^{\alpha}}$, and the optimal retraining times are therefore obtained by solving the constraint optimization problem, imposing that $0 = \widehat{t}_0 \leq \ldots \leq \widehat{t}_{M+1} \leq T$, leading to a regret of the order of $T^{\frac{1}{2\beta(M+1)}}$. Note that this is, for all values of α, better than the canonical regret rate of bandits ($\sqrt{T}$) as expected since the optimal arm becomes more and more easy to identify across the interaction.

The rate obtained in Theorem 1 is optimal with respect to T, as shown by the following theorem.

Theorem 2. *Let ν be a bandit instance, whose rewards, for every possible action and time step, follow the structure detailed in Equation* (1)*. Let $\mathcal{A}$ be a random algorithm that selects both an arm to be played at every timestep t, and a set of M retraining times $\tau = \{\widehat{t}_1, \dots, \widehat{t}_M\}$. Then, there exists a bandit instance ν for which, every choice of τ, the expected regret of algorithm $\mathcal{A}$ with horizon T is lower bounded by*

$$\mathcal{R}_{\mathcal{A},\nu,T} \geq \Omega\left(\left(\frac{\sigma\sqrt{K}}{M+1}\right)^{\frac{2}{2+\alpha}} \max\left\{T^{\frac{1}{2\beta(M+1)}}, \overline{c}^{\frac{\alpha}{2+\alpha}} T^{\frac{1}{2+\alpha}}\right\}\right).$$

Finally, we analyzed a full-information setting, in which at every round the agent observes the rewards for all arms. In this case, an analogous estimator to the one defined in Equation (4) and a greedy routine allow to recover the following regret bound, which has the same dependency in T of Theorem 1, and only improves in terms of K.

Theorem (Informal statement)**.** *Let $c^* = \tilde{\mathcal{O}}(T^{\frac{1/2-\beta(M)}{1+\alpha\beta(M)}})$ and $\lambda = \tilde{\mathcal{O}}(T/\max\{c^{*2}, \overline{c}^2\})^{\frac{1}{(2+\alpha)\beta(M)}}$. Then the regret of a greedy algorithm with retraining times $\widehat{t}_j = \lambda^{\beta(j)}$ is*

$$\mathcal{R}_T = \tilde{\mathcal{O}}\left(\max\left\{T^{\frac{1}{2\beta(M+1)}}, \overline{c}T^{\frac{1}{2+\alpha}}\right\}\right)$$

All the details regarding this model, such as the definition of the estimator, the algorithm, and the results can be found in Appendix A.1

4 Budgeted Improving Bandits with Decreasing Biases (BIB-DB)

In this section, we analyze the BIB-DB setting, where we assume that the agent interacts with a more challenging bandit instance, in which for every played action, she receives a sample of the reward further corrupted by some bias. This setting represents the case in which the reward resulting from a certain action is mediated by the model itself, which, for instance, translates a qualitative feedback into a quantitative reward, as could be in a reviewing system. In this case, the resulting value is exposed to some degree of subjectivity and likely presents some bias. An improvement of the algorithm represents a correction of the embedding mechanism, such that after its deployment, the rewards observed are more representative of the true value. Another example is the case where the information collected comes, for instance, from a distinct population, whose individual characteristics impact the distribution of the reward, as might be the case for surveys or clinical trials. In this scenario, a retaining step corresponds to an opportunity for the service provider to improve the algorithm by sourcing data from heterogeneous datasets. By comparing diverse batches of information, it is therefore possible to estimate the bias and reduce it. Again, we

assume the impact of this correction to be proportional to the time in which the retraining is executed, motivated by the idea that a longer wait yields richer datasets. The final regret rates will depend on the initial magnitude of the bias, the number of retrainings allowed, and the influence each one has thanks to the parameter α, as highlighted by the following theorems.

Analogously to the analysis in Section 3, we use the instance of Successive-Elimination defined in Algorithm 2, which employs an estimator of the average defined as follows. Fix an arm i, and assume there exists some constants γ_j such that, if i is still active at time $\widehat{t}_{j+1}$ then $\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\} > \gamma_j$ almost surely; we will later specify these constants. Then, we can define

$$\widehat{r}_t^j(i) = \frac{1}{N_j^\tau(i) \vee 1} \sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}\wedge t} o_p^\tau(i) \mathbb{1}\{i_p = i \wedge N_j^\tau(i) \leq \gamma_j\}$$

where $N_j^\tau(i) = \sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}\wedge t} \mathbb{1}\{i_p = i \wedge N_j^\tau(i) \leq \gamma_j\}$ counts the times an arm has been played during each epoch, stopped at γ_j. Note that $\widehat{r}_t^j(i)$ corresponds to the average of the observations obtained during epoch j where only the initial γ_j information is considered, and the following observations are discarded. The final estimator is again obtained as a weighted average across epochs. Specifically, let

$$\Pi_t^\tau(i) = \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{1}{\sigma^2(\lfloor t \rfloor_\tau \vee 1)^{-\alpha} + {\phi^2}/{N_j^\tau(i)}},$$

then, the estimator is defined as

$$\widehat{r}_t^\tau(i) = \frac{1}{\Pi_t^\tau(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\widehat{r}_t^j(i)}{\sigma^2(\lfloor t \rfloor_\tau \vee 1)^{-\alpha} + {\phi^2}/{N_j^\tau(i)}}, \tag{8}$$

where we use the convention ${1}/{0} = \infty$ and ${1}/{\infty} = 0$. By the linearity of expectation, it is straightforward to see that for i.i.d. samples, $\mathbb{E}\left[\widehat{r}_t^\tau(i)\right] = \mu(i)$. We prove that a similar result holds for the adaptive case as well, as stated in the following lemma.

Lemma 2. *Consider an arm i and the sequence of random observations $(o_t^\tau(i_t))_{t\geq 1}$, which is adapted with respect to the filtration $\mathcal{F}_t = \sigma\left((\xi_{\widehat{t}_0}(i), \dots, \xi_{\lfloor t \rfloor_\tau}(i)), (i_1, \dots, i_t), (\epsilon_1(i_1), \dots, \epsilon_t(i_t))\right)$. Let $\widehat{r}_t^\tau(i)$, then the estimator defined in Equation* (8)*, then for every $\delta \in (0,1)$, it holds*

$$\mathbb{P}\left(|\widehat{r}_t^\tau(i) - \mu(i)| \leq R\sqrt{\frac{32\log({4\gamma_{\iota(\lfloor t \rfloor_\tau)}}/{\delta})}{\Pi_t^\tau(i)(\sigma^2+\phi^2)}}\right) \geq 1-\delta$$

where R is a bound on the support of the distribution of the observations.

This allows us to define upper and lower confidence bounds necessary for Algorithm 2. Specifically,

$$\begin{aligned} UCB_t^\tau(i,\delta) &= \widehat{r}_t^\tau(i) + R\sqrt{\frac{32\log({4\gamma_{\iota(\lfloor t \rfloor_\tau)}}/{\delta})}{\Pi_t^\tau(i)(\sigma^2+\phi^2)}} \\ LCB_t^\tau(i,\delta) &= \widehat{r}_t^\tau(i) - R\sqrt{\frac{32\log({4\gamma_{\iota(\lfloor t \rfloor_\tau)}}/{\delta})}{\Pi_t^\tau(i)(\sigma^2+\phi^2)}}. \end{aligned} \tag{9}$$

Utilizing this concentration result, we derive the following regret bound.

Theorem 3. *Let κ be the function $\kappa(j) = \frac{1-({\alpha}/{2})^j}{1-{\alpha}/{2}}$, and λ the quantity $\lambda = T^{{1}/{\kappa(M+1)}}$. Consider the estimates defined as in Equation* (8)*, with $\gamma_j = \frac{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}}{K} - 1$. Then, the regret of Algorithm 2, using these estimates, with confidence intervals as in Equation* (9) *and confidence parameter $\delta = {1}/{K^2T^4}$, is bounded by*

$$\mathcal{R}_T \leq \tilde{\mathcal{O}}\left(MK\left(\sigma T^{\frac{1}{\kappa(M+1)}} + \phi\sqrt{KT}\right)\right)$$

with probability $1 - {1}/{T}$, when the retraining times are picked as $\tilde{t}_j = \lambda^{\kappa(j)}, \forall j = 1, \dots, M$.

As before, the retraining times specified in the statement represent preliminary retraining steps The true value of each $\widehat{t}_j$ is obtained by rounding the corresponding $\tilde{t}_j$ to the end of the loop where all the still active arms are played in a round-robin way, as done in Section 3.

Remark 3. *Note that the choice of γ_j as expressed in the statement is valid. Indeed, consider the true retraining time $\widehat{t}_j$, defined adaptively. Let i be an arm still active at the end of the j-th epoch and $\mathcal{A}_j$ be the set of active arms at time $\widehat{t}_j$. Since the algorithm plays the arms in a round-robin way and the number of active arms during epoch j can only decrease, we have*

$$\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\} \geq \frac{\widehat{t}_{j+1} - \widehat{t}_j}{|\mathcal{A}_j|}\,.$$

Moreover, since $|\mathcal{A}_j| \leq K$.

$$\sum_{p=\widehat{t}_j+1}^{\widehat{t}_j} \mathbb{1}\{i_p = i\} \geq \frac{\widehat{t}_{j+1} - \widehat{t}_j}{K} = \frac{\lceil \tilde{t}_{j+1} \rceil^{\mathcal{A}} - \lceil \tilde{t}_j \rceil^{\mathcal{A}}}{K} \geq \frac{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}}{K} - 1$$

where $\lceil \cdot \rceil^{\mathcal{A}}$ represents the rounding to the end of the round-robin loop. The last inequality follows from the fact $\forall j,\ \tilde{t}_j \leq \lceil \tilde{t}_j \rceil^{\mathcal{A}} \leq \tilde{t}_j + K$.

Sketch of proof. (See Appendix B.2.1 for the full proof) Similarly as before, it is possible to decompose the total regret as $\mathcal{R}_T = \sum_j \mathcal{R}_j$, where $\mathcal{R}_j$ is the regret accumulated during epoch j. Using the regret decomposition lemma [Lattimore and Szepesvári, 2020, Lemma 4.5] and the concentration bound of Lemma 2, we find that each $\mathcal{R}_j$ is of the order of $\tilde{N}_j(i)/\sqrt{\tilde{\Pi}^{\tau}_T(i)}$, where $\tilde{N}_j(i)$ is the total amount of times arm i has been played during epoch j and $\tilde{\Pi}^{\tau}_T(i)$ the term in $\Pi^{\tau}_T(i)$ relative to epoch j. By further expliciting the value of $\tilde{\Pi}^{\tau}_T(i)$, the terms rewrite as $\tilde{N}_j(i)/\sqrt{N_j(i)}(\sigma\sqrt{N_j(i)/\widehat{t}_j^{\alpha}} + \phi)$. Note that, due to its definition, $N_j(i) \leq (\widehat{t}_{j+1} - \widehat{t}_j/K) - 1$ and when the inequality is strict, we have $N_j(i) = \tilde{N}_j(i) \leq \widehat{t}_{j+1} - \widehat{t}_j$. Hence in both cases, if replaced in the regret decomposition, we obtain that $\mathcal{R}_j$ is bounded by $\sigma\left(\widehat{t}_{j+1}/\sqrt{\widehat{t}_j^{\alpha}}\right) + \phi\sqrt{\widehat{t}_{j+1}}$. The final result follows by considering a sum over all the epochs and replacing the retraining times as in the statement. □

From the regret bound of Theorem 3, it is possible to gather some intuition regarding the nature of this model. The bound is composed of two independent terms $\sigma T^{1/\kappa(M+1)}$ and $\phi\sqrt{T}$. The latter corresponds to the standard regret rate of a bandit instance where the distribution of the reward is ϕ-subgaussian, while the former is distinctive of our model and is caused by the bias. Note that in general, $\sigma T^{1/\kappa(M+1)}$ is the leading term, as expected since the obfuscated observations that the agent receives in this setting make the problem harder than in the canonical multi-armed-bandit framework. The only cases in which the regret bound obtained matches the standard $\sqrt{T}$ rate, are for large values of M if $\alpha = 1$, or if $\sigma \ll \phi$. Both cases represent occurrences in which the bias term is eventually negligible, either because its variance σ is substantially smaller than the noise variance ϕ, or because each of the frequent retraining steps leads to major improvements, again significantly shrinking σ.

To prove the optimality of this bound, we provide the following result.

Theorem 4. *Let ν be a bandit instance, whose rewards, for every possible action and time step follow the structure detailed in Equation* (2)*. Assume that for every time t, the agent receives an observation from every arm, in a full-information fashion. Let $\mathcal{A}$ be a random algorithm that selects both an arm to be played at every time t, and a set of M retraining times $\tau = \{\widehat{t}_1, \dots, \widehat{t}_M\}$. Then, there exists a bandit instance ν, for which, for every choice of τ, the expected regret of algorithm $\mathcal{A}$ with horizon T is lower bounded by*

$$\mathcal{R}_{\mathcal{A},\nu,T} \geq \Omega\left(\max\left\{\sigma T^{1/\kappa(M+1)}, \phi\sqrt{T}\right\}\right)\,.$$

Note that Theorem 4 considers a full-information version of the BIB-DB setting. The result also implies an analogous bound in terms of T for the case in which the agent can only rely on bandit feedback, since bandit feedback is more restrictive than full information. Still, the bandit case often

has a worse dependency on the number of arms K that is not captured in this lower bound, which also exists in our upper bound.

For the full-information version of this model, whose details can be found in Appendix B.1, we also provide an upper bound on the regret rate. The algorithm we outline for this is an instance of a greedy routine that utilizes a similar estimator to the one defined in Equation (8), updating it for all the arms at every round. The bound of the regret in this case has the same dependency on T as in Theorem 3, as shown by the following.

Theorem (Informal statement)**.** *In a BIB-DB setting with full information, the regret of a greedy algorithm is bounded by*

$$\mathcal{R}_T \leq \tilde{\mathcal{O}}(\sigma T^{1/\kappa(M+1)} + \phi\sqrt{T})$$

when retraining times are defined as $\widehat{t}_j = \lambda^{\kappa(j)}$ *and* $\lambda = T^{1/\kappa(M+1)}$.

5 Conclusion

In this work, we presented a novel multi-armed bandit framework, called budgeted improving bandits, in which the agent can modify the distribution of all the arms for a fixed number of times M. We analyzed two different ways in which such action impacts the distribution of the arms. In the first, we assume each retraining step to augment the average of the reward distribution and decrease its variance when the retraining is executed. In the second, we assume that the observations obtained by the agent are corrupted by some bias, and each retraining step serves to shrink this bias. In both models, the longer the agent waits before retraining, the better its gain, and agents must decide whether to update early, gaining small improvement for a longer duration, or wait with the update in order to get a larger performance boost for a short period of time. For both variants, we provide algorithms that exhibit optimal rates with respect to the time horizon T. We leave characterizing the optimal dependence of the regret on the number of arms K to future work.

Given the novelty of the model, many possible research directions stem from this work. It would be interesting, for instance, to understand what regret rates are achievable when the improvement that each retraining yields is not known beforehand. In particular, one could study whether, given enough retraining steps, the improvement parameter α could be estimated on the fly. Alternatively, our work can be extended to support more general improvement profiles $\mathbb{s}$, including non-polynomial parametric families or even assuming it only belongs to a specific class of functions, as in monotone Lipschitz or Hölder functions. In terms of performance, we provide minimax regret bounds that are order optimal w.r.t. the interaction length. It is worthwhile to characterize the optimal regret also w.r.t. the number of arms and update steps, as well as derive instance-dependent bounds. Lastly, it is possible to extend other formulations that model retraining, including problems in which the number of retraining steps is unlimited but each comes with some cost, considering retraining for each arm separately, situations where the agent has partial control over the extent to which models improve, and more.

References

Peter Auer, Yifang Chen, Pratik Gajane, Chung-Wei Lee, Haipeng Luo, Ronald Ortner, and Chen-Yu Wei. Achieving optimal dynamic regret for non-stationary bandits without prior information. In *Conference on Learning Theory*, pages 159–163. PMLR, 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab Mirrokni. Regret bounds for batched bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7340–7348, 2021.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Sudipto Guha, Kamesh Munagala, and Peng Shi. Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM)*, 58(1):1–50, 2010.

Hoda Heidari, Michael J Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *IJCAI*, pages 1562–1570, 2016.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B Oliva, Jeff Schneider, and Barnabás Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. *Advances in neural information processing systems*, 29, 2016a.

Kirthevasan Kandasamy, Gautam Dasarathy, Barnabas Poczos, and Jeff Schneider. The multi-fidelity multi-armed bandit. *Advances in neural information processing systems*, 29, 2016b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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

Alberto Maria Metelli, Francesco Trovo, Matteo Pirola, and Marcello Restelli. Stochastic rising bandits. In *International Conference on Machine Learning*, pages 15421–15457. PMLR, 2022.

Alessandro Montenegro, Marco Mussi, Francesco Trovò, Marcello Restelli, and Alberto Maria Metelli. Stochastic rising bandits: A best arm identification approach. In *Sixteenth European Workshop on Reinforcement Learning*, 2023. URL `https://openreview.net/forum?id=Ctq0d9LEuT`.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory*, pages 214–228. Springer, 2012.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. In *Conference on Learning Theory*, pages 1456–1456. PMLR, 2015.

Riccardo Poiani, Rémy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. Optimal multi-fidelity best-arm identification. *arXiv preprint arXiv:2406.03033*, 2024.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Tamojeet Roychowdhury, Kota Srinivas Reddy, Krishna P Jagannathan, and Sharayu Moharir. Fixed-confidence best arm identification with decreasing variance. *arXiv preprint arXiv:2502.07199*, 2025.

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

Matilde Tullii, Solenne Gaucher, Nadav Merlis, and Vianney Perchet. Improved algorithms for contextual dynamic pricing. *Advances in Neural Information Processing Systems*, 37:126088–126117, 2024.

Xuchuang Wang, Qingyun Wu, Wei Chen, and John Lui. Multi-fidelity multi-armed bandits revisited. *Advances in Neural Information Processing Systems*, 36:31570–31600, 2023.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on learning theory*, pages 4300–4354. PMLR, 2021.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

Algorithm 3 Follow the Leader with full-information

Input: Time horizon T, arms $\mathcal{K}$, parameter α, retaining times τ, estimates $\widehat{r}_0^\tau(i) = 0,\ \forall i \in \mathcal{K}$

while $t \leq T$ **do**
  Agent picks action $i_t = \operatorname{argmax}_{i \in \mathcal{K}} \widehat{r}_{t-1}^\tau(i)$, breaking ties arbitrarily
  if $t = t_j \in \tau$ **then**
    Set $\lfloor t \rfloor_\tau = \widehat{t}_j$
  end if
  BIB-IA Agent receives reward $r_t^\tau(i_t)$ observes rewards $r_t^\tau(i)\,,\ \forall i \neq i_t \in \mathcal{K}$
  BIB-DB Agent receives reward $r_t^\tau(i_t)$ observes $o_t^\tau(i)\,,\ \forall i \in \mathcal{K}$
  Agent updates estimate $\widehat{r}_t^\tau(i)$
  $t \leftarrow t + 1$
end while

A BIB-IA Model

A.1 BIB-IA with Full-Information

In this section, we analyze a modified version of the BIB-IA model where we assume the agent obtains, at every round, a sample from the distribution of the reward of each arm. In this case, the estimator used is obtained by modifying the one defined in Equation (4) such that it takes into account the additional information received. Specifically, we define the estimator as

$$\widehat{r}_t^\tau(i) = \sum_{p \leq t} \frac{\frac{r_p(i)}{\overline{\sigma^2(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}}}}{\sum_{p \leq t} \frac{1}{\overline{\sigma^2(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}}}}\,. \tag{10}$$

The algorithm proposed in this case is an instance of Follow-the-Leader, that maintains and updates the estimate $\widehat{r}_t^\tau(i)$ for all the arms $i \in \mathcal{K}$ and greedily chooses the biggest one at every round. The details of the routine are included in Algorithm 3.

We present the following result which bounds the regret rates of Algorithm 3. Note that the algorithm exhibits the same dependency in T, as the version with bandit feedback of Section 3, while a different dependency with respect to the number of arms K.

Theorem 5. *Define the function* $\beta(j) = \frac{1-\alpha^j}{1-\alpha}$ *and the quantities*

$$c^* \coloneqq 8\sigma\sqrt{2\log(KT^2)}\left(\frac{T}{M^2}\right)^{\frac{1/2-\beta(M)}{1+\alpha\beta(M)}}$$

and

$$\lambda \coloneqq \left[128\sigma^2\log(KT^2)\frac{T}{M^2}\min\left\{\frac{1}{c^{*2}}, \frac{1}{\overline{c}^2}\right\}\right]^{\frac{1}{(2+\alpha)\beta(M)}}\,.$$

Then, the regret of algorithm Algorithm 3, when using the estimates defined in Equation (10)*, and retraining times* $\widehat{t}_j = \lceil \lambda^{\beta(j)} \rceil,\ j = 1, \ldots, M$*, is bounded by*

$$\mathcal{R}_T \leq 4\left(8\sigma\sqrt{2\log(KT^2)}\right)^{1+\frac{1}{(2+\alpha)\beta(M)}} M \max\left\{T^{\frac{1}{2\beta(M+1)}}, \overline{c}T^{\frac{1}{2+\alpha}}\right\}$$

with probability at least $1 - 1/T$.

Proof. In what follows, we denote by $\mu_t^\tau(i^*)$ the average reward of arm i at time t with respect to retraining times τ. Similarly, $\mu_t^{\tau^*}(i)$ indicates the average reward at time t of arm i when the retraining times are chosen optimally. Recall that, as motivated in Section 2, for this model the optimal times correspond to $\tau^* = \{1, \ldots, M\}$, since they are the ones for which it is possible to benefit for the longest time of the increased average rewards. We start our analysis by considering the following lemma.

Lemma 3. *The regret accumulated by an algorithm interacting with a bandit problem whose rewards follow the structure detailed in Equation* (1) *with respect to some retraining times τ, is bounded by*

$$\mathcal{R}_T \leq \bar{c} M \widehat{t}_M + \sum_{t=1}^{T} \mu(i^*) - \mu(i_t)$$

Hence, we focus on the second term $\sum_{t=1}^{T} \mu(i^*) - \mu(i_t)$. Note that, by adding and subtracting $\widehat{r}_t^{\tau}(i_t)$, it holds

$$\begin{aligned}\sum_{t=1}^{T} \mu(i^*) - \mu(i_t) &= \sum_{t=1}^{T} \mu(i^*) - \widehat{r}_t^{\tau}(i_t) + \widehat{r}_t^{\tau}(i_t) - \mu(i_t) \\ &\leq \sum_{t=1}^{T} \mu(i^*) - \widehat{r}_t^{\tau}(i^*) + \widehat{r}_t^{\tau}(i_t) - \mu(i_t)\,,\end{aligned}$$

where in the last inequality we used the fact that $\widehat{r}_t^{\tau}(i^*) \leq \widehat{r}_t^{\tau}(i_t)$, which holds due to the greedy nature of the algorithm. Let's now define the following *good* event

$$\mathcal{G}_t = \left\{ |\widehat{r}_t^{\tau}(i) - \mu(i)| \leq \sigma \sqrt{\frac{2\log(K/\delta_t)}{\sum_{p \leq t} \frac{1}{(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}}}}, \quad \forall i \in \mathcal{K} \right\}$$

Lemma 4. *The event $\mathcal{G}_t$ happens with probability at least $1 - \delta_t$.*

Considering this event with $\delta_t = 1/T^2$, and applying the bound therein both on i^* and i_t, we have

$$\begin{aligned}&\sum_{t=1}^{T} \mu(i^*) - \mu(i_t) \\ &\qquad \leq \sum_{t=1}^{T} 2\sigma \sqrt{\frac{2\log(KT^2)}{\sum_{p \leq t} \frac{1}{(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}}}} \\ &\qquad \overset{(i)}{=} \sum_{t=1}^{T} 2\sigma \sqrt{\frac{2\log(KT^2)}{\sum_{j=0}^{(\iota(\lfloor t \rfloor_\tau)-1)\vee 0} \frac{\widehat{t}_{j+1} - \widehat{t}_j}{(\widehat{t}_j \vee 1)^{-\alpha}} + \frac{t - \lfloor t \rfloor_\tau}{(\widehat{t}_j \vee 1)^{-\alpha} 2}}} \\ &\qquad \overset{(ii)}{=} \sum_{t=1}^{\widehat{t}_1} \frac{2\sigma\sqrt{2\log(KT^2)}}{\sqrt{t}} + \sum_{i=1}^{M} \sum_{t=\widehat{t}_i+1}^{\widehat{t}_{i+1}} \frac{2\sigma\sqrt{2\log(KT^2)}}{\sqrt{\sum_{j=0}^{i-1} (\widehat{t}_{j+1} - \widehat{t}_j)(\widehat{t}_j^{\alpha} \vee 1) + (t - \widehat{t}_i)\widehat{t}_i^{\alpha}}} \\ &\qquad = 2\sigma\sqrt{2\log(KT^2)}\sqrt{\widehat{t}_1} + \sum_{i=1}^{M} \sum_{s=0}^{\widehat{t}_{i+1} - \widehat{t}_i - 1} \frac{2\sigma\sqrt{2\log(KT^2)}}{\sqrt{\sum_{j=0}^{i-1} (\widehat{t}_{j+1} - \widehat{t}_j)(\widehat{t}_j^{\alpha} \vee 1) + s\widehat{t}_i^{\alpha}}}\,,\end{aligned}$$

with probability at least $1 - 1/T^2$. Note that (i) follows from decomposing the sum in the denominator in the different epochs and observing that during epoch j the value $\lfloor p \rfloor_\tau$ is constant and equal to $\widehat{t}_j$, while (ii) splitting the regret in the one accumulated in each epoch and treating the first epoch separately. The internal sum in the second term of the expression above can be further bounded by using integral approximation as follows

$$\sum_{s=0}^{\widehat{t}_{i+1} - \widehat{t}_i - 1} \frac{1}{\sqrt{\sum_{j=0}^{i-1} (\widehat{t}_{j+1} - \widehat{t}_j)(\widehat{t}_j^{\alpha} \vee 1) + s\widehat{t}_i^{\alpha}}} \leq \int_0^{\widehat{t}_{i+1} - \widehat{t}_i - 1} \frac{1}{\sqrt{\sum_{j=0}^{i-1} (\widehat{t}_{j+1} - \widehat{t}_j)(\widehat{t}_j^{\alpha} \vee 1) + s\widehat{t}_i^{\alpha}}} ds$$

$$\leq \frac{2}{\widehat{t}_i^{\alpha} \vee 1} \sqrt{\sum_{j=0}^{i} (\widehat{t}_{j+1} - \widehat{t}_j)(\widehat{t}_j^{\alpha} \vee 1)} \,.$$

By replacing this in the above bound, we obtain

$$\sum_{t=1}^{T} \mu(i^*) - \mu(i_t) \leq 2\sigma\sqrt{2\log(KT^2)}\sqrt{\widehat{t}_1} + 2\sigma\sqrt{2\log(KT^2)} \sum_{i=1}^{M} \frac{2}{\widehat{t}_i^{\alpha} \vee 1} \sqrt{\sum_{j=0}^{i} (\widehat{t}_{j+1} - \widehat{t}_j)(\widehat{t}_j^{\alpha} \vee 1)}$$

$$\leq 4\sigma\sqrt{2\log(KT^2)} \sum_{i=0}^{M} \sqrt{\frac{\widehat{t}_{i+1}}{\widehat{t}_i^{\alpha} \vee 1}} \,.$$

Replacing this in the initial bound obtained stated in Lemma 3, leads to

$$\mathcal{R}_T \leq \overline{c} M \widehat{t}_M + \sum_{i=0}^{M} 4\sigma\sqrt{2\log(KT^2)} \sqrt{\frac{\widehat{t}_{i+1}}{\widehat{t}_i^{\alpha} \vee 1}}$$

The regret bound is then obtained by replacing the values of $\widehat{t}_i$ as defined in the statement of the theorem, and considering the relation they satisfy highlighted by the following lemma

Lemma 5. *Let $x, u \in \mathbb{R}$ and $n \in \mathbb{N}$, then*

$$x^{\frac{1-u^n}{1-u}} = x \cdot x^{u\frac{1-u^{n-1}}{1-u}} \,.$$

Applying this respectively to λ, α, and i, we found that the values $\widehat{t}_i$, defined the statement satisfy

$$\widehat{t}_{i+1} = \left\lceil \lambda^{\beta(i+1)} \right\rceil = \left\lceil \lambda \cdot \lambda^{\alpha\beta(i)} \right\rceil \leq \lceil \lambda \rceil \widehat{t}_i \,.$$

Thus the regret can be bounded by

$$\mathcal{R}_T \leq \overline{c} M \left(\lambda^{\beta(M)} + 1 \right) + 8\sigma\sqrt{2\log(KT^2)} M \left(\sqrt{\lambda} + 1 \right) + 8\sigma\sqrt{2\log(KT^2)} \sqrt{\frac{T}{\lambda^{\alpha\beta(M)}}} \,, \quad (11)$$

where we have used the fact that $\lambda \leq \lceil \lambda \rceil \leq \lambda + 1$. The final bound is obtained by replacing the value of λ as specified in the statement. Note that, in one case we have $\overline{c} \leq c^*$ and

$$\lambda = \left[128\sigma^2 \log(KT^2) \frac{T}{c^{*2} M^2} \right]^{1/(2+\alpha)\beta(M)} = \left(\frac{T}{M^2} \right)^{1/1+\alpha\beta(M)} \,,$$

which replaced in eq. (11) gives

$$\begin{aligned}
\mathcal{R}_T \leq & c^* M \left(\frac{T}{M^2} \right)^{\frac{\beta(M)}{1+\alpha\beta(M)}} \\
& + 8\sigma\sqrt{2\log(KT^2)} \left[M \left(\frac{T}{M^2} \right)^{\frac{\beta(M)}{2(1+\alpha\beta(M))}} + M^{\frac{\alpha\beta(M)}{1+\alpha\beta(M)}} T^{\frac{1}{2(1+\alpha\beta(M))}} \right] \\
& + c^* M + 8\sigma\sqrt{2\log(KT^2)} M \,,
\end{aligned}$$

further replacing $\overline{c} \leq c^*$, leads to

$$\mathcal{R}_T \leq 32\sigma\sqrt{2\log(KT^2)} M T^{\frac{1}{2(1+\alpha\beta(M))}} \,.$$

On the other end, if $\overline{c} > c^*$, the value of λ, by definition, becomes

$$\lambda = \left[128\sigma^2 \log(KT^2) \frac{T}{(\overline{c} M)^2} \right]^{1/(2+\alpha)\beta(M)} \,,$$

which replaced in eq. (11), gives

$$\begin{aligned}\mathcal{R}_T \leq & \left(8\sigma\sqrt{2\log(KT^2)}\right)^{\frac{2}{2+\alpha}} (\bar{c}M)^{\frac{\alpha}{2+\alpha}} T^{\frac{1}{2+\alpha}} \\ & + \left(8\sigma\sqrt{2\log(KT^2)}\right)^{1+\frac{1}{(2+\alpha)\beta(M)}} M \left(\frac{T}{(\bar{c}M)^2}\right)^{\frac{1}{2(2+\alpha)\beta(M)}} \\ & + \left(8\sigma\sqrt{2\log(KT^2)}\right)^{\frac{2}{2+\alpha}} (\bar{c}M)^{\frac{\alpha}{2+\alpha}} T^{\frac{1}{2+\alpha}} \\ & + \bar{c}M + 8\sigma\sqrt{2\log(KT^2)}\,,\end{aligned}$$

hence

$$\mathcal{R}_T \leq 4\left(8\sigma\sqrt{2\log(KT^2)}\right)^{1+\frac{1}{(2+\alpha)\beta(M)}} M\bar{c}T^{\frac{1}{2+\alpha}}$$

Combining the two results yields

$$\mathcal{R}_T \leq 4\left(8\sigma\sqrt{2\log(KT^2)}\right)^{1+\frac{1}{(2+\alpha)\beta(M)}} M\max\left\{T^{\frac{1}{2(1+\alpha\beta(M))}}, \bar{c}T^{\frac{1}{2+\alpha}}\right\}$$

we finally obtain the desired result by considering an union bound among all the events $\mathcal{G}_t$.

□

A.2 Proofs of Section 3

A.2.1 Proof of Theorem 1

We start by considering again the initial bound on the regret given by Lemma 3 considering the rewards obtained with retraining times τ, and as before we focus on the second term.

Let t be the last time step where arm i was active before being eliminated. Consider the event

$$\tilde{\mathcal{G}} = \bigcap_{t\leq T}\left\{UCB_t^\tau(i^*,\delta) > \max_i LCB_t^\tau(i,\delta)\right\}$$

from, the concentration bound, this intersection holds with probability at least $1 - 1/2T$, where we have replaced $\delta = 1/4KT^{5+\alpha}$, as mentioned in the statement. Thus, under this event, we know that at time t the optimal arm i^* is still active. Then, from the elimination condition of Successive Elimination and the concentration result in Lemma 1 for the estimators, it holds that

$$\mu(i^*) - \mu(i) \leq 2\left(2\sqrt{\frac{\log(1/\delta_t)}{\Sigma_t^\tau(i^*)}} + 2\sqrt{\frac{\log(1/\delta_t)}{\Sigma_t^\tau(i)}}\right),$$

with probability at least $1 - 2tT^\alpha\delta_t$. Note that, from the definition of Algorithm 2, the elimination of arms happens only at the end of loops during which all the still active arms are played in a round-robin fashion. From this follows that every couple of arms still active at any time t has been played the same amount of times. Furthermore, also by definition, the retraining steps happen outside of round-robin routines, hence arms that are still active during one of these loops have an equal variance term in the expression of Σ_t^τ. Thus

$$\Sigma_t^\tau(i^*) = \sum_{p\leq t}\frac{1}{\sigma^2(\lfloor p\rfloor_\tau^{-\alpha}\vee 1)}\mathbb{1}\{i_p = i^*\} = \sum_{p\leq t}\frac{1}{\sigma^2(\lfloor p\rfloor_\tau^{-\alpha}\vee 1)}\mathbb{1}\{i_p = i\} = \Sigma_t^\tau(i)\,,$$

moreover, since arm i is not played again after time t, it holds $\Sigma_t^\tau(i) = \Sigma_T^\tau(i)$. From this, we can define the following event, which holds with probability at least $1 - 2tT^\alpha\delta_t$

$$\mathcal{G}_t^i = \left\{|\mu(i^*) - \mu(i)| \leq 8\sqrt{\frac{\log(1/\delta_t)}{\Sigma_T^\tau(i)}}\right\}.$$

Let $N_T(i) = \sum_{t=1}^{T} \mathbb{1}\{i_t = i\}$, and $\delta_t = 1/4KT^{5+\alpha}$, under the event $\mathcal{G}_t^i \cap \tilde{\mathcal{G}}$ using the regret decomposition lemma, it holds

$$\begin{aligned}
\sum_{t=1}^{T} \mu(i^*) - \mu(i_t) &= \sum_{i \in \mathcal{K}} \Delta_i N_T(i) \\
&\leq \sum_{i \in \mathcal{K}} 8\sqrt{\frac{\log(4KT^{5+\alpha})}{\Sigma_T^\tau(i)}} N_T(i) \\
&\leq 8\sigma\sqrt{\log(4KT^{5+\alpha})} \sum_{i \in \mathcal{K}} \frac{\sum_{p \leq T} \mathbb{1}\{i_p = i\}}{\sqrt{\sum_{p \leq T} \frac{1}{(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}} \mathbb{1}\{i_p = i\}}} \\
&\leq 8\sigma\sqrt{\log(4KT^{5+\alpha})} \sum_{i \in \mathcal{K}} \sum_{j=0}^{M} \frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}}{\sqrt{\sum_{p \leq T} \frac{1}{(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}} \mathbb{1}\{i_p = i\}}}.
\end{aligned}$$

We can further bound this by disregarding all the terms in the denominator but the ones relative to the j-th epoch, where $1/\lfloor p \rfloor_\tau^{-\alpha} = \widehat{t}_j^{\alpha}$, so that the bound becomes

$$\begin{aligned}
\sum_{t=1}^{T} \mu(i^*) - \mu(i_t) &\leq 8\sigma\sqrt{\log(4KT^{5+\alpha})} \sum_{i \in \mathcal{K}} \sum_{j=0}^{M} \frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}}{\sqrt{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} (\widehat{t}_j^{\alpha} \vee 1) \mathbb{1}\{i_p = i\}}} \\
&\leq 8\sigma\sqrt{\log(4KT^{5+\alpha})} \sum_{i \in \mathcal{K}} \sum_{j=0}^{M} \frac{\sqrt{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}}}{\sqrt{\widehat{t}_j^{\alpha} \vee 1}} \\
&\overset{(i)}{\leq} 8\sigma K\sqrt{\log(4KT^{5+\alpha})} \sum_{j=0}^{M} \frac{\sqrt{\sum_{i \in \mathcal{K}} \sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}}}{\sqrt{K(\widehat{t}_j^{\alpha} \vee 1)}} \\
&\leq 8\sigma\sqrt{K\log(4KT^{5+\alpha})} \sum_{j=0}^{M} \sqrt{\frac{\widehat{t}_{j+1}}{\widehat{t}_j^{\alpha} \vee 1}}
\end{aligned}$$

where in (i) we used Jensen's inequality. Replacing this in Lemma 3, and isolating the last term of the sum, for which we replace $\widehat{t}_{M+1} = T$ that again we assume by convention, we obtain

$$\mathcal{R}_T \leq 8\sigma\sqrt{K\log(4KT^{5+\alpha})} \sum_{j=0}^{M-1} \sqrt{\frac{\widehat{t}_{j+1}}{\widehat{t}_j^{\alpha} \vee 1}} + 8\sigma\sqrt{K\log(4KT^{5+\alpha})} \sqrt{\frac{T}{\widehat{t}_M^{\alpha}}} + \overline{c} M \widehat{t}_M$$

Now we use the fact that, from the definition of $\widehat{t}_j$ we have that $\tilde{t}_j \leq \widehat{t}_j \leq \tilde{t}_j + K$, since the number of active arms is always at most K. Furthermore, as made explicit by Lemma 5, it holds $\tilde{t}_j = \lambda \tilde{t}_{j-1}^{\alpha}$ Using the notation $A = 8\sigma\sqrt{K\log(4KT^{5+\alpha})}$, the expression above is

$$\begin{aligned}
\mathcal{R}_T &\leq A \sum_{j=0}^{M-1} \sqrt{\frac{\tilde{t}_{j+1} + K}{\tilde{t}_j^{\alpha} \vee 1}} + A\sqrt{\frac{T}{\tilde{t}_M^{\alpha}}} + \overline{c} M(\tilde{t}_M + K) \\
&\leq AM\left(\sqrt{\lambda} + \sqrt{K}\right) + A\sqrt{\frac{T}{\lambda^{\alpha\beta(M)}}} + \overline{c} M\left(\lambda^{\beta(M)} + K\right)
\end{aligned}$$

$$\leq \underbrace{AM\sqrt{\lambda}}_{(A)} + \underbrace{A\sqrt{\frac{T}{\lambda^{\alpha\beta(M)}}}}_{(B)} + \underbrace{\overline{c}M\lambda^{\beta(M)}}_{(C)} + \underbrace{M\left(A\sqrt{K} + \overline{c}K\right)}_{(D)}$$

To obtain the optimal regret bound we are left with the task of optimizing λ between (A), (B) and (C). Note that, for small values of $\overline{c}$, (C), is negligible, hence the optimal value λ^* will be obtained by optimizing with respect to λ the first two terms, namely

$$AM\sqrt{\lambda^*} = A\sqrt{\frac{T}{(\lambda^*)^{\alpha\beta(M)}}}$$

$$M^2(\lambda^*)^{1+\alpha\beta(M)} = T$$

$$\lambda^* = \left(\frac{T}{M^2}\right)^{\frac{1}{1+\alpha\beta(M)}} .$$

This holds true for values of $\overline{c}$ for which $(A) \geq (C)$, which corresponds to

$$MA\left(\frac{T}{M^2}\right)^{\frac{1/2}{1+\alpha\beta(M)}} \geq \overline{c}M\left(\frac{T}{M^2}\right)^{\frac{\beta(M)}{1+\alpha\beta(M)}} \tag{12}$$

$$c^* \coloneqq A\left(\frac{T}{M^2}\right)^{\frac{1/2-\beta(M)}{1+\alpha\beta(M)}} \geq \overline{c}\,. \tag{13}$$

Hence if $\overline{c} \leq c^*$ it holds

$$\begin{aligned}
\mathcal{R}_T &\leq (A) \ + \ (B) \ + \ (C) \ + \ (D) \\
&\leq 3(A) \ + \ (D) \\
&= 3MA\left(\frac{T}{M^2}\right)^{\frac{1/2}{1+\alpha\beta(M)}} + M\left(A\sqrt{K} + \overline{c}K\right) \\
&\leq 4MKA\left(\frac{T}{M^2}\right)^{\frac{1/2}{1+\alpha\beta(M)}} .
\end{aligned}$$

On the other end, if $\overline{c} \geq c^*$ then the optimal value λ^* is obtained by balancing terms (B) and (C), such that

$$A\sqrt{\frac{T}{(\lambda^*)^{\alpha\beta(M)}}} = \overline{c}M(\lambda^*)^{\beta(M)}$$

$$A^2T = \overline{c}^2M^2(\lambda^*)^{\beta(M)(2+\alpha)}$$

$$\lambda^* = \left(A^2\frac{T}{\overline{c}^2M^2}\right)^{\frac{1}{\beta(M)(2+\alpha)}} .$$

Therefore, the final regret bound becomes

$$\begin{aligned}
\mathcal{R}_T &\leq (A) + (B) + (C) + (D) \\
&\leq (A) + 2(B) + (D) \\
&= AM\left(A^2\frac{T}{\overline{c}^2M^2}\right)^{\frac{1}{2\beta(M)(2+\alpha)}} + 2\overline{c}M\left(A^2\frac{T}{\overline{c}^2M^2}\right)^{\frac{\beta(M)}{\beta(M)(2+\alpha)}} + M\left(A\sqrt{K} + \overline{c}K\right)
\end{aligned}$$

$$\leq 4MKA^{1+\frac{1}{\beta(M)(2+\alpha)}}\overline{c}T^{\frac{1}{2+\alpha}} .$$

Combining the results found, we conclude that the optimal value for λ corresponds to

$$\lambda^* = \min\left\{\left(\frac{T}{M^2}\right)^{\frac{1}{1+\alpha\beta(M)}}, \left(A^2\frac{T}{\overline{c}^2M^2}\right)^{\frac{1}{\beta(M)(2+\alpha)}}\right\}$$

$$= \left[A^2\frac{T}{M^2}\min\left\{\frac{1}{c^{*2}}, \frac{1}{\overline{c}^2}\right\}\right]^{\frac{1}{(2+\alpha)\beta(M)}}$$

and the regret is bounded by

$$\mathcal{R}_T \leq \max\left\{4MKA\left(\frac{T}{M^2}\right)^{\frac{1/2}{1+\alpha\beta(M)}}, 4MKA^{1+\frac{1}{\beta(M)(2+\alpha)}}\overline{c}T^{\frac{1}{2+\alpha}}\right\}$$

A.2.2 Proof of Theorem 2

To prove the result, we adapt the standard notion of a random strategy to our setting. A strategy is a family of functions $\psi = (\psi_t)_{t\leq T}$, that indicates the choices made by the algorithm at every time step. Formally,

$$\psi_t : I_t \to \{1, \dots, K\} \times \{0, \dots, M\}$$

such that the history $I_t = (U_0, r_t^\tau(i_1), \dots, U_t)$, given by the internal randomization of previous rounds and rewards collected, is mapped in a couple (i_{t+1}, n_{t+1}) representing the action and retraining index. The latter is a counter of how many times the model was retrained. To incorporate the fact that this value can only grow, and it increases by at most 1 between consecutive rounds, we limit the set of feasible strategies to the ones for which the second component satisfies

$$\begin{aligned}
&\text{(a)} \quad n_0 = 0 \text{ and } n_T = M \\
&\text{(b)} \quad n_s \leq n_t \quad \forall t > s \\
&\text{(c)} \quad n_{t+1} - n_t \leq 1
\end{aligned}$$

For the sake of generality, we allow this index to be arbitrarily chosen as a function of the past information under the aforementioned constraints. Note that, using this new notation, the set of retraining times τ induced by the strategy ψ, is the set of $\widehat{t}_j$ for which $n_{\widehat{t}_j} = j$ and $n_{\widehat{t}_j} - n_{\widehat{t}_j - 1} = 1$.
As done in the proof for the upper-bounds, the regret relative to the bandit instance ν can be initially decomposed as follows

$$\mathcal{R}_{\psi,\nu,T} = \sum_{t=1}^{T}\mathbb{E}_\nu[\mu_t^{\tau^*}(i^*) - \mu_t^\tau(i^*)] + \sum_{t=1}^{T}\mathbb{E}_\nu\left[\mu(i^*) - \mu(i_t)\right] ,$$

where, for the sake of clarity, we omit the dependency on ψ in the expectation. As done previously, we bound the two terms independently.
We consider a bandit instance in which each increment of the average c_j is equal to $\overline{c}$, therefore by replacing the expression of the optimal retraining times τ^*, it holds that

$$\sum_{t=1}^{T}\mu_t^{\tau^*}(i^*) - \mu_t^\tau(i^*) = \sum_{t=1}^{T}\left(\mu(i^*) + \sum_{j=0}^{t\wedge M}c_j\right) - \left(\mu(i^*) + \sum_{j=0}^{n_t}c_j\right)$$

$$= \sum_{t=1}^{T}\mu(i^*) + \overline{c}(t\wedge M) - (\mu(i^*) + \overline{c}n_t)$$

$$\overline{c}\left(\sum_{t=1}^{M}t - n_t + \sum_{t=M+1}^{M}M - n_t\right)$$

We consider two possible bounds for this quantity. On one hand, because of the constraints we imposed on the strategy, we have $n_t \leq t$, thus the expression above is always bigger or equal to zero. On the other hand, if $\tau \neq \tau^*$, then the first sum is always bigger than 1, while in the second, since $t < \widehat{t}_M$, we have $M - n_t \geq 1$.

$$\sum_{t=1}^{T} \mu_t^{\tau^*}(i^*) - \mu_t^{\tau}(i^*) \geq \overline{c} + \overline{c}(\widehat{t}_M - M - 1) = \overline{c}(\widehat{t}_M - M) \geq \frac{\overline{c}}{2}\widehat{t}_M \,,$$

where the last inequality holds since $\widehat{t}_M \geq M + 1 \geq 2$. Hence, introducing the notation $\mathbb{E}_\nu\left[\widehat{t}_j\right] = \mathbb{t}_j$

$$\sum_{t=1}^{T} \mathbb{E}_\nu[\mu_t^{\tau^*}(i^*) - \mu_t^{\tau}(i^*)] \geq \frac{\overline{c}}{2}\mathbb{E}_\nu\left[\widehat{t}_M\right] = \frac{\overline{c}}{2}\mathbb{t}_M \,.$$

Now we focus on bounding the second term of the regret, corresponding to $\tilde{\mathcal{R}}_{\psi,\nu,T} = \sum_{t=1}^{T} \mathbb{E}_\nu\left[\mu(i^*) - \mu(i_t)\right]$. To this end, we initially recover a bound on the regret halted at the beginning of epoch j, defined as

$$\tilde{\mathcal{R}}_{\psi,\nu,j} = \mathbb{E}_\nu\left[\sum_{t=1}^{\widehat{t}_j} \mu(i^*) - \mu(i_t)\right]$$

and later observe that $\tilde{\mathcal{R}}_{\psi,\nu,j} \leq \tilde{\mathcal{R}}_{\psi,\nu,T}$ for all epochs j. In order to properly define this regret, we consider an equivalent bandit problem modified as follows. Assume the reward obtained by choosing arm i at time t to be

$$r_t^{\tau}(i)\mathbb{1}\{t \leq \widehat{t}_j\} \tag{14}$$

where $r_t^{\tau}(i)$ defined as in Equation (1). Moreover, it is possible to define an accordingly modified arbitrary strategy $\psi^j = (\psi_t^j)_{t \leq T}$ such that $\psi_t^j = \psi_t$ for all $t \leq \widehat{t}_j$ and $\psi_t^j(I'_t) = (1, n_{\widehat{t}_j})$ if $t > \widehat{t}_j$, where I'_t corresponds to the history of the modified model up to time t. Finally denote ν^j the probability measure given by the interaction of strategy ψ^j with the bandit instance whose rewards are defined as in Equation (14). Then the expected regret generated by a suboptimal choice of the arm by the strategy ψ^j, arrested at the start of the j-th epoch is defined as

$$\tilde{\mathcal{R}}_{\psi^j,\nu^j,j} = \mathbb{E}_{\nu^j}\left[\sum_{t=1}^{\widehat{t}_j} \mu(i^*) - \mu(i_t)\right] \tag{15}$$

where again we are omitting the dependency on the strategy in the expectation. Observe that due to a coupling argument, the expectation with respect to the probability measure ν^j is equivalent to the one with respect to ν, therefore it holds

$$\tilde{\mathcal{R}}_{\psi^j,\nu^j,j} = \tilde{\mathcal{R}}_{\psi,\nu,j} \,.$$

Thus, we continue our analysis considering $\tilde{\mathcal{R}}_{\psi^j,\nu^j,j}$. The following lemma states a version of the regret decomposition lemma adapted to our case.

Lemma 6. *Let $\Delta_i := \mu(i^*) - \mu(i)$ the sub-optimality gap of arm i, and $N_i(\widehat{t}_j) = \sum_{t \leq \widehat{t}_j} \sum_{h \in [[0,j-1]]} \mathbb{1}\left\{\psi_{t-1}^j(I'_{t-1}) = (i,h)\right\}$ the number of times arm i was chosen by the strategy ψ^j, then*

$$\tilde{\mathcal{R}}_{\psi^j,\nu^j,j} = \sum_{i \in \mathcal{K}} \Delta_i \mathbb{E}_{\nu^j}\left[N_i(\widehat{t}_j)\right]$$

Consider now the following bandit instances. Let $\nu = (\nu_i)_{i \in \mathcal{K}}$ be a family of gaussian distributions where ν_1 has average Δ and all the others have average 0. Note that, since $\sum_{i \in \mathcal{K}} N_i(\widehat{t}_j) = \widehat{t}_j$ almost surely, by taking the expectation on both sides, we have that $\sum_{i \in \mathcal{K}} \mathbb{E}_{\nu^j}[N_i(\widehat{t}_j)] = \mathbb{t}_j$. This implies that there must be an arm k for which $\mathbb{E}_{\nu^j}[N_i(\widehat{t}_j)] \leq \mathbb{t}_j/K$. From this, we define the instance $\nu' = (\nu'_i)_{i \in \mathcal{K}}$ as the family of distributions for which $\nu'_i = \nu_i, \forall i \neq k$ and ν'_k is a gaussian

with average 2Δ. For both instances the variance is initially σ^2 and then it evolves as detailed by Equation (1). Note that the definition of k might be different for different $\widehat{t}_j$, but without loss of generality, we can consider it to be the same for all intermediate epochs.

Moreover, define ν^j and ν'^j as the modified version respectively of ν and ν', stopped at the start at the j-th epoch, as described above. Then, from Lemma 6, we have that

$$\tilde{\mathcal{R}}_{\psi^j,\nu^j,j} \geq \Delta \left\lfloor \frac{\mathbb{t}_j}{2} \right\rfloor \mathbb{P}_{\nu^j} \left(N_1(\widehat{t}_j) \leq \frac{\mathbb{t}_j}{2} \right)$$

and

$$\tilde{\mathcal{R}}_{\psi^j,\nu'^j,\widehat{t}_j} \geq \Delta \left\lfloor \frac{\mathbb{t}_j}{2} \right\rfloor \mathbb{P}_{\nu'^j} \left(N_1(\widehat{t}_j) > \frac{\mathbb{t}_j}{2} \right).$$

Summing the two equations and using the fact that $\left\lfloor \frac{\mathbb{t}_j}{2} \right\rfloor \geq \frac{\mathbb{t}_j}{4}$ since $\mathbb{t}_j \geq 1$ because, by the definition of the arbitrary strategy $\widehat{t}_j \geq 1$ almost surely, it holds that

$$\max\left\{\tilde{\mathcal{R}}_{\psi^j,\nu^j,j}, \tilde{\mathcal{R}}_{\psi^j,\nu'^j,j}\right\} \geq \frac{\mathbb{t}_j\Delta}{8}\left(\mathbb{P}_{\nu^j}\left(N_1(\widehat{t}_j) \leq \frac{\mathbb{t}_j}{2}\right) + \mathbb{P}_{\nu'^j}\left(N_1(\widehat{t}_j) > \frac{\mathbb{t}_j}{2}\right)\right).$$

where we have used the fact that $a+b \leq 2\max\{a,b\}$. Moreover, using Bretagnolle-Huber inequality on the expression above, we obtain

$$\max\left\{\tilde{\mathcal{R}}_{\psi^j,\nu^j,j}, \tilde{\mathcal{R}}_{\psi^j,\nu'^j,j}\right\} \geq \frac{\mathbb{t}_j\Delta}{16}\exp\left(-KL(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j})\right). \tag{16}$$

To bound the KL divergence above, we make use of the following lemma.

Lemma 7. *Let ν^j and ν'^j the bandit instances defined by halting ν and ν' at epoch j, and $\mathbb{P}_{\nu^j}$ and $\mathbb{P}_{\nu'^j}$ the respective distributions, then*

$$KL\left(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}\right) \leq \mathbb{E}_{\nu^j}[N_k(\widehat{t}_j)]\frac{\Delta^2(\mathbb{t}_{j-1}^{\alpha} \vee 1)}{2\sigma^2}$$

Replacing this in Equation (16) and choosing $\Delta = \sqrt{\frac{2\sigma^2}{\mathbb{E}_{\nu^j}[N_k(\widehat{t}_j)](\mathbb{t}_{j-1}^{\alpha}\vee 1)}}$, leads to

$$\max\left\{\tilde{\mathcal{R}}_{\psi^j,\nu^j,j}, \tilde{\mathcal{R}}_{\psi^j,\nu'^j,j}\right\} \geq \frac{\sigma}{8\sqrt{2}e}\frac{\mathbb{t}_j}{\sqrt{\mathbb{E}_{\nu^j}[N_k(\widehat{t}_j)](\mathbb{t}_{j-1}^{\alpha}\vee 1)}} \geq \frac{\sigma\sqrt{K}}{8\sqrt{2}e}\sqrt{\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}^{\alpha}\vee 1}},$$

where the last inequality is obtained by replacing the assumption $\mathbb{E}_{\nu^j}[N_i(\widehat{t}_j)] \leq \mathbb{t}_j/K$. Finally, recall that $\tilde{\mathcal{R}}_{\psi,\nu,T} \geq \tilde{\mathcal{R}}_{\psi,\nu,\widehat{t}_j}$, $\forall j$, hence the following holds

$$\begin{aligned}
\max_{\nu}\tilde{\mathcal{R}}_{\psi,\nu,T} &\geq \max_{j=1,\dots,M+1}\left\{\max\{\tilde{\mathcal{R}}_{\psi^j,\nu^j,j}, \tilde{\mathcal{R}}_{\psi^j,\nu'^j,j}\}\right\} \\
&\geq \max_{j=1,\dots,M+1}\left\{\frac{\sigma\sqrt{K}}{8\sqrt{2}e}\sqrt{\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}^{\alpha}\vee 1}}\right\} \\
&\geq \frac{\sigma\sqrt{K}}{8\sqrt{2}e(M+1)}\sum_{j=1}^{M+1}\sqrt{\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}^{\alpha}\vee 1}},
\end{aligned}$$

where we are using again the convention $\mathbb{t}_0 = 0$ and $\mathbb{t}_{M+1} = T$. The final bound will be obtained by combining this result with the lower bounds of the first term that correspond either to 0 or $\overline{c}\mathbb{t}_M/2$. Thus, on one end, it holds

$$\max_{\nu}\mathcal{R}_{\psi,\nu,T} \geq \frac{\sigma\sqrt{K}}{8\sqrt{2}e(M+1)}\sum_{j=1}^{M+1}\sqrt{\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}^{\alpha}\vee 1}} + \frac{\overline{c}}{2}\mathbb{t}_M.$$

Keeping only the last term of the sum, we have

$$\max_{\nu} \mathcal{R}_{\psi,\nu,T} \geq \frac{\sigma\sqrt{K}}{8\sqrt{2}e(M+1)} \sqrt{\frac{T}{\mathbb{t}_M^\alpha}} + \frac{\overline{c}}{2}\mathbb{t}_M \,,$$

and optimizing with respect to $\mathbb{t}_M$, leads to

$$\max_{\nu} \mathcal{R}_{\psi,\nu,T} \geq \frac{1}{2}\left(\frac{\sigma\sqrt{K}}{4\sqrt{2}e(M+1)}\right)^{\frac{2}{2+\alpha}} \overline{c}^{\frac{\alpha}{2+\alpha}} T^{\frac{1}{2+\alpha}} \tag{17}$$

On the other end, if we assume the first term to be lower bounded by 0, the complete expression of the regret corresponds to

$$\max_{\nu} \mathcal{R}_{\psi,\nu,T} \geq \frac{\sigma\sqrt{K}}{8\sqrt{2}e(M+1)} \sum_{j=1}^{M+1} \sqrt{\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}^\alpha \vee 1}}$$

which can be lower bounded by optimally choosing $\mathbb{t}_j$. This can be done by solving the following optimization problem, which takes into account the assumptions made for the $\mathbb{t}_j$

$$\begin{aligned} \underset{\mathbb{t}_0,\ldots,\mathbb{t}_{M+1} \in \mathbb{R}}{\text{minimize}} \quad & \sum_{j=1}^{M+1} \sqrt{\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}^\alpha \vee 1}} \\ \text{subject to} \quad & \mathbb{t}_0 = 0, \\ & \mathbb{t}_{M+1} = T, \\ & \mathbb{t}_j \leq \mathbb{t}_{j+1} \ \ \forall j \end{aligned} \tag{18}$$

To solve this, we start by finding a solution of the unconstrained problem, which can be done by finding the values that nullify the gradient of the function $\sum_{j=1}^{M+1} \sqrt{\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}^\alpha \vee 1}}$. Thus note that, the derivative of the sum with respect to $\mathbb{t}_j$ is equal to

$$\frac{1}{2}\frac{1}{\sqrt{\mathbb{t}_j \mathbb{t}_{j-1}^\alpha}} - \frac{\alpha}{2}\sqrt{\mathbb{t}_{j+1}}\mathbb{t}_j^{-\alpha/2-1} = 0$$

$$\sqrt{\frac{\mathbb{t}_{j+1}}{\mathbb{t}_j^\alpha}} = \frac{1}{\alpha}\sqrt{\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}^\alpha}}$$

Iterating this relation tells us that the retraining times that minimize the sum are the ones that satisfy the relation

$$\sqrt{\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}^\alpha}} = \frac{m}{\alpha^j} \,,$$

which can be rewritten as

$$\begin{aligned} \mathbb{t}_{j+1} &= m^2\alpha^{2j}\mathbb{t}_j^\alpha \\ &= m^2\alpha^{2j}(m^2\alpha^{2(j-1)}\mathbb{t}_{j-1}^\alpha)^\alpha \\ &= m^{2+2\alpha}\alpha^{2j+2(j-1)\alpha}\mathbb{t}_j^{\alpha^2} \,, \end{aligned}$$

from which, by iterating, we found

$$\mathbb{t}_{j+1} = m^{2\sum_{p=0}^{j}\alpha^p}\alpha^{2\sum_{p=0}^{j}p\alpha^{j-p}}(\mathbb{t}_0^{\alpha^j} \vee 1)$$

Computing this for $\mathbb{t}_{M+1}$, while imposing $\mathbb{t}_{M+1} = T$ and $\mathbb{t}_0 = 0$, gives

$$T = m^{2\sum_{p=0}^{M}\alpha^p}\alpha^{2\sum_{p=0}^{M}p\alpha^{M-p}}$$

hence

$$m = \left(\frac{T}{\alpha^{2\sum_{p=0}^{M}p\alpha^{M-p}}}\right)^{\frac{1}{2}\frac{1-\alpha}{1-\alpha^{M+1}}} \,.$$

Note now, that the average retraining times $\mathbb{t}_1, \dots, \mathbb{t}_M$ found in this way, with the value of m as above minimize the sum while satisfying $\mathbb{t}_1 < \ldots < \mathbb{t}_M$ and $\mathbb{t}_0 = 0$, $\mathbb{t}_{M+1} = T$, hence they constitute a valid solution of the constrained problem Equation (18). Therefore

$$\max_{\nu} \mathcal{R}_{\psi,\nu,T} \geq \frac{\sigma\sqrt{K}P}{8\sqrt{2}e(M+1)} T^{\frac{1}{2}\frac{1-\alpha}{1-\alpha^{M+1}}} \tag{19}$$

where P corresponds to

$$P = \left(\frac{1}{\alpha^{2\sum_{p=0}^{M} p\alpha^{M-p}}} \right)^{\frac{1}{2}\frac{1-\alpha}{1-\alpha^{M+1}}} \sum_{j=1}^{M+1} \frac{1}{\alpha^j}$$

Combining the bounds obtained in Equation (17) and Equation (19) , we finally recover

$$\max_{\nu} \mathcal{R}_{\psi,\nu,T} \geq \max \left\{ \frac{1}{2} \left(\frac{\sigma\sqrt{K}}{4\sqrt{2}e(M+1)} \right)^{\frac{2}{2+\alpha}} \overline{c}^{\frac{\alpha}{2+\alpha}} T^{\frac{1}{2+\alpha}}, \ \frac{\sigma\sqrt{K}P}{8\sqrt{2}e(M+1)} T^{\frac{1}{2}\frac{1-\alpha}{1-\alpha^{M+1}}} \right\}.$$

A.3 Proofs of Lemmas of Section 3

A.3.1 Proof of Lemma 1

This lemma constitutes a version of the standard Hoeffding-Azuma inequality adapted to our case, and the general scheme of the proof is taken by Lemma 12 of Tullii et al. [2024]. Consider the following sum

$$Z_t(i) = \sum_{p \leq t} \frac{r_p^\tau(i) - \left(\mu(i) + \sum_{j=0}^{\iota(\lfloor p \rfloor_\tau)} c_j\right)}{\sigma^2(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}} \mathbb{1}\{i_p = i\},$$

note that both the functions $\lfloor t \rfloor_\tau$, and $\mathbb{1}\{i_t = i\}$ are $\mathcal{F}_{t-1}$ measurable. Moreover, it holds that $\mu(i) + \sum_{j=0}^{\iota(\lfloor p \rfloor_\tau)} c_j = \mathbb{E}[r_p^\tau(i) \mid \mathcal{F}_{p-1}] \ \forall p$, so that the random variables in $Z_t(i)$ correspond to a martingale difference sequence adapted to $\mathcal{F}_t$. We now consider $\forall x$ the martingale $M_t = \exp(xZ_t(i) - \frac{x^2}{2}\Sigma_t^\tau(i))$ and we argue that this is indeed a super-martingale. To prove it, consider the sequence of random variables

$$\frac{r_t^\tau(i) - \left(\mu(i) + \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} c_j\right)}{\sigma^2(\lfloor t \rfloor_\tau \vee 1)^{-\alpha}} \mathbb{1}\{i_t = i\} = \frac{\xi_t \mathbb{1}\{i_t = i\}}{\sigma^2(\lfloor t \rfloor_\tau \vee 1)^{-\alpha}} \tag{20}$$

since ξ_t is $\sigma\mathfrak{s}(\lfloor t \rfloor_\tau)$ conditionally subgaussian, and by assumption $\mathfrak{s}(\lfloor t \rfloor_\tau) \leq (\lfloor t \rfloor_\tau \vee 1)^{-\alpha/2}$, then the random variable Equation (20) is a $1/\sigma(\lfloor t \rfloor_\tau \vee 1)^{-\alpha/2}$ -subgaussian martingale difference with respect to the filtration $\mathcal{F}_{t-1}$ Thus, it holds

$$\mathbb{E}\left[e^{x \frac{r_t^\tau(i) - \left(\mu(i) + \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} c_j\right)}{\sigma^2(\lfloor t \rfloor_\tau \vee 1)^{-\alpha}} \mathbb{1}\{i_t = i\}} \,\middle|\, \mathcal{F}_{t-1} \right] \leq e^{\frac{x^2}{2} \frac{\mathbb{1}\{i_t = i\}}{\sigma^2(\lfloor t \rfloor_\tau \vee 1)^{-\alpha}}}.$$

Noticing that

$$M_t = M_{t-1} e^{x \frac{r_t^\tau(i) - \left(\mu(i) + \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} c_j\right)}{\sigma^2(\lfloor t \rfloor_\tau \vee 1)^{-\alpha}} \mathbb{1}\{i_t = i\} - \frac{x^2}{2} \frac{\mathbb{1}\{i_t = i\}}{\sigma^2(\lfloor t \rfloor_\tau \vee 1)^{-\alpha}}}$$

we can conclude that M_t is indeed a super-martingale, hence $\mathbb{E}[M_t] \leq \mathbb{E}[M_0] = 1$.

Now for all $\varepsilon > 0$ and $0 < a < b \in \mathbb{R}$, and all $x > 0$

$$\begin{aligned}
\mathbb{P}\left(Z_t(i) \geq \varepsilon \text{ and } \Sigma_t^\tau(i) \in [a,b]\right) &= \mathbb{P}\left(\mathbb{1}\{\Sigma_t^\tau(i) \in [a,b]\} e^{xZ_t(i)} \geq e^{x\varepsilon} \right) \\
&\leq e^{-x\varepsilon} \mathbb{E}\left[\mathbb{1}\{\Sigma_t^\tau(i) \in [a,b]\} e^{xZ_t(i)} \right] \\
&\leq e^{-x\varepsilon + \frac{x^2}{2}b} \mathbb{E}\left[\mathbb{1}\{\Sigma_t^\tau(i) \in [a,b]\} e^{xZ_t(i) - \frac{x^2}{2}b} \right].
\end{aligned}$$

We can bound the expectation above as follows

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\{\Sigma_t^\tau(i)\in[a,b]\}e^{xZ_t(i)-\frac{x^2}{2}b}\right] &\leq \mathbb{E}\left[\mathbb{1}\{\Sigma_t^\tau(i)\in[a,b]\}e^{xZ_t(i)-\frac{x^2}{2}\Sigma_t^\tau(i)}\right]\\ &\leq \mathbb{E}\left[e^{xZ_t(i)-\frac{x^2}{2}\Sigma_t^\tau(i)}\right]\\ &= \mathbb{E}[M_t]\leq 1\,,\end{aligned}$$

from which we recover

$$\mathbb{P}(Z_t(i)\leq\varepsilon \text{ and } \Sigma_t^\tau(i)\in[a,b])\leq e^{-x\varepsilon+\frac{x^2}{2}b}\,.$$

Since this holds for every $x>0$, we choose the value that minimizes the right-hand term, namely $x=\frac{\varepsilon}{b}$. Furthermore choosing $\varepsilon=\sqrt{2b\log(1/\delta_t)}$, the expression above becomes

$$\mathbb{P}\left(Z_t(i)\geq\sqrt{2b\log(1/\delta_t)} \text{ and } \Sigma_t^\tau(i)\in[a,b]\right)\leq\delta_t\,.$$

Observe that

$$\Sigma_t^\tau(i)=\sum_{p\leq t}\frac{\lfloor p\rfloor_\tau^\alpha}{\sigma^2}\mathbb{1}\{i_p=i\}\leq\frac{tT^\alpha}{\sigma^2}\,,$$

and, from its definition, if $\Sigma_t^\tau(i)>0$, it holds $\Sigma_t^\tau(i)\geq 1/\sigma^2$, $\forall t$. Then for every t, when it is positive, we know $\Sigma_t^\tau(i)$ lives in the interval $\left[\frac{1}{\sigma^2},\frac{tT^\alpha}{\sigma^2}\right]$. Therefore we can recover the probability for the generic bound with $\Sigma_t^\tau(i)>0$, by considering an union bound over smaller intervals as follows

$$\begin{aligned}&\mathbb{P}\Big(Z_t(i)\geq 2\sqrt{\Sigma_t^\tau(i)\log(1/\delta_t)} \text{ and } \Sigma_t^\tau(i)>0\Big)\\ &\quad=\mathbb{P}\left(Z_t(i)\geq 2\sqrt{\Sigma_t^\tau(i)\log(1/\delta_t)} \text{ and } \Sigma_t^\tau(i)\geq\frac{1}{\sigma^2}\right)\\ &\quad\leq\sum_{k=1}^{tT^\alpha-1}\mathbb{P}\left(Z_t(i)\geq 2\sqrt{\Sigma_t^\tau(i)\log(1/\delta_t)} \text{ and } \Sigma_t^\tau(i)\in\left[\frac{k}{\sigma^2},\frac{k+1}{\sigma^2}\right]\right)\\ &\quad\leq\sum_{k=1}^{tT^\alpha-1}\mathbb{P}\left(Z_t(i)\geq\sqrt{4(k/\sigma^2)\log(1/\delta_t)} \text{ and } \Sigma_t^\tau(i)\in\left[\frac{k}{\sigma^2},\frac{k+1}{\sigma^2}\right]\right)\\ &\quad\leq\sum_{k=1}^{tT^\alpha-1}\mathbb{P}\left(Z_t(i)\geq\sqrt{2(k+1/\sigma^2)\log(1/\delta_t)} \text{ and } \Sigma_t^\tau(i)\in\left[\frac{k}{\sigma^2},\frac{k+1}{\sigma^2}\right]\right)\\ &\quad\leq\sum_{k=1}^{tT^\alpha-1}\delta_t\leq T^\alpha t\delta_t\end{aligned}$$

where we have used the fact that $2(k/\sigma^2)\geq k+1/\sigma^2$.

With the same argument, we can prove also that

$$\mathbb{P}\left(-Z_t(i)\geq 2\sqrt{\Sigma_t^\tau(i)\log(1/\delta_t)} \text{ and } \Sigma_t^\tau(i)>0\right)\leq T^\alpha t\delta_t\,.$$

We can now conclude by normalizing by Σ_t^τ and observing that $Z_t(i)/\Sigma_t^\tau=\widehat{r}_t^\tau(i)-\mu(i)$.

A.3.2 Proof of Lemma 3

Starting from the definition of the regret, as in Equation (3), by adding subtracting $\mu^{\tau}(i^*)$, it holds

$$\begin{aligned}\mathcal{R}_T &= \sum_{t=1}^{T} \mu_t^{\tau^*}(i^*) - \mu_t^{\tau}(i_t) \\ &= \sum_{t=1}^{T} \mu_t^{\tau^*}(i^*) - \mu_t^{\tau}(i^*) + \mu_t^{\tau}(i^*) - \mu_t^{\tau}(i_t) \\ &= \underbrace{\sum_{t=1}^{T} \mu_t^{\tau^*}(i^*) - \mu_t^{\tau}(i^*)}_{(A)} + \underbrace{\sum_{t=1}^{T} \mu_t^{\tau}(i^*) - \mu_t^{\tau}(i_t)}_{(B)} .\end{aligned}$$

Note that, the two terms A and B above represent respectively the regret the algorithm incurs by choosing incorrectly the retraining instants τ, and the one caused by a suboptimal choice of the arm to be played. We proceed to bound the two quantities independently. In the first case, we replace the explicit values the optimal times in τ^*, that, as explained in the introduction, correspond to $\{1, \dots, M\}$

$$\begin{aligned}\text{(A)} &= \sum_{t=1}^{T} \left(\mu(i^*) + \sum_{j=0}^{t \wedge M} c_j \right) - \left(\mu(i^*) + \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau} c_j \right) \\ &= \sum_{t=1}^{\widehat{t}_M} \left(\sum_{j=0}^{t \wedge M} c_j - \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} c_j \right) \\ &\overset{(i)}{=} \sum_{t=1}^{\widehat{t}_M} \sum_{j=\iota(\lfloor t \rfloor_\tau)}^{t \wedge M} c_j \\ &\leq \overline{c} \sum_{t=1}^{\widehat{t}_M} M - \iota(\lfloor t \rfloor_\tau) \\ &\leq \overline{c} M \widehat{t}_M\end{aligned}$$

where (i) is well defined as $\iota(\lfloor t \rfloor_\tau) \leq t \wedge M, \forall t$.

For the second term, since the retraining times τ are the same, it is sufficient to replace the explicit definitions of $\mu_t^{\tau}(i^*)$ and $\mu_t^{\tau}(i_t)$, such that

$$\text{(B)} = \sum_{t=1}^{T} \left(\mu(i^*) + \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} c_j \right) - \left(\mu(i_t) + \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} c_j \right) = \sum_{t=1}^{T} \mu(i^*) - \mu(i_t) .$$

Considering both the bounds for A and B yields the result.

A.3.3 Proof of Lemma 4

Proving Lemma 4 is equivalent to showing that $|\widehat{r}_t(i) - \mu(i)|$ is a subgaussian random variable. To this end, consider the following lemma

Lemma 8 ([Lattimore and Szepesvári, 2020, Lemma 5.4])**.** *Suppose that X_1 and X_2 are independent σ_1 and σ_2-subgaussian random variables, respectively then:*

(a) γX_1 is $|\gamma|\sigma$-subgaussian for all $\gamma \in \mathbb{R}$

(b) $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-subgaussian

Note that the estimate $\widehat{r}_t(i)$, defined as in eq. (10) is a convex combination of the reward of arm i at each time step, where the weights are

$$\frac{\frac{1}{\sigma^2(\lfloor t \rfloor_\tau \vee 1)^{-\alpha}}}{\sum_{p \leq t} \frac{1}{\sigma^2(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}}}.$$

Hence, an application of Lemma 8 shows that $\widehat{r}_t(i)$ is a $\tilde{\sigma}$-subgaussian random variable for every $i \in \mathcal{K}$, for

$$\tilde{\sigma} = \sqrt{\frac{1}{\sum_{p \leq t} \frac{1}{\sigma^2(\lfloor p \rfloor_\tau \vee 1)^{-\alpha}}}}.$$

A.3.4 Proof of Lemma 5

The proof follows from a straightforward algebraic manipulation as follows

$$x^{\frac{1-u^n}{1-u}} = x^{\frac{1-u+u-u^n}{1-u}} = x^{1+\frac{u-u^n}{1-u}} = x \cdot x^{u\frac{1-u^{n-1}}{1-u}}$$

A.3.5 Proof of Lemma 6

Let $\tilde{\mathcal{R}}_{\psi^j,\nu^j,j}$ be the regret interrupted at epoch j with respect to the modified instance and strategy ν^j and ψ^j. It is possible to rewrite it as follows

$$\begin{aligned}
\tilde{\mathcal{R}}_{\psi^j,\nu^j,j} &= \mathbb{E}_{\nu^j}\left[\sum_{t=1}^{\widehat{t}_j} \mu(i^*) - \mu(i_t)\right] \\
&= \mathbb{E}_{\nu^j}\left[\sum_{i\in\mathcal{K}}\sum_{h=1}^{j-1}\left(\sum_{t=1}^{\widehat{t}_j} \mu(i^*) - \mu(i_t)\right)\mathbb{1}\left\{\psi_{t-1}^j(I'_{t-1}) = (i,h)\right\}\right] \\
&= \mathbb{E}_{\nu^j}\left[\sum_{i\in\mathcal{K}}\sum_{h=1}^{j-1}\left(\sum_{t=1}^{\widehat{t}_j} \mu(i^*) - \mu(i)\right)\mathbb{1}\left\{\psi_{t-1}^j(I'_{t-1}) = (i,h)\right\}\right] \\
&= \sum_{i\in\mathcal{K}}(\mu(i^*) - \mu(i))\,\mathbb{E}_{\nu^j}\left[\sum_{t=1}^{\widehat{t}_j}\sum_{h=1}^{j-1}\mathbb{1}\left\{\psi_{t-1}^j(I'_{t-1}) = (i,h)\right\}\right] \\
&= \sum_{i\in\mathcal{K}}\Delta_i\mathbb{E}_{\nu^j}[N_i(\widehat{t}_j)]
\end{aligned}$$

where the last inequality is obtained by replacing the definitions of Δ_i and $N_i(t_j)$.

A.3.6 Proof of Lemma 7

This proof is adapted from Lemma 15.1 of Lattimore and Szepesvári [2020]. We start by noticing that,

$$KL(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}) = \mathbb{E}_\nu\left[\log\left(\frac{\mathrm{d}\mathbb{P}_{\nu^j}}{\mathrm{d}\mathbb{P}_{\nu'^j}}\right)\right]$$

For every h, we denote by P_i^h and $P_i^{'h}$ the conditional distributions of arm i, conditioned on $\widehat{t}_h$ for the bandit instances ν^j and ν'^j respectively. Furthermore, we denote by f_i^h and $f_i^{'h}$ the respective densities. By consequence, it is possible to define the measure $\lambda^{j-1} = \sum_{i\in K}\sum_{h=0}^{j-1} P_i^h + P_i^{'h}$, such that $\lambda^{j-1}(A) = \sum_{i\in K}\sum_{h=0}^{j-1} P_i^h(A) + P_i^{'h}(A)$ for any measurable set A. Let now ρ^{j-1} be the counting measure on $\mathcal{K} \times [[0, j-1]]$, then from classical results on the distribution of the canonical bandit model, it follows that the Radon-Nikodym derivatives of $\mathbb{P}_{\nu^j}$ with respect to $\lambda^{j-1} \times \rho^{j-1}$ can be written as

$$\frac{\mathrm{d}\mathbb{P}_{\nu^j}}{\mathrm{d}(\lambda^{j-1}\times\rho^{j-1})}\left((A_1,N_1), r_1^\tau(A_1), \ldots, (A_{\widehat{t}_j}, N_{\widehat{t}_j}), r_{\widehat{t}_j}^\tau(A_{\widehat{t}_j})\right) = \prod_{t=1}^{\widehat{t}_j}\mathbb{P}((A_t,N_t)|I'_{t-1})f_{A_t}^{N_t}(r_t^\tau(A_t)),$$

where I'_{t-1} represents the history up to time $t-1$. An analogous result holds for $\mathbb{P}_{\nu'^j}$. So that, we can rewrite the KL divergence as

$$\begin{aligned}
KL(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j} &= \mathbb{E}_{\nu^j}\left[\log\left(\frac{\mathrm{d}\mathbb{P}_{\nu^j}}{\mathrm{d}\mathbb{P}_{\nu'^j}}\right)\right] \\
&= \mathbb{E}_{\nu^j}\left[\mathbb{E}_{\nu^j}\left[\log\left(\frac{\mathrm{d}\mathbb{P}_{\nu^j}}{\mathrm{d}\mathbb{P}_{\nu'^j}}\right)\middle|(\widehat{t}_h)_{h=1}^{j-1}\right]\right] \\
&\overset{(i)}{=} \mathbb{E}_{\nu^j}\left[\mathbb{E}_{\nu^j}\left[\sum_{t=1}^{\widehat{t}_j}\log\left(\frac{f_{A_t}^{N_t}(r_t^\tau(A_t))}{f_{A_t}^{'N_t}(r_t^\tau(A_t))}\right)\middle|(\widehat{t}_h)_{h=1}^{j-1}\right]\right] \\
&= \mathbb{E}_{\nu^j}\left[\sum_{t=1}^{\widehat{t}_j}\mathbb{E}_{\nu^j}\left[\log\left(\frac{f_{A_t}^{N_t}(r_t^\tau(A_t))}{f_{A_t}^{'N_t}(r_t^\tau(A_t))}\right)\middle|(\widehat{t}_h)_{h=1}^{j-1}\right]\right]
\end{aligned}$$

where in (i) we have used the chain rule and the explicit definition of the distribution conditioned on $(\widehat{t}_h)_{h=1}^{j-1}$ made explicit above. We focus now on the internal expectation. We begin by observing the following

$$\begin{aligned}
\mathbb{E}_{\nu^j}\left[\log\left(\frac{f_{A_t}^{N_t}(r_t^\tau(A_t))}{f_{A_t}^{'N_t}(r_t^\tau(A_t))}\right)\middle|(\widehat{t}_h)_{h=1}^{j-1}\right] &= \mathbb{E}_{\nu^j}\left[\mathbb{E}_{\nu^j}\left[\log\left(\frac{f_{A_t}^{N_t}(r_t^\tau(A_t))}{f_{A_t}^{'N_t}(r_t^\tau(A_t))}\right)\middle|(A_t,N_t)\right]\middle|(\widehat{t}_h)_{h=1}^{j-1}\right] \\
&= KL\left(P_{A_t}^{N_t}, P_{A_t}^{'N_t}\right),
\end{aligned}$$

where we have used the fact that $\mathbb{P}_{\nu^j}(\cdot|(A_t,N_t),(\widehat{t}_h)_{h=1}^{j-1})$ is $\mathrm{d}P_{A_t}^{N_t} = f_{A_t}^{N_t}\mathrm{d}\lambda^{j-1}$, from which

$$\begin{aligned}
\sum_{t=1}^{\widehat{t}_j}\mathbb{E}_{\nu^j}\left[\log\left(\frac{f_{A_t}^{N_t}(r_t^\tau(A_t))}{f_{A_t}^{'N_t}(r_t^\tau(A_t))}\right)\middle|(\widehat{t}_h)_{h=1}^{j-1}\right] &= \sum_{t=1}^{\widehat{t}_j} KL\left(P_{A_t}^{N_t}, P_{A_t}^{'N_t}\right) \\
&= \sum_{t=1}^{\widehat{t}_j}\sum_{i\in\mathcal{K}}\sum_{h=0}^{j-1}\mathbb{1}\left\{(A_t,N_t)=(i,h)\right\} KL\left(P_{A_t}^{N_t}, P_{A_t}^{'N_t}\right) \\
&= \sum_{t=1}^{\widehat{t}_j}\sum_{i\in\mathcal{K}}\sum_{h=0}^{j-1}\mathbb{1}\left\{(A_t,N_t)=(i,h)\right\} KL\left(P_i^h, P_i^{'h}\right)
\end{aligned}$$

Furthermore, note that

$$\begin{aligned}
KL\left(P_i^h, P_i^{'h}\right) &= \mathbb{E}_{\nu^j}\left[\log\frac{f_i^h(X_t)}{f_i^{'h}(X_t)}\middle|(\widehat{t}_h)_{h=0}^{j-1}\right] \\
&\overset{(i)}{=} \mathbb{E}_{\nu^j}\left[\log\frac{f_i^h(X_t)}{f_i^{'h}(X_t)}\middle|\widehat{t}_h\right] \\
&\overset{(ii)}{=} \frac{(\mu(i)-\mu'(i))^2}{2\sigma^2\mathbb{s}(\widehat{t}_h)^2} \\
&\leq \frac{(\mu(i)-\mu'(i))^2(\widehat{t}_h^{\alpha}\vee 1)}{2\sigma^2}
\end{aligned}$$

where (i) follows from the fact that during epoch h, the densities f_i^h and $f_i^{'h}$ only depend on the retraining time $\widehat{t}_h$, while (ii) follows from the explicit expression of the KL divergence between two

gaussian distributions with averages $\mu(i)$ and $\mu'(i)$ and variance $\sigma^2 \mathbb{s}(\widehat{t}_h)^2$, that are the distributions we are considering thanks to the definitions of the two instances. And again, due to the definition of ν^j and ν'^j, we note that the difference $\mu(i) - \mu'(i)$ is equal to zero for all arms but arm k for which has value Δ. Therefore, replacing this in the expression above and taking the expectation on both sides, we have

$$
\begin{aligned}
KL(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}) &= \mathbb{E}_{\nu} \left[\sum_{t=1}^{\widehat{t}_j} \mathbb{E}_{\nu^j} \left[\log \left(\frac{f_{A_t}^{N_t}(X_t)}{f_{A_t}'^{N_t}(X_t)} \right) \middle| (\widehat{t}_h)_{h=1}^{j-1} \right] \right] \\
&= \mathbb{E}_{\nu^j} \left[\sum_{t=1}^{\widehat{t}_j} \sum_{h=0}^{j-1} \mathbb{1}\left\{ (A_t, N_t) = (k, h) \right\} KL\left(P_k^h, P_k'^h \right) \right] \\
&\leq \mathbb{E}_{\nu^j} \left[\sum_{t=1}^{\widehat{t}_j} \sum_{h=0}^{j-1} \mathbb{1}\left\{ (A_t, N_t) = (k, h) \right\} \frac{\Delta^2 (\widehat{t}_h^{\alpha} \vee 1)}{2\sigma^2} \right] \\
&\overset{(i)}{\leq} \mathbb{E}_{\nu^j} \left[\sum_{t=1}^{\widehat{t}_j} \sum_{h=0}^{j-1} \mathbb{1}\left\{ (A_t, N_t) = (k, h) \right\} \frac{\Delta^2 (\widehat{t}_{j-1}^{\alpha} \vee 1)}{2\sigma^2} \right] \\
&\overset{(ii)}{=} \mathbb{E}_{\nu^j} \left[\sum_{t=1}^{\widehat{t}_j} \sum_{h=0}^{j-1} \mathbb{1}\left\{ (A_t, N_t) = (k, h) \right\} \right] \mathbb{E}_{\nu^j} \left[\frac{\Delta^2 (\widehat{t}_{j-1}^{\alpha} \vee 1)}{2\sigma^2} \right]
\end{aligned}
$$

where, (i) follows from the fact that $\widehat{t}_h \leq \widehat{t}_{j-1}$ almost surely, while (ii) from an application of Wald's lemma. Lastly, using the definition of $N_k(\widehat{t}_j)$ and the fact that $\mathbb{E}_{\nu^j}[\widehat{t}_{j-1}^{\alpha}] \leq \mathbb{t}_{j-1}^{\alpha}$ because of Jensen's inequality since $\alpha \leq 1$, we obtain

$$
KL(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j})) = \mathbb{E}_{\nu^j} \left[N_k(\widehat{t}_j) \right] \frac{\Delta^2 (\mathbb{t}_{j-1}^{\alpha} \vee 1)}{2\sigma^2}
$$

B BIB-DB Model

B.1 BIB-DB with Full Information

In this section, we study the full-information version of the BIB-DB model, where the agent obtains a reward drawn from the distribution of the arm chosen at every round, but observes a noisy sample from all of them. The estimators for the averages used in this case again is defined through a two-step process. Let $\tau = \{\widehat{t}_1, \dots, \widehat{t}_M\}$ be the retraining times, initially let

$$\widehat{r}_t^j(i) = \frac{1}{N_j^\tau} \sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}\wedge t} o_p^\tau(i)$$

where $N_j^\tau = (\widehat{t}_{j+1} \wedge t) - \widehat{t}_j$. This corresponds to the average of the rewards obtained during the epoch in which the model has been retrained exactly j times, then the final estimator is obtained as

$$\widehat{r}_t^\tau(i) = \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\frac{\widehat{r}_t^j(i)}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau}}{\sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{1}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau}} \tag{21}$$

In this case, there is no need to arrest the updating of the estimator, due to a different concentration bound that grants a good behavior of the estimator with respect to its average.

Algorithm 3, as before an *ad hoc* version of Follow-the-Leader, contains the details of the algorithm used, which relies on the estimates Equation (21).

Thus, we have the following result. Note that, once again the dependency with respect to T is the same as in the bandit-feedback case.

Theorem 6. *The regret of Algorithm 3, that picks at every round the biggest estimate defined as in Equation* (21)*, is bounded by*

$$\mathcal{R}_T \leq 4M\sqrt{2\log(KT^2)} \left(\sigma T^{\frac{1-\alpha/2}{1-(\alpha/2)^{M+1}}} + \phi\sqrt{T} \right)$$

with probability $1 - 1/T$*, when the retraining times are picked as* $\widehat{t}_j = \lceil \lambda^{\kappa(j)} \rceil$*,* $\forall j = 1, \dots, M$*, with* $\kappa(j) = \frac{1-(\alpha/2)^j}{1-\alpha/2}$ *and*

$$\lambda = T^{\frac{1-\alpha/2}{1-(\alpha/2)^{M+1}}} . \tag{22}$$

Proof. Note that, as stated in Section 2, since both the noise at time t and the bias are centered random variables, the average reward and the optimal strategy are independent of the choice of the retraining times. Adding and subtracting $\widehat{r}_t^\tau(i_t)$ in the explicit expression of the regret gives

$$\begin{aligned}
\mathcal{R}_T &= \sum_{t=1}^T \mu(i^*) - \mu(i_t) \\
&= \sum_{t=1}^T \mu(i^*) - \widehat{r}_t^\tau(i_t) + \widehat{r}_t^\tau(i_t) - \mu(i_t) \\
&\leq \sum_{t=1}^T \mu(i^*) - \widehat{r}_t^\tau(i^*) + \widehat{r}_t^\tau(i_t) - \mu(i_t)
\end{aligned}$$

where the last inequality follows from the selection rule of the algorithm i.e. $\widehat{r}_t^\tau(i^*) \leq \widehat{r}_t^\tau(i)$. Let the event $\mathcal{G}_t$ be

$$\mathcal{G}_t = \left\{ \left| \widehat{r}_t^\tau(i) - \mu(i) \right| \leq \sqrt{\frac{2\log(K/\delta_t)}{\sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{1}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau}}}, \quad \forall i \in \mathcal{K} \right\} .$$

Lemma 9. *The event* $\mathcal{G}_t$ *holds with probability at least* $1 - \delta_t$*.*

Under this event, and considering $\delta_t = 1/T^2$, we have

$$\mathcal{R}_T \leq 2\sum_{t=1}^{T}\sqrt{\frac{2\log(KT^2)}{\sum_{j=0}^{\iota(\lfloor t\rfloor_\tau)}\frac{1}{\sigma^2(\widehat{t}_j\vee 1)^{-\alpha}+\phi^2/N_j^\tau}}},$$

which follows from the concentration bound applied to i^* and i_t. From this, replacing the explicit expressions of $\mathbb{s}$ and N_j^τ, we have

$$\mathcal{R}_T \leq 2\sqrt{2\log(KT^2)}\left(\sum_{t=1}^{\widehat{t}_1}\sqrt{\frac{1}{\frac{1}{\sigma^2+\frac{\phi^2}{t}}}} + \underbrace{\sum_{i=1}^{M}\sum_{t=\widehat{t}_i+1}^{\widehat{t}_{i+1}}\sqrt{\frac{1}{\sum_{j=0}^{i-1}\frac{1}{\sigma^2(\widehat{t}_j\vee 1)^{-\alpha}+\frac{\phi^2}{\widehat{t}_{j+1}-\widehat{t}_j}}+\frac{1}{\sigma^2\widehat{t}_i^{-\alpha}+\frac{\phi^2}{t-\widehat{t}_i}}}}}_{(\star)}\right),$$

where, again we have isolated the regret corresponding to the first epoch that we bound separately. We can rewrite $(\star)$ as

$$\begin{aligned}
(\star) &= \sum_{i=1}^{M}\sum_{t=\widehat{t}_i+1}^{\widehat{t}_{i+1}}\sqrt{\frac{1}{\sum_{j=0}^{i-1}\frac{(\widehat{t}_j^\alpha\vee 1)(\widehat{t}_{j+1}-\widehat{t}_j)}{\sigma^2(\widehat{t}_{j+1}-\widehat{t}_j)+\phi^2(\widehat{t}_i^\alpha\vee 1)}+\frac{(\widehat{t}_i^\alpha\vee 1)(t-\widehat{t}_i)}{\sigma^2(t-\widehat{t}_i)+\phi^2(\widehat{t}_i^\alpha\vee 1)}}} \\
&\overset{(i)}{\leq} \sum_{i=1}^{M}\sum_{t=\widehat{t}_i+1}^{\widehat{t}_{i+1}}\sqrt{\frac{1}{\sum_{j=0}^{i-1}\frac{(\widehat{t}_j^\alpha\vee 1)(\widehat{t}_{j+1}-\widehat{t}_j)}{\sigma^2\widehat{t}_{j+1}+\phi^2(\widehat{t}_j^\alpha\vee 1)}+\frac{(\widehat{t}_i^\alpha\vee 1)(t-\widehat{t}_i)}{\sigma^2 t+\phi^2(\widehat{t}_i^\alpha\vee 1)}}} \\
&\overset{(ii)}{\leq} \sum_{i=1}^{M}\sum_{t=\widehat{t}_i+1}^{\widehat{t}_{i+1}}\sqrt{\frac{\sigma^2\widehat{t}_{i+1}+\phi^2\widehat{t}_i^\alpha}{\sum_{j=0}^{i-1}(\widehat{t}_j^\alpha\vee 1)(\widehat{t}_{j+1}-\widehat{t}_j)+\widehat{t}_i^\alpha(t-\widehat{t}_i)}} \\
&\leq \sum_{i=1}^{M}\sqrt{\sigma^2\widehat{t}_{i+1}+\phi^2\widehat{t}_i^\alpha}\sum_{t=\widehat{t}_i+1}^{\widehat{t}_{i+1}}\sqrt{\frac{1}{\sum_{j=0}^{i-1}(\widehat{t}_j^\alpha\vee 1)(\widehat{t}_{j+1}-\widehat{t}_j)+\widehat{t}_i^\alpha(t-\widehat{t}_i)}}
\end{aligned}$$

where (i) and (ii) are obtained by first removing the negative term in the denominator, then replacing it with the biggest term corresponding to $\sigma^2\widehat{t}_{i+1}+\phi^2\widehat{t}_i^\alpha$. Now the most internal sum in the term above can be bounded again using integral approximation as done in the proof of Theorem 5, such that

$$\begin{aligned}
\sum_{t=\widehat{t}_i+1}^{\widehat{t}_{i+1}}\sqrt{\frac{1}{\sum_{j=0}^{i-1}(\widehat{t}_j^\alpha\vee 1)(\widehat{t}_{j+1}-\widehat{t}_j)+\widehat{t}_i^\alpha(t-\widehat{t}_i)}} &\leq \sum_{s=0}^{\widehat{t}_{i+1}-\widehat{t}_i-1}\frac{1}{\sqrt{\sum_{j=0}^{i-1}(\widehat{t}_{j+1}-\widehat{t}_j)(\widehat{t}_j^\alpha\vee 1)+s\widehat{t}_i^\alpha}} \\
&\leq \int_0^{\widehat{t}_{i+1}-\widehat{t}_i-1}\frac{1}{\sqrt{\sum_{j=0}^{i-1}(\widehat{t}_{j+1}-\widehat{t}_j)(\widehat{t}_j^\alpha\vee 1)+s\widehat{t}_i^\alpha}}ds \\
&\leq \frac{2}{\widehat{t}_i^\alpha}\sqrt{\sum_{j=0}^{i}(\widehat{t}_{j+1}-\widehat{t}_j)(\widehat{t}_j^\alpha\vee 1)} \\
&\leq 2\sqrt{\frac{\widehat{t}_{i+1}}{\widehat{t}_i^\alpha}}.
\end{aligned}$$

Therefore

$$(\star) \leq \sum_{i=1}^{M} \sqrt{\sigma^2 \widehat{t}_{i+1} + \phi^2 \widehat{t}_i^{\alpha}} \left(2\sqrt{\frac{\widehat{t}_{i+1}}{\widehat{t}_i^{\alpha}}} \right) \leq \sum_{i=1}^{M} \sigma \frac{\widehat{t}_{i+1}}{\sqrt{\widehat{t}_i^{\alpha}}} + \phi\sqrt{\widehat{t}_{i+1}}\,,$$

and using an analogous argument, we can prove that

$$\sum_{t=1}^{\widehat{t}_1} \sqrt{\frac{1}{\frac{1}{\sigma^2 + \frac{\phi^2}{t}}}} \leq \sigma \widehat{t}_1 + \phi\sqrt{\widehat{t}_1}\,.$$

such that the complete bound on the regret becomes

$$\mathcal{R}_T \leq 2\sqrt{2\log(KT^2)} \sum_{i=0}^{M} \sigma \frac{\widehat{t}_{i+1}}{\sqrt{\widehat{t}_i^{\alpha} \vee 1}} + \phi\sqrt{\widehat{t}_{i+1}}\,.$$

We now replace the values of $\widehat{t}_i$, as specified in the statement.

$$\widehat{t}_{i+1} = \left\lceil \lambda^{\kappa(i+1)} \right\rceil = \left\lceil \lambda \cdot \lambda^{\frac{\alpha}{2}\kappa(i)} \right\rceil \leq \lceil \lambda \rceil \widehat{t}_i\,.$$

where we have used Lemma 5. Replacing these values in the expression of the regret, together with $\widehat{t}_0 = 0$ and $\widehat{t}_{M+1} = T$, we obtain

$$\mathcal{R}_T \leq 2\sqrt{2\log(KT^2)} \sum_{i=0}^{M} \sigma\lceil\lambda\rceil + \phi\sqrt{T} \leq 4M\sqrt{2\log(KT^2)} \left(\sigma T^{\frac{1-\alpha/2}{1-(\alpha/2)^{M+1}}} + \phi\sqrt{T} \right)$$

where the last inequality is obtained replacing the value of λ as in eq. (22). Considering an union bound over all the events $\mathcal{G}_t$, we obtained the desired result.

□

B.2 Proofs of Section 4

B.2.1 Proof of Theorem 3

Fix an arm i and let t be the last time step where arm i was active before being eliminated. Consider the event

$$\tilde{\mathcal{G}} = \bigcap_{t \leq T} \left\{ UCB_t^{\tau}(i^*, \delta) > \max_i LCB_t^{\tau}(i, \delta) \right\}$$

from, the concentration bound, this intersection holds with probability at least $1 - 1/2T$, where we have replaced $\delta = 1/2KT^4$, as mentioned in the statement. Under this event, we can assume that with high probability arm i^* is still active at time t. Then, from the elimination condition of the algorithm and the definition of the upper/lower confidence intervals, we have that

$$\mu(i^*) - \mu(i) \leq 2 \left(R\sqrt{\frac{32\log(4\gamma_{\iota(\lfloor t \rfloor_\tau)}/\delta)}{\Pi_t^{\tau}(i)(\sigma^2 + \phi^2)}} + R\sqrt{\frac{32\log(4\gamma_{\iota(\lfloor t \rfloor_\tau)}/\delta)}{\Pi_t^{\tau}(i)(\sigma^2 + \phi^2)}} \right).$$

As previously, since all the retraining and the elimination of the arms happen at the end of a round-robin routine, when all the still-active arms have been played, we can conclude that $\Pi_t^{\tau}(i^*) = \Pi_t^{\tau}(i)$. Furthermore, since t is the last time action i was played, it holds $\Pi_t^{\tau}(i) = \Pi_T^{\tau}(i)$. We can then define the following event that holds with probability at least $1 - \delta$

$$\mathcal{G}_t^i = \left\{ |\mu(i^*) - \mu(i)| \leq 4R\sqrt{\frac{32\log(4\gamma_{\iota(\lfloor t \rfloor_\tau)}/\delta)}{\Pi_t^{\tau}(i)(\sigma^2 + \phi^2)}} \right\}.$$

Assuming the event $\mathcal{G}_t^i$ holds, using the regret decomposition formula, we obtain

$$\mathcal{R}_T = \sum_{i \in \mathcal{K}} \Delta_i \sum_{p \leq T} \mathbb{1}\{i_p = i\} \leq 4R\sqrt{\frac{32\log(4\gamma_{\iota(\lfloor t \rfloor_\tau)}/\delta)}{(\sigma^2 + \phi^2)}} \sum_{i \in \mathcal{K}} \frac{\sum_{p \leq T} \mathbb{1}\{i_p = i\}}{\sqrt{\Pi_T^{\tau}(i)}}$$

from which, replacing the explicit expression of $\Pi_T^\tau(i)$ and splitting the sum into the M epochs, we have

$$\mathcal{R}_T \leq 4R\sqrt{\frac{32\log\left(4\gamma_{\iota(\lfloor t\rfloor_\tau)}/\delta\right)}{(\sigma^2+\phi^2)}}\sum_{i\in\mathcal{K}}\sum_{j=0}^{M}\frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}}\mathbb{1}\{i_p=i\}}{\sqrt{\sum_{j=0}^{M}\frac{1}{\frac{\sigma^2}{\widehat{t}_j^\alpha\vee 1}+\frac{\phi^2}{N_j^\tau(i)}}}}$$

$$= 4R\sqrt{\frac{32\log\left(4\gamma_{\iota(\lfloor t\rfloor_\tau)}/\delta\right)}{(\sigma^2+\phi^2)}}\sum_{i\in\mathcal{K}}\left(\frac{\sum_{p=1}^{\widehat{t}_1}\mathbb{1}\{i_p=i\}}{\sqrt{\sum_{j=0}^{M}\frac{(\widehat{t}_j^\alpha\vee 1)N_j^\tau(i)}{\sigma^2N_j^\tau(i)+\phi^2(\widehat{t}_j^\alpha\vee 1)}}}+\sum_{j=1}^{M}\frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}}\mathbb{1}\{i_p=i\}}{\sqrt{\sum_{j=0}^{M}\frac{(\widehat{t}_j^\alpha\vee 1)N_j^\tau(i)}{\sigma^2N_j^\tau(i)+\phi^2(\widehat{t}_j^\alpha\vee 1)}}}\right)$$

$$\leq 4R\sqrt{\frac{32\log\left(4\gamma_{\iota(\lfloor t\rfloor_\tau)}/\delta\right)}{(\sigma^2+\phi^2)}}\sum_{i\in\mathcal{K}}\left(\frac{\sum_{p=1}^{\widehat{t}_1}\mathbb{1}\{i_p=i\}}{\sqrt{\frac{N_1^\tau(i)}{\sigma^2N_1^\tau(i)+\phi^2}}}+\sum_{j=1}^{M}\frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}}\mathbb{1}\{i_p=i\}}{\sqrt{\frac{\widehat{t}_j^\alpha N_j^\tau(i)}{\sigma^2N_j^\tau(i)+\phi^2\widehat{t}_j^\alpha}}}\right)$$

where the last inequality is obtained by omitting all the terms in the sum in the denominator but the one relative to the j-th epoch, since they are all positive. This can be further simplified as

$$\mathcal{R}_T \leq 4R\sqrt{\frac{32\log\left(4\gamma_{\iota(\lfloor t\rfloor_\tau)}/\delta\right)}{(\sigma^2+\phi^2)}}\sum_{i\in\mathcal{K}}\left(\frac{\sum_{p=1}^{\widehat{t}_1}\mathbb{1}\{i_p=i\}}{\sqrt{N_1^\tau(i)}}\sqrt{\sigma^2N_1^\tau(i)+\phi^2}\right.$$
$$\left.+\sum_{j=1}^{M}\frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}}\mathbb{1}\{i_p=i\}}{\sqrt{N_j^\tau(i)}}\sqrt{\frac{\sigma^2N_j^\tau(i)+\phi^2\widehat{t}_j^\alpha}{\widehat{t}_j^\alpha}}\right)$$

and

$$\mathcal{R}_T \leq 4R\sqrt{\frac{32\log\left(4\gamma_{\iota(\lfloor t\rfloor_\tau)}/\delta\right)}{(\sigma^2+\phi^2)}}\sum_{i\in\mathcal{K}}\left(\frac{\sum_{p=1}^{\widehat{t}_1}\mathbb{1}\{i_p=i\}}{\sqrt{N_1^\tau(i)}}\left(\sigma\sqrt{N_1^\tau(i)}+\phi\right)\right.$$
$$\left.+\sum_{j=1}^{M}\frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}}\mathbb{1}\{i_p=i\}}{\sqrt{N_j^\tau(i)}}\left(\sigma\sqrt{\frac{N_j^\tau(i)}{\widehat{t}_j^\alpha}}+\phi\right)\right).$$

Now note that, due to the way it is defined, $N_j^\tau(i)\leq\frac{\lambda^{\kappa(j+1)}-\lambda^{\kappa(j)}}{K}-1$. We start by analyzing the case $N_j^\tau(i)=\frac{\lambda^{\kappa(j+1)}-\lambda^{\kappa(j)}}{K}-1$, therefore using the fact that by definition $\tilde{t}_j\leq\widehat{t}_j\leq\tilde{t}_j+K$, and that $\tilde{t}_j=\lambda^{\kappa(j)}$ hence it holds

$$\frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}}\mathbb{1}\{i_p=i\}}{\sqrt{N_j^\tau(i)}}\leq\frac{\tilde{t}_{j+1}+K-\tilde{t}_j}{\sqrt{N_j^\tau(i)}}$$
$$\leq\frac{\lambda^{\kappa(j+1)}-\lambda^{\kappa(j)}+K}{\sqrt{\frac{\lambda^{\kappa(j+1)}-\lambda^{\kappa(j)}}{K}-1}}$$
$$\leq\frac{2(\lambda^{\kappa(j+1)}-\lambda^{\kappa(j)}+K)}{\sqrt{\frac{\lambda^{\kappa(j+1)}-\lambda^{\kappa(j)}}{K}}}$$
$$\leq 2\sqrt{K}\sqrt{\lambda^{\kappa(j+1)}-\lambda^{\kappa(j)}}+2K\sqrt{\frac{K}{\lambda^{\kappa(j+1)}-\lambda^{\kappa(j)}}}$$

from which

$$\sum_{j=1}^{M} \frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}}{\sqrt{N_j^\tau(i)}} \left(\sigma \sqrt{\frac{N_j^\tau(i)}{\widehat{t}_j^\alpha}} + \phi \right)$$

$$\leq 2 \sum_{j=1}^{M} \left(\sqrt{K} \sqrt{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}} + K \sqrt{\frac{K}{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}}} \right) \left(\frac{\sigma}{\sqrt{K}} \sqrt{\frac{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}}{\lambda^{\alpha\kappa(j)}}} + \phi \right)$$

$$\leq 2 \sum_{j=1}^{M} \sigma \frac{\lambda^{\kappa(j+1)}}{\sqrt{\lambda^{\alpha\kappa(j)}}} + \phi \sqrt{K \lambda^{\kappa(j+1)}} + \sigma \frac{K}{\sqrt{\lambda^{\alpha\kappa(j)}}} + \phi K \sqrt{\frac{K}{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}}}$$

$$\leq 2 \sum_{j=1}^{M} \left(\sigma \frac{\lambda^{\kappa(j+1)}}{\sqrt{\lambda^{\alpha\kappa(j)}}} + \phi \sqrt{K \lambda^{\kappa(j+1)}} \right) + 2MK\sqrt{K}(\sigma + \phi)$$

Instead, when $N_j^\tau(i) < \frac{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}}{K} - 1$, observe that it holds

$$N_j^\tau(i) = \sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\left\{ i_p = i \ \wedge N_j^\tau(i) \leq \frac{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}}{K} - 1 \right\} = \sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}$$

hence

$$\sum_{j=1}^{M} \frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}}{\sqrt{N_j^\tau(i)}} \left(\sigma \sqrt{\frac{N_j^\tau(i)}{\widehat{t}_j^\alpha}} + \phi \right)$$

$$= \sum_{j=1}^{M} \sqrt{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}} \left(\sigma \sqrt{\frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}}{\widehat{t}_j^\alpha}} + \phi \right)$$

$$\leq \sum_{j=1}^{M} \sigma \frac{\widehat{t}_{j+1}}{\sqrt{\widehat{t}_j^\alpha}} + \phi \sqrt{\widehat{t}_{j+1}}\,,$$

from which, replacing the values of $\widehat{t}_j$ as defined in the statement, and using again the fact that $\lceil \lambda^{\kappa(j)} \rceil \leq \lambda^{\kappa(j)} + K$

$$\sum_{j=1}^{M} \sigma \frac{\widehat{t}_{j+1}}{\sqrt{\widehat{t}_j^\alpha}} + \phi \sqrt{\widehat{t}_{j+1}} \leq \sum_{j=1}^{M} \sigma \frac{\lambda^{\kappa(j+1)} + K}{\sqrt{\lambda^{\alpha\kappa(j)}}} + \phi \sqrt{\lambda^{\kappa(j+1)} + K}$$

$$\leq \sum_{j=1}^{M} \left(\sigma \frac{\lambda^{\kappa(j+1)}}{\sqrt{\lambda^{\alpha\kappa(j)}}} + \phi \sqrt{\lambda^{\kappa(j+1)}} \right) + MK\,.$$

Combining the bounds for both $N_j^\tau(i) < \frac{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}}{K} - 1$ and $N_j^\tau(i) = \frac{\lambda^{\kappa(j+1)} - \lambda^{\kappa(j)}}{K} - 1$, we obtain that $\forall i \in \mathcal{K}$

$$\sum_{j=1}^{M} \frac{\sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i\}}{\sqrt{N_j^\tau(i)}} \left(\sigma \sqrt{\frac{N_j^\tau(i)}{\widehat{t}_j^\alpha}} + \phi \right)$$

$$\leq 4\sum_{j=1}^{M}\sigma\left(\frac{\lambda^{\kappa(j+1)}}{\sqrt{\lambda^{\alpha\kappa(j)}\vee 1}}+\phi\sqrt{K\lambda^{\kappa(j+1)}}\right)+4MK\sqrt{K}(\sigma+\phi)\,.$$

Notice how with an analogous argument to the one above it holds that

$$\frac{\sum_{p=1}^{\widehat{t}_1}\mathbb{1}\{i_p=i\}}{\sqrt{N_1^\tau(i)}}\left(\sigma\sqrt{N_1^\tau(i)}+\phi\right)\leq 4\sigma\lambda^{\kappa(1)}+4\phi\sqrt{K\lambda^{\kappa(1)}}+4K\sqrt{K}(\sigma+\phi)\,.$$

Therefore the regret becomes

$$\mathcal{R}_T\leq 16R\sqrt{\frac{32\log(4\gamma_{\iota(\lfloor t\rfloor_\tau)}/\delta)}{(\sigma^2+\phi^2)}}\sum_{i\in\mathcal{K}}\sum_{j=0}^{M}\sigma\frac{\lambda^{\kappa(j+1)}}{\sqrt{\lambda^{\alpha\kappa(j)}\vee 1}}+\phi\sqrt{K\lambda^{\kappa(j+1)}}+K\sqrt{K}(\sigma+\phi)\,.$$

From which using the fact that $\lambda^{\kappa(j+1)}=\lambda\sqrt{\lambda^{\alpha\kappa(j)}}$, as shown in Lemma 5, and $\lambda^{\kappa(j+1)}\leq T,\ \forall j$

$$\mathcal{R}_T\leq 16R\sqrt{\frac{32\log(4\gamma_{\iota(\lfloor t\rfloor_\tau)}/\delta)}{(\sigma^2+\phi^2)}}\sum_{i\in\mathcal{K}}\sum_{j=0}^{M}\sigma\lambda+\phi\sqrt{KT}+K\sqrt{K}(\sigma+\phi)$$

$$\leq 16R\sqrt{\frac{32\log(4\gamma_{\iota(\lfloor t\rfloor_\tau)}/\delta)}{(\sigma^2+\phi^2)}}(M+1)\left(K(\sigma\lambda+\phi\sqrt{KT})+K^2\sqrt{K}(\sigma+\phi)\right)$$

with probability at least $1-\delta$. We conclude by replacing the value of λ as specified in the statement. Lastly, note that it is possible to bound the quantity $\gamma_{\iota(\lfloor t\rfloor_\tau)}$, as

$$\gamma_{\iota(\lfloor t\rfloor_\tau)}=\frac{\lambda^{\kappa(j+1)}-\lambda^{\kappa(j)}}{K}-1\leq\frac{T}{K}\,.$$

Considering an union bound on the events $\mathcal{G}_t^i$, and $\tilde{\mathcal{G}}$ finally yields the result.

B.2.2 Proof of Theorem 4

We begin by considering a random strategy ψ as the one defined in the proof of Theorem 2. Specifically, $\psi=(\psi_t)_{t\leq T}$ is a family of functions, such that

$$\psi_t: I_t\to\{1,\dots,K\}\times\{0,\dots,M\}$$

thus, where the history I_t, which contains the past observations and internal randomization of the algorithm, gets mapped into a couple (i_{t+1},n_{t+1}), action retaining index. Furthermore, we limit again the set of feasible strategies to the ones for which the sequence of retraining indexes satisfies

$$\begin{aligned}&\text{(a)}\quad n_0=0\ \text{ and }\ n_T=M\\&\text{(b)}\quad n_s\leq n_t\quad\forall t>s\\&\text{(c)}\quad n_{t+1}-n_t\leq 1\end{aligned}$$

to incorporate the fact that such indexes can only augment, and increase at most by one between consecutive rounds. The retraining times therefore correspond to the instants of times $\widehat{t}_j$, such that $n_{\widehat{t}_j}=j$ and $n_{\widehat{t}_j}-n_{\widehat{t}_j-1}=1$, and they are arbitrarily selected as a function of the past through the strategy ψ.

In order to prove the result we will again consider the regret accumulated up to the j-th epoch, which corresponds to

$$\mathcal{R}_{\psi,\nu,j}=\mathbb{E}_\nu\left[\sum_{t=1}^{\widehat{t}_j}\mu(i)-\mu(i_t)\right]\,,$$

and observe that, for all j, it holds $\mathcal{R}_{\psi,\nu,j}\leq\mathcal{R}_{\psi,\nu,T}$. Moreover, we can consider an equivalent bandit problem in which for every time t, the agent observes

$$o_t^\tau(i)\mathbb{1}\{t\leq\widehat{t}_j\}\,,\ \forall i\in\mathcal{K}\tag{23}$$

where $o_t^\tau(i)$ are observations defined as in Equation (2), and analogously receives $r_t(i_t)\mathbb{1}\{t \leq \widehat{t}_j\}$. The associated arbitrary strategy $\psi^j = (\psi_t^j)_{t \leq T}$ satisfies $\psi_t^j = \psi_t$ if $t \leq \widehat{t}_j$ and $\psi_t^j(I'_t) = (1, n_{\widehat{t}_j})$ otherwise, where I'_t corresponds to the history up to time t of this modified bandit instance. We denote then by ν^j the probability measure obtained by the interaction between the random strategy ψ^j and the bandit instance whose reward are given in Equation (23).

We can now define the regret incurred by strategy ψ^j, with respect to the bandit instance ν^j up to time halted at the j-th epoch as

$$\mathcal{R}_{\psi^j,\nu^j,j} = \mathbb{E}_{\nu^j}\left[\sum_{t=1}^{\widehat{t}_j} \mu(i) - \mu(i_t)\right] .$$

Note that, due to a coupling argument the probability measures ν and ν^j coincide up to time $\widehat{t}_j$, therefore

$$\mathcal{R}_{\psi,\nu,j} = \mathcal{R}_{\psi^j,\nu^j,j}$$

hence we will prove the lower bound using the expression of the regret on the right-hand side. Applying Lemma 6, we find that an analogous version of the regret decomposition lemma holds in this case as well, such that

$$\mathcal{R}_{\psi^j,\nu^j,j} = \sum_{i \in \mathcal{K}} \Delta_i \mathbb{E}_{\nu^j}\left[N_i(\widehat{t}_j)\right]$$

where $N_i(\widehat{t}_j) = \sum_{t \leq \widehat{t}_j} \sum_{h \in [[0,j-1]]} \mathbb{1}\left\{\psi_{t-1}^j(I'_{t-1}) = (i,h)\right\}$.

We specify now two bandit instances ν and ν' that generate the observations for each arm. Recall that, for every arm i the observation $o_t^\tau(i)$ is obtained as the aggregation of the reward $r_t^\tau(i)$ and the bias term $\mathbb{s}(\widehat{t}_{n_t})\xi_{\widehat{t}_{n_t}}(i)$. For the instance ν, we will assume all the rewards to be distributed as a gaussian $\mathcal{N}(\mu(i), \phi^2)$, where

$$\mu(i) = \begin{cases} \Delta & \text{if } i = 1 \\ 0 & \text{if } i \neq 1 \end{cases}$$

where Δ is a constant to be specified later, and the bias term $\xi_{\widehat{t}_{n_t}}(i) \sim \mathcal{N}(0, \sigma^2)$.
Note that, since almost surely $\sum_{i \in \mathcal{K}} N_i(\widehat{t}_j) = \widehat{t}_j$, by taking the expectation on both sides it follows that

$$\sum_{i \in \mathcal{K}} \mathbb{E}_{\nu^j}[N_i(\widehat{t}_j)] = \mathbb{t}_j ,$$

where we have introduced the notation $\mathbb{t}_j \coloneqq \mathbb{E}_{\nu^j}[\widehat{t}_j]$. Therefore there must exists an arm k such that $\mathbb{E}_{\nu^j}[N_k(\widehat{t}_j)] \leq \mathbb{t}_j/K$. From this we define the instance ν' as $\nu'_i = \nu_i$ for all $i \neq k$, and $\nu'_k \sim \mathcal{N}(2\Delta, \phi^2) + \mathcal{N}(0, \mathbb{s}(\widehat{t}_{n_t})\sigma^2)$.

Considering such instances and Lemma 6, it follows that

$$\mathcal{R}_{\psi^j,\nu^j,j} \geq \mathbb{P}_{\nu^j}\left(N_1(\widehat{t}_j) \leq \frac{\mathbb{t}_j}{2}\right) \Delta \left\lfloor \frac{\mathbb{t}_j}{2} \right\rfloor$$

and

$$\mathcal{R}_{\psi^j,\nu'^j,j} \geq \mathbb{P}_{\nu'^j}\left(N_1(\widehat{t}_j) > \frac{\mathbb{t}_j}{2}\right) \Delta \left\lfloor \frac{\mathbb{t}_j}{2} \right\rfloor$$

where ν^j and ν'^j correspond respectively to the instances ν and ν' halted at $\widehat{t}_j$, as previously described. Now we use the fact that $\left\lfloor \frac{\mathbb{t}_j}{2} \right\rfloor \geq \frac{\mathbb{t}_j}{4}$ since $\mathbb{t}_j \geq 1$ due to the fact that $\widehat{t}_j \geq 1$ almost surely, given how the strategy is defined. Hence, we obtain

$$\max\{\mathcal{R}_{\psi^j,\nu^j,j}, \mathcal{R}_{\psi^j,\nu'^j,j}\} \geq \frac{\mathbb{t}_j \Delta}{8}\left(\mathbb{P}_{\nu^j}\left(N_1(\widehat{t}_j) \leq \frac{\mathbb{t}_j}{2}\right) + \mathbb{P}_{\nu'^j}\left(N_1(\widehat{t}_j) > \frac{\mathbb{t}_j}{2}\right)\right) ,$$

where we have made use of the fact that $a + b \leq 2\max\{a, b\}$. Furthermore, by applying Bretagnolle-Huber inequality, the expression above becomes

$$\max\{\mathcal{R}_{\psi^j,\nu^j,j}, \mathcal{R}_{\psi^j,\nu'^j,j}\} \geq \frac{\mathbb{t}_j \Delta}{16} \exp\left(-KL(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j})\right) \tag{24}$$

The following lemma provides a bound on the KL divergence between $\mathbb{P}_{\nu^j}$ and $\mathbb{P}_{\nu'^j}$.

Lemma 10. *Let ν^j and ν'^j the bandit instances obtained by halting ν and ν' at time $\widehat{t}_j$, and $\mathbb{P}_{\nu^j}$ and $\mathbb{P}_{\nu'^j}$ the corresponding distributions, then*

$$KL(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}) \leq 2\Delta^2 \min\left\{\frac{\mathbb{t}_j}{\phi^2}, \frac{\mathbb{t}_{j-1}^{\alpha} \vee 1}{\sigma^2}\log(\mathbb{t}_j)\right\}$$

We study these two possible bounds separately.

Initially consider the case $KL(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}) \leq {}^{2\Delta^2\mathbb{t}_j}/_{\phi^2}$. Replacing this bound in Equation (24) and choosing $\Delta = {}^{\phi}/_{\sqrt{2\mathbb{t}_j}}$, gives

$$\max\{\mathcal{R}_{\psi^j,\nu^j,j}, \mathcal{R}_{\psi^j,\nu'^j,j}\} \geq \frac{\phi\sqrt{\mathbb{t}_j}}{16\sqrt{2}e}\,.$$

Recalling that $\mathcal{R}_{\psi^j,\nu^j,j} \leq \mathcal{R}_{\psi,\nu,T}$, $\forall j$, we have

$$\begin{aligned}
\max_{\nu} \mathcal{R}_{\psi,\nu,T} &\geq \max_{j=1,\dots,M+1} \left\{\max\{\mathcal{R}_{\psi^j,\nu^j,j}, \mathcal{R}_{\psi^j,\nu'^j,j}\}\right\} \\
&\geq \max_{j=1,\dots,M+1} \left\{\frac{\phi\sqrt{\mathbb{t}_j}}{16\sqrt{2}e}\right\} \\
&= \frac{\phi\sqrt{T}}{16\sqrt{2}e}
\end{aligned} \tag{25}$$

where we have used the fact that, by construction of the strategy $\mathbb{t}_h \leq \mathbb{t}_j$, $\forall h \leq j$ and the convention $\mathbb{t}_{M+1} = T$.

On the other end, if $KL(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}) \leq {}^{2\Delta^2(\mathbb{t}_{j-1}^{\alpha}\vee 1)\log(\mathbb{t}_j)}/_{\sigma^2}$, then choosing $\Delta = {}^{\sigma}/_{\sqrt{2(\mathbb{t}_{j-1}^{\alpha}\vee 1)\log(\mathbb{t}_j)}}$ gives

$$\max\{\mathcal{R}_{\psi^j,\nu^j,j}, \mathcal{R}_{\psi^j,\nu'^j,j}\} \geq \frac{\sigma\mathbb{t}_j}{16\sqrt{2}e\sqrt{(\mathbb{t}_{j-1}^{\alpha}\vee 1)\log(\mathbb{t}_j)}}$$

from which

$$\begin{aligned}
\max_{\nu} \mathcal{R}_{\psi,\nu,T} &\geq \max_{j=1,\dots,M+1} \left\{\max\{\mathcal{R}_{\psi^j,\nu^j,j}, \mathcal{R}_{\psi^j,\nu'^j,j}\}\right\} \\
&\geq \max_{j=1,\dots,M+1} \left\{\frac{\sigma\mathbb{t}_j}{16\sqrt{2}e\sqrt{(\mathbb{t}_{j-1}^{\alpha}\vee 1)\log(\mathbb{t}_j)}}\right\} \\
&= \frac{\sigma}{16\sqrt{2}e(M+1)\log(T)} \sum_{j=1}^{M+1} \frac{\mathbb{t}_j}{\sqrt{(\mathbb{t}_{j-1}^{\alpha}\vee 1)}}\,.
\end{aligned}$$

The minimum of the right-hand side can be found by optimally choosing $\mathbb{t}_j$, which corresponds to solving the following constrained problem

$$\begin{aligned}
\underset{\mathbb{t}_0,\dots,\mathbb{t}_{M+1} \in \mathbb{R}}{\text{minimize}} \quad & \sum_{j=1}^{M+1} \frac{\mathbb{t}_j}{\sqrt{\mathbb{t}_{j-1}^{\alpha}\vee 1}} \\
\text{subject to} \quad & \mathbb{t}_0 = 0, \\
& \mathbb{t}_{M+1} = T, \\
& \mathbb{t}_j \leq \mathbb{t}_{j+1} \;\; \forall j
\end{aligned} \tag{26}$$

As done in the proof of Theorem 2, we begin by recovering a solution of the unconstrained problem, found by finding the values of the expected retraining times that do make the gradient of the function $\sum_{j=1}^{M+1} \frac{\mathbb{t}_j}{\sqrt{\mathbb{t}_{j-1}^{\alpha} \vee 1}}$ equal to zero. Thus we consider the derivative of the sum with respect to $\mathbb{t}_{j+1}$, which corresponds to

$$\frac{1}{\sqrt{\mathbb{t}_j^{\alpha}}} - \frac{\alpha}{2} \mathbb{t}_{j+2} \mathbb{t}_{j+1}^{\alpha/2-1} = 0$$

Solving for $\mathbb{t}_{j+2}$ and iterating, it gives

$$\begin{aligned}
\mathbb{t}_{j+2} &= \frac{2}{\alpha} \left(\frac{\mathbb{t}_{j+1}}{\mathbb{t}_j} \right)^{\alpha/2} \mathbb{t}_{j+1} \\
&= \frac{2}{\alpha} \left(\frac{\mathbb{t}_{j+1}}{\mathbb{t}_j} \right)^{\alpha/2} \frac{2}{\alpha} \left(\frac{\mathbb{t}_j}{\mathbb{t}_{j-1}} \right)^{\alpha/2} \mathbb{t}_j \\
&= \left(\frac{2}{\alpha} \right)^2 \left(\frac{\mathbb{t}_{j+1}}{\mathbb{t}_{j-1}} \right)^{\alpha/2} \mathbb{t}_j
\end{aligned}$$

from which, replacing again the expression for $\mathbb{t}_j$ and iterating, we find

$$\mathbb{t}_{j+2} = \left(\frac{2}{\alpha} \right)^{j+1} \mathbb{t}_{j+1}^{\alpha/2} \mathbb{t}_1 \,.$$

Further using this recursive formula for $\mathbb{t}_{j+1}$, gives

$$\begin{aligned}
\mathbb{t}_{j+2} &= \left(\frac{2}{\alpha} \right)^{j+1} \mathbb{t}_1 \left(\left(\frac{2}{\alpha} \right)^j \mathbb{t}_j^{\frac{\alpha}{2}} \mathbb{t}_1 \right)^{\frac{\alpha}{2}} \\
&= \left(\frac{2}{\alpha} \right)^{(j+1)+j\frac{\alpha}{2}} \mathbb{t}_j^{\left(\frac{\alpha}{2}\right)^2} \mathbb{t}_1^{1+\frac{\alpha}{2}} \,,
\end{aligned}$$

from which, iterating again we finally found

$$\mathbb{t}_{j+2} = \eta_{j+1} \mathbb{t}_1^{\sum_{h=0}^{i+1} \left(\frac{\alpha}{2}\right)^h}$$

where

$$\eta_{j+1} = \left(\frac{2}{\alpha} \right)^{\sum_{h=0}^{j+1} (j+1-h)\left(\frac{\alpha}{2}\right)^j}$$

Lastly, imposing the condition $\mathbb{t}_{M+1} = T$, we can obtain the optimal value for $\mathbb{t}_1$, which is given by

$$\mathbb{t}_1 = \left(\frac{1}{\eta_M} T \right)^{\frac{1-\alpha/2}{1-(\alpha/2)^{M+1}}} \,.$$

Note that the values derived in this way do satisfy the constraints of Equation (26), hence they constitute also a valid solution of the constrained optimization problem. We can now replace these values in the initial sum, observing that each $\mathbb{t}_j$ satisfies the relation

$$\mathbb{t}_j = (2/\alpha)^j \sqrt{(\mathbb{t}_{j-1}^{\alpha} \vee 1)} \mathbb{t}_1 \,,$$

such that

$$\sum_{j=1}^{M+1} \frac{\mathbb{t}_j}{\sqrt{\mathbb{t}_{j-1}^{\alpha} \vee 1}} = \sum_{j=1}^{M+1} \frac{(\alpha/2)^j \sqrt{(\mathbb{t}_{j-1}^{\alpha} \vee 1)} \mathbb{t}_1}{\sqrt{\mathbb{t}_{j-1}^{\alpha} \vee 1}} = \mathbb{t}_1 \sum_{j=1}^{M+1} \left(\frac{2}{\alpha} \right)^j$$

hence

$$\max_{\nu} \mathcal{R}_{\psi,\nu,T} \geq \frac{\sigma}{16\sqrt{2}e(M+1)\log(T)} \mathbb{t}_1 \sum_{j=1}^{M+1} \left(\frac{2}{\alpha} \right)^j = \frac{\sigma B}{16\sqrt{2}e(M+1)\log(T)} T^{\frac{1-\alpha/2}{1-(\alpha/2)^{M+1}}} \tag{27}$$

where

$$B = \frac{1}{\eta_M} \sum_{j=1}^{M+1} \left(\frac{2}{\alpha}\right)^j .$$

The final lower bound on the regret is obtained by considering the bounds obtained in Equation (25) and Equation (27), so that

$$\max_{\nu} \mathcal{R}_{\psi,\nu,T} \geq \frac{1}{16\sqrt{2}e} \max\left\{\phi\sqrt{T}, \frac{\sigma B}{(M+1)\log(T)} T^{\frac{1-\alpha/2}{1-(\alpha/2)^{M+1}}}\right\}$$

B.3 Proofs of Lemmas of Section 4

B.3.1 Proof of Lemma 2

We start our analysis by observing that the estimator $\widehat{r}_t^{\tau}(i)$ can be rewritten as

$$\widehat{r}_t^{\tau}(i) = \mu(i) + \frac{1}{\Pi_t^{\tau}(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\mathbb{s}(\widehat{t}_j)\xi_{\widehat{t}_j}(i)}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^{\tau}(i)} + \frac{1}{\Pi_t^{\tau}(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\epsilon_j}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^{\tau}(i)} ,$$

where

$$\epsilon_j = \frac{1}{N_j^{\tau}(i) \vee 1} \sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}\wedge t} \varepsilon_p(i_p) \mathbb{1}\{i_p = i \wedge N_j^{\tau}(i) \leq \gamma_j\} .$$

Hence, it is possible to rewrite the concentration inequality as follows

$$\mathbb{P}\big(|\widehat{r}_t^{\tau}(i) - \mu(i)| \geq u\big)$$

$$= \mathbb{P}\left(\left|\frac{1}{\Pi_t^{\tau}(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\mathbb{s}(\widehat{t}_j)\xi_{\widehat{t}_j}(i)}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^{\tau}(i)} + \frac{1}{\Pi_t^{\tau}(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\epsilon_j}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^{\tau}(i)}\right| \geq u\right)$$

where the right hand term can be bounded by

$$\mathbb{P}\left(\left|\frac{1}{\Pi_t^{\tau}(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\mathbb{s}(\widehat{t}_j)\xi_{\widehat{t}_j}(i)}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^{\tau}(i)}\right| \geq \frac{u}{2}\right) + \mathbb{P}\left(\left|\frac{1}{\Pi_t^{\tau}(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\epsilon_j}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^{\tau}(i)}\right| \geq \frac{u}{2}\right) . \tag{28}$$

Observe that, $\forall j < \iota(\lfloor t \rfloor_\tau)$ we have $t > \widehat{t}_{j+1}$, hence, from the way the constant γ_j is chosen, the following hold

$$N_j^{\tau}(i) = \sum_{p=\widehat{t}_j+1}^{\widehat{t}_{j+1}} \mathbb{1}\{i_p = i \ \wedge \ N_j^{\tau}(i) \leq \gamma_j\} = \gamma_j .$$

In the case of $j = \iota(\lfloor t \rfloor_\tau)$, instead $t < \widehat{t}_{\iota(\lfloor t \rfloor_\tau)+1}$, hence it holds

$$N_{\iota(\lfloor t \rfloor_\tau)}^{\tau}(i) = \sum_{p=\widehat{t}_{\iota(\lfloor t \rfloor_\tau)}+1}^{t} \mathbb{1}\{i_p = i \ \wedge \ N_j^{\tau}(i) \leq \gamma_{\iota(\lfloor t \rfloor_\tau)}\} \leq \gamma_{\iota(\lfloor t \rfloor_\tau)} .$$

Now fix $0 < l \leq \gamma_{\iota(\lfloor t \rfloor_\tau)}$ such that $N_{\iota(\lfloor t \rfloor_\tau)}^{\tau}(i) = l$, and consider the random variables

$$\frac{1}{\Pi_t^{\tau}(i)} \frac{\mathbb{s}(\widehat{t}_j)\xi_{\widehat{t}_j}(i)}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^{\tau}(i)}$$

which are adapted with respect to the filtration $\mathcal{F} = \sigma(\xi_{\widehat{t}_0}(i), \ldots, \xi_{\widehat{t}_j}(i)) \subset \mathcal{F}_t$. Note that this holds because the functions $\lfloor t \rfloor_\tau$ and $\mathbb{1}\{i_p = i \ \wedge \ N_j^{\tau}(i) \leq \gamma_j\}$ $\forall p \leq t$, hidden in the definition of $\frac{1}{\Pi_t^{\tau}(i)}$

are $\mathcal{F}_{t-1}$ measurable. Moreover, this corresponds to the product between the random variable ξ_j and a deterministic constant. The difference between two consecutive terms is bounded by

$$z_j \coloneqq \left| \frac{1}{\Pi_t^\tau(i)} \frac{\mathbb{s}(\widehat{t}_j)\xi_{\widehat{t}_j}(i)}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)} \right| \leq \frac{R}{\sigma\Pi_t^\tau(i)} \left| \frac{(\widehat{t}_j \vee 1)^{-\alpha/2}\sigma}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)} \right|$$

$$\overset{(i)}{\leq} \frac{R}{\sigma\Pi_t^\tau(i)} \left| \frac{\sqrt{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)}}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)} \right|$$

$$\leq \frac{R}{\sigma\Pi_t^\tau(i)} \left| \frac{1}{\sqrt{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)}} \right|$$

where we have used the fact that random variables have support bounded by R. Hence, we can apply the standard Hoeffding-Azuma Inequality, by noticing before that

$$\sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} z_j^2 = \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{R^2}{\sigma^2\Pi_t^\tau(i)^2} \left(\frac{1}{\sqrt{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)}} \right)^2$$

$$\leq \frac{4R^2}{\sigma^2\Pi_t^\tau(i)^2} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{1}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)}$$

$$\leq \frac{4R^2}{\sigma^2\Pi_t^\tau(i)} .$$

Therefore,

$$\mathbb{P}\left(\left| \frac{1}{\Pi_t^\tau(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\mathbb{s}(\widehat{t}_j)\xi_{\widehat{t}_j}(i)}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)} \right| \geq \omega \right) \leq \exp\left(-\frac{\omega^2\sigma^2\Pi_t^\tau(i)}{8R^2} \right) .$$

Hence, considering an union-bound among all values of l, it holds

$$\mathbb{P}\left(\left| \frac{1}{\Pi_t^\tau(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\mathbb{s}(\widehat{t}_j)\xi_{\widehat{t}_j}(i)}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)} \right| \geq \frac{u}{2} \right)$$

$$\leq \sum_{l=1}^{\gamma_{\iota(\lfloor t \rfloor_\tau)}} \mathbb{P}\left(\mathbb{1}\left\{ N_{\iota(\lfloor t \rfloor_\tau)}^\tau(i) = l \right\} \left\{ \left| \frac{1}{\Pi_t^\tau(i)} \sum_{j=0}^{\iota(\lfloor t \rfloor_\tau)} \frac{\mathbb{s}(\widehat{t}_j)\xi_{\widehat{t}_j}(i)}{(\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)} \right| \geq \frac{u}{2} \right\} \right)$$

$$\overset{(i)}{\leq} \sum_{l=1}^{\gamma_{\iota(\lfloor t \rfloor_\tau)}} 2e^{-\frac{u^2}{32}\frac{\sigma^2\Pi_t^\tau(i)}{R^2}} \leq 2\gamma_{\iota(\lfloor t \rfloor_\tau)} e^{-\frac{u^2}{32}\frac{\sigma^2\Pi_t^\tau(i)}{R^2}} .$$

An analogous argument holds for the variable

$$\frac{1}{\Pi_t^\tau(i)} \frac{\epsilon_j}{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)}$$

which is again adapted to the filtration $\mathcal{F} = \sigma\left((\epsilon_p(i_p))_{p \leq t}, (i_p)_{p \leq t} \right) \subset \mathcal{F}_t$, and for which the difference between two consecutive terms is again bounded by

$$\tilde{z}_j \coloneqq \frac{R}{\phi\Pi_t^\tau(i)} \left| \frac{1}{\sqrt{\sigma^2(\widehat{t}_j \vee 1)^{-\alpha} + \phi^2/N_j^\tau(i)}} \right|$$

So that it holds

$$\mathbb{P}\left(\left|\frac{1}{\Pi_t^\tau(i)}\sum_{j=0}^{\iota(\lfloor t\rfloor_\tau)}\frac{\epsilon_j}{\sigma^2(\widehat{t}_j\vee 1)^{-\alpha}+{\phi^2}/{N_j^\tau(i)}}\right|\geq\omega\right)\leq\exp\left(-\frac{\omega^2\phi^2\Pi_t^\tau(i)}{8R^2}\right).$$

From this, we can conclude

$$\mathbb{P}\left(|\widehat{r}_t^\tau(i)-\mu(i)|\geq u\right)\leq 4\gamma_{\iota(\lfloor t\rfloor_\tau)}e^{-\frac{u^2}{32}\frac{\Pi_t^\tau(i)(\sigma^2+\phi^2)}{R^2}}.$$

B.3.2 Proof of Lemma 9

The proof follows the same steps of the proof of Lemma 4, again by applying lemma 8 to the estimator $\widehat{r}_t^j(i)$. Note that this is obtained by initially computing an empirical average to obtain $\widehat{r}_t^j(i)$, which is therefore ${\phi}/{N_j^\tau}$-subgaussian. And then further modified to obtain the final estimator by considering a convex combination, whose weights are

$$\frac{1}{\frac{\sigma^2}{\widehat{t}_j^\alpha}+\frac{\phi^2}{N_j^\tau}},$$

hence the final estimator is $\tilde{\sigma}$-subgaussian, where

$$\tilde{\sigma}=\sqrt{\frac{1}{\sum_{j=0}^{\iota(\lfloor t\rfloor_\tau)}\frac{1}{\frac{\sigma^2}{\widehat{t}_j^\alpha}+\frac{\phi^2}{N_j^\tau}}}}$$

B.3.3 Proof of Lemma 10

The proof is this lemma is adapted to our case from the proof of Lemma 15.1 of Lattimore and Szepesvári [2020]. We start by observing that

$$KL\left(\mathbb{P}_{\nu^j},\mathbb{P}_{\nu'^j}\right)=\mathbb{E}_{\nu^j}\left[\log\left(\frac{\mathrm{d}\mathbb{P}_{\nu^j}}{\mathrm{d}\mathbb{P}_{\nu'^j}}\right)\right]$$

It is possible to define $P_{i,h}^t$ and $P_{i,h}^{'t}$ as the conditional distributions of the observation of arm i, at time t given $\widehat{t}_h$, and $f_{i,h}^t$ and $f_{i,h}^{'t}$ their respective densities. These distributions are well defined since the switching times $\widehat{t}_h$ are fixed. Moreover, denote by ρ^{j-1} the counting measure on $[[0,j-1]]$, and $\lambda^{j-1}=\sum_{i\mathcal{K}}\sum_{h=0}^{j-1}\sum_{t=\widehat{t}_h+1}^{\widehat{t}_{h+1}}P_{i,h}^t+P_{i,h}^{'t}$. Then, for every fixed arm i and fixed retraining times $\widehat{t}_1,\ldots,\widehat{t}_{j-1}$ the Radon-Nikodym derivative of $\mathbb{P}_{\nu^j}$ is

$$\frac{\mathrm{d}\mathbb{P}_{\nu^j}}{\mathrm{d}(\lambda^{j-1}\times\rho^{j-1})}\left((i,N_1),o_1^\tau(i),\ldots,(i,N_{\widehat{t}_j}),o_{\widehat{t}_j}^\tau(i)\right)=\prod_{t=1}^{\widehat{t}_j}\mathbb{P}\left((i,N_t)\middle|I'_{t-1}\right)f_{i,N_t}^t(o_t^\tau(i))$$

where I'_{t-1} is the history up to time $t-1$. Therefore, by conditioning on the previous retraining times and applying the chain rule the following holds

$$\begin{aligned}KL\left(\mathbb{P}_{\nu^j},\mathbb{P}_{\nu'^j}\right)&=\mathbb{E}_{\nu^j}\left[\log\left(\frac{\mathrm{d}\mathbb{P}_{\nu^j}}{\mathrm{d}\mathbb{P}_{\nu'^j}}\right)\right]\\&=\mathbb{E}_{\nu^j}\left[\mathbb{E}_{\nu^j}\left[\log\left(\frac{\mathrm{d}\mathbb{P}_{\nu^j}}{\mathrm{d}\mathbb{P}_{\nu'^j}}\right)\middle|(\widehat{t}_h)_{h=1}^{j-1}\right]\right]\\&=\mathbb{E}_{\nu^j}\left[\sum_{t=1}^{\widehat{t}_j}\sum_{i\in\mathcal{K}}\mathbb{E}_{\nu^j}\left[\log\left(\frac{f_{i,N_t}^t(o_t^\tau(i))}{f_{i,N_t}^{'t}(o_t^\tau(i))}\right)\middle|(\widehat{t}_h)_{h=1}^{j-1}\right]\right]\end{aligned}$$

Furthermore it holds

$$\mathbb{E}_{\nu^j}\left[\log\left(\frac{f_{i,N_t}^t(o_t^\tau(i))}{f_{i,N_t}^{'t}(o_t^\tau(i))}\right)\middle|(\widehat{t}_h)_{h=1}^{j-1}\right]=KL(P_{i,N_t}^t,P_{i,N_t}^{'t}),$$

where we have used the fact that, under $\mathbb{P}_{\nu^j}$ the distribution of $o_t^\tau(i)$ is $dP_{i,N_t}^t = f_{i,N_t}^t d\lambda^{j-1}$. The expression above then becomes

$$KL\left(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}\right) = \sum_{t=1}^{\widehat{t}_j} \sum_{i \in \mathcal{K}} \mathbb{E}_{\nu^j}\left[KL(P_{i,N_t}^t, P_{i,N_t}^{'t})\right]$$

$$= \sum_{t=1}^{\widehat{t}_j} \sum_{i \in \mathcal{K}} \sum_{h=0}^{j-1} \mathbb{E}_{\nu^j}\left[\mathbb{1}\{N_t = h\} KL(P_{i,N_t}^t, P_{i,N_t}^{'t})\right]$$

$$= \sum_{t=1}^{\widehat{t}_j} \sum_{i \in \mathcal{K}} \sum_{h=0}^{j-1} \mathbb{E}_{\nu^j}\left[\mathbb{1}\{N_t = h\} KL(P_{i,h}^t, P_{i,h}^{'t})\right] \tag{29}$$

We now analyze the term $KL\left(P_{i,h}^t, P_{i,h}^{'t}\right)$, by deriving the explicit expression of the distributions involved thanks to the following lemma.

Lemma 11. *Let ν be a bandit instance, where the rewards of each arm follow the structure of Equation (2), with $\varepsilon_t \sim \mathcal{N}(0, \phi^2)$ and $\xi_{\lfloor t \rfloor_\tau} \sim \mathcal{N}(0, \sigma^2)$, and ν^j the distribution obtained by halting ν at time $\widehat{t}_j$. Then the conditional distribution $P_{i,h}^t$ of the observation $o_t^\tau(i)$, received by arm i at time t given $\widehat{t}_h$, follows a gaussian law with parameters $\hat{\mu}(i)$, $\hat{\sigma}^2$, where*

$$\hat{\mu}(i) = \mu(i) + \sum_{p=\widehat{t}_h+1}^{t-1} \frac{o_p^\tau(i) - \mu(i)}{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t - \widehat{t}_h - 2)}$$

and

$$\hat{\sigma}^2 = \frac{\phi^2\left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t - \widehat{t}_h - 1)\right)}{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t - \widehat{t}_h - 2)}\,.$$

Therefore, the KL divergence between $P_{i,h}^t$ and $P_{i,h}^{'t}$ can be explicitly computed thanks to the formula of the KL divergence between gaussian distributions with the same variance. Furthermore note that the averages of ν^j and ν'^j only differ with respect to arm k, hence the same holds for the averages of $P_{i,h}^t$ and $P_{i,h}^{'t}$. From this, we deduce that the only non-zero divergence is the one corresponding to arm k, for which the averages of the two distributions are

$$\hat{\mu}(k) = \sum_{p=\widehat{t}_h+1}^{t-1} \frac{o_p^\tau(i)}{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t - \widehat{t}_h - 2)} \qquad \hat{\mu}'(k) = 2\Delta + \sum_{p=\widehat{t}_h+1}^{t-1} \frac{o_p^\tau(i) - 2\Delta}{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t - \widehat{t}_h - 2)}\,.$$

Therefore

$$KL\Big(P_{k,h}^t, P_{k,h}^{'t}\Big) = \frac{(\hat{\mu}(k) - \hat{\mu}'(k))^2}{2\hat{\sigma}^2}$$

Note that, the difference between the averages corresponds to

$$\left(\sum_{p=\widehat{t}_h+1}^{t-1} \frac{o_p^\tau(i)}{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t - \widehat{t}_h - 2)} - 2\Delta - \sum_{p=\widehat{t}_h+1}^{t-1} \frac{o_p^\tau(i) - 2\Delta}{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t - \widehat{t}_h - 2)}\right)$$

$$= \left(-2\Delta + \sum_{p=\widehat{t}_h+1}^{t-1} \frac{2\Delta}{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t - \widehat{t}_h - 2)}\right)$$

Thus

$$KL\Big(P_{k,h}^t, P_{k,h}^{'t}\Big)$$

$$= \left(-2\Delta + \sum_{p=\widehat{t}_h+1}^{t-1} \frac{2\Delta}{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-2)} \right)^2 \frac{1}{2} \frac{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-2)}{\phi^2 \left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1) \right)}$$

$$= 4\Delta^2 \frac{\left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} \right)^2}{\left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-2) \right)^2} \frac{1}{2} \frac{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-2)}{\phi^2 \left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1) \right)}$$

$$= 2\Delta^2 \frac{\frac{\phi^2}{\left(\mathfrak{s}(\widehat{t}_h)^2\sigma^2\right)^2}}{\left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-2) \right)} \frac{1}{\left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1) \right)}$$

$$= \frac{2\Delta^2 \frac{\phi^2}{\left(\mathfrak{s}(\widehat{t}_h)^2\sigma^2\right)^2}}{\left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1) \right)^2 + \left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1) \right)}$$

We observe now, that the KL divergence can be further bounded into two different ways.

Case 1 We have

$$KL(P^t_{k,h}, P'^t_{k,h}) = \frac{2\Delta^2 \frac{\phi^2}{\left(\mathfrak{s}(\widehat{t}_h)^2\sigma^2\right)^2}}{\left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1) \right)^2 + \left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1) \right)}$$

$$\leq \frac{2\Delta^2 \frac{\phi^2}{\left(\mathfrak{s}(\widehat{t}_h)^2\sigma^2\right)^2}}{\left(\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1) \right)^2 + \frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2}}$$

$$\leq \frac{2\Delta^2 \frac{\phi^2}{\left(\mathfrak{s}(\widehat{t}_h)^2\sigma^2\right)^2}}{\frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2}(t-\widehat{t}_h-1) + \frac{\phi^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2}}$$

$$= \frac{2\Delta^2}{\mathfrak{s}(\widehat{t}_h)^2\sigma^2(t-\widehat{t}_h)}$$

therefore, by replacing this in Equation (29), together with $\mathfrak{s}(\widehat{t}_h)^2 \leq (\widehat{t}_h \vee 1)^{-\alpha}$ we recover

$$KL\left(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}\right) \leq \sum_{t=1}^{\widehat{t}_j} \sum_{h=0}^{j-1} \mathbb{E}_{\nu^j} \left[\mathbb{1}\{N_t = h\} \frac{2\Delta^2(\widehat{t}_h^{\alpha} \vee 1)}{\sigma^2(t-\widehat{t}_h)} \right]$$

from which, the fact that $\widehat{t}_h \leq \widehat{t}_{j-1}$ almost surely for $h \leq j-1$ and an application of Wald's lemma, gives

$$KL\left(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}\right) \leq \frac{2\Delta^2}{\sigma^2} \mathbb{E}_{\nu^j}[\widehat{t}_{j-1}^{\alpha} \vee 1] \sum_{t=1}^{\widehat{t}_j} \sum_{h=0}^{j-1} \mathbb{E}_{\nu^j} \left[\frac{\mathbb{1}\{N_t = h\}}{t-\widehat{t}_h} \right]$$

$$\leq \frac{2\Delta^2}{\sigma^2} \mathbb{E}_{\nu^j}[\widehat{t}_{j-1}^{\alpha} \vee 1] \mathbb{E}_{\nu^j}\left[\log(\widehat{t}_j)\right]$$

and again by using Jensen's inequality for both the expectations and the notation $\mathbb{E}_{\nu^j}[\widehat{t}_{j-1}] = \mathbb{t}_{j-1}$, we conclude

$$KL\left(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}\right) \leq \frac{2\Delta^2(\mathbb{t}_{j-1}^{\alpha} \vee 1)}{\sigma^2} \log(\mathbb{t}_j)\,. \tag{30}$$

Case 2 In this case, we bound the KL divergence between $P^t_{k,h}$ and $P^{'t}_{k,h}$ as

$$KL(P^t_{k,h}, P^{'t}_{k,h}) = \frac{2\Delta^2 \frac{\phi^2}{\left(\mathbb{s}(\widehat{t}_h)^2\sigma^2\right)^2}}{\left(\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1)\right)^2 + \left(\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1)\right)}$$

$$\leq \frac{2\Delta^2 \frac{\phi^2}{\left(\mathbb{s}(\widehat{t}_h)^2\sigma^2\right)^2}}{\left(\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1)\right)^2}$$

$$\leq \frac{2\Delta^2 \frac{\phi^2}{\left(\mathbb{s}(\widehat{t}_h)^2\sigma^2\right)^2}}{\left(\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2}\right)^2}$$

$$\leq \frac{2\Delta^2}{\phi^2}$$

Which replaced above gives

$$KL\left(\mathbb{P}_{\nu^j}, \mathbb{P}_{\nu'^j}\right) \leq \sum_{t=1}^{\widehat{t}_j} \sum_{h=0}^{j-1} \mathbb{E}_{\nu^j}\left[\mathbb{1}\{N_t = h\}\frac{2\Delta^2}{\phi^2}\right] = \frac{2\Delta^2}{\phi^2}\mathbb{t}_j \tag{31}$$

The final bound follows by considering the min between eq. (30) and eq. (31)

B.3.4 Proof of Lemma 11

By the definition of the bandit instance ν, we can show that the vector of the observations obtained during epoch h up to time t by arm i has a gaussian distribution. Indeed, let $\underline{\varepsilon}(i) = (\varepsilon_p(i))_{p\in[[\widehat{t}_h+1,t]]}$ be the vector of all the noise terms obtained from the start of the h-th epoch, where, by construction, each component is an independent drawn by a gaussian with parameters 0 and ϕ^2. Let $\mathbb{s}(\widehat{t}_h)\xi_{\widehat{t}_h}(i)$ the bias term of epoch h, which, again by definition of the bandit instance, is distributed as a gaussian with parameters 0 and $\mathbb{s}(\widehat{t}_h)^2\sigma^2$. Let $\underline{\xi}_h(i) = (\mathbb{s}(\widehat{t}_h)\xi_{\widehat{t}_h}(i))_{p\in[[\widehat{t}_h+1,t]]}$ the vector containing $(t-\widehat{t}_h-1)$ copies of the bias term. Then, the vector of the observations up to time t by arm i, $\underline{o}^\tau(i) = (o^\tau_p(i))_{p\in[[\widehat{t}_h+1,p]]}$, can be written as

$$\underline{o}^\tau(i) = \underline{\mu}(i) + \underline{\xi}_h(i) + \underline{\varepsilon}(i)\,,$$

where $\underline{\mu}(i)$ is a vector containing $(t-\widehat{t}_h-1)$ copies of the average $\mu(i)$. Since gaussian distributions are maintained by linear transformations, we have that the distribution of $\underline{o}^\tau(i)$, is also gaussian with average $\underline{\mu}(i)$ and covariance matrix Σ, defined as

$$\Sigma = \begin{bmatrix} \mathbb{s}(\widehat{t}_h)^2\sigma^2 + \phi^2 & \mathbb{s}(\widehat{t}_h)^2\sigma^2 \cdots & \mathbb{s}(\widehat{t}_h)^2\sigma^2 \\ \vdots & \ddots & \vdots \\ \mathbb{s}(\widehat{t}_h)^2\sigma^2 & \cdots & \mathbb{s}(\widehat{t}_h)^2\sigma^2 + \phi^2 \end{bmatrix}.$$

The distribution $P^t_{i,h}$ corresponds to the distribution of the last component of the vector $\underline{o}^\tau(i)$ having observed all the previous ones. Therefore, using the formulas for the conditional distribution of a subvector of a gaussian multivariate distribution, we find that $o^\tau_t(i)|(o^\tau_p(i))^{t-1}_{p=\widehat{t}_h+1} \sim \mathcal{N}(\hat{\mu}(i), \hat{\sigma}^2)$, where

$$\hat{\mu}(i) = \mu(i) + \Sigma_1\Sigma_2^{-1} \cdot (o^\tau_p(i) - \mu(i))^{t-1}_{p=\widehat{t}_h+1} \tag{32}$$

and

$$\hat{\sigma}^2 = (\mathbb{s}(\widehat{t}_h)^2\sigma^2 + \phi^2) - \Sigma_1\Sigma_2^{-1}\Sigma_1^\top\,, \tag{33}$$

where Σ_1 is a row vector containing $(t-\widehat{t}_j-2)$ copies of ϕ^2, and Σ_2 is a squared $(t-\widehat{t}_j-2)$ matrix obtained from Σ by removing the last row and column. Note that Σ_2 can be written as

$$\Sigma_2 = \begin{bmatrix} \mathbb{s}(\widehat{t}_h)^2\sigma^2 & \cdots & \widehat{t}_h^{\mathbb{s}}(\widehat{t}_h)^2\sigma^2 \\ \vdots & \ddots & \vdots \\ \mathbb{s}(\widehat{t}_h)^2\sigma^2 & \cdots & \mathbb{s}(\widehat{t}_h)^2\sigma^2 \end{bmatrix} + \begin{bmatrix} \phi^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi^2 \end{bmatrix}$$

therefore we can compute its inverse by using the fact that if B is a matrix of rank 1, and A a generic matrix, then

$$(A+B)^{-1} = B^{-1} - \frac{1}{1+g}B^{-1}AB^{-1}$$

where $g = tr(AB^{-1})$ the trace of the product. In our case AB^{-1} corresponds to

$$\begin{bmatrix} \mathbb{s}(\widehat{t}_h)^2\sigma^2 & \cdots & \mathbb{s}(\widehat{t}_h)^2\sigma^2 \\ \vdots & \ddots & \vdots \\ \mathbb{s}(\widehat{t}_h)^2\sigma^2 & \cdots & \mathbb{s}(\widehat{t}_h)^2\sigma^2 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\phi^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\phi^2} \end{bmatrix} = \begin{bmatrix} \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2} & \cdots & \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2} \\ \vdots & \ddots & \vdots \\ \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2} & \cdots & \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2} \end{bmatrix},$$

therefore its trace corresponds to $\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)$, thus it is possible to compute the inverse of Σ_2 as

Σ_2^{-1}

$$= \begin{bmatrix} \frac{1}{\phi^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\phi^2} \end{bmatrix} - \frac{1}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \begin{bmatrix} \frac{1}{\phi^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\phi^2} \end{bmatrix} \cdot \begin{bmatrix} \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2} & \cdots & \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2} \\ \vdots & \ddots & \vdots \\ \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2} & \cdots & \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\phi^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\phi^2} \end{bmatrix} - \frac{1}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \begin{bmatrix} \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4} & \cdots & \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4} \\ \vdots & \ddots & \vdots \\ \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4} & \cdots & \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\phi^2} - \frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} & \cdots & -\frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \\ \vdots & \ddots & \vdots \\ -\frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} & \cdots & \frac{1}{\phi^2} - \frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \end{bmatrix}.$$

Moreover

$\Sigma_1\Sigma_2^{-1}$

$$= \begin{bmatrix} \mathbb{s}(\widehat{t}_h)^2\sigma^2 & \cdots & \mathbb{s}(\widehat{t}_h)^2\sigma^2 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\phi^2} - \frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} & \cdots & -\frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \\ \vdots & \ddots & \vdots \\ -\frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} & \cdots & \frac{1}{\phi^2} - \frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \end{bmatrix}$$

$$= \left(\mathbb{s}(\widehat{t}_h)^2\sigma^2 \left(\frac{1}{\phi^2} - \frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \right) - \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}(t-\widehat{t}_h-3)}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \right)[1,\ldots,1]$$

$$= \left[\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4} - \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2 \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}(t-\widehat{t}_h-2)}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \quad \cdots \quad \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4} - \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2 \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^4}(t-\widehat{t}_h-2)}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \right]$$

$$= \left[\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}\left(1 - \frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)}\right) \quad \cdots \quad \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}\left(1 - \frac{\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)}\right) \right]$$

$$= \left[\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}\frac{1}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \quad \cdots \quad \frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}\frac{1}{1+\frac{\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\phi^2}(t-\widehat{t}_h-2)} \right]$$

$$= \left[\frac{1}{\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2}+(t-\widehat{t}_h-2)} \quad \cdots \quad \frac{1}{\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2}+(t-\widehat{t}_h-2)} \right]$$

Therefore, replacing this computations in Equation (32), we obtain

$$\hat{\mu}(i) = \mu(i) + \sum_{p=\widehat{t}_h+1}^{t-1} \frac{o_p^\tau(i) - \mu(i)}{\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-2)}$$

and, since

$$\Sigma_1\Sigma_2^{-1}\Sigma_1^\top = \left[\frac{1}{\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2}+(t-\widehat{t}_h-2)} \quad \cdots \quad \frac{1}{\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2}+(t-\widehat{t}_h-2)} \right] \cdot \begin{bmatrix} \mathbb{s}(\widehat{t}_h)^2\sigma^2 \\ \vdots \\ \mathbb{s}(\widehat{t}_h)^2\sigma^2 \end{bmatrix}$$

$$= \frac{(t-\widehat{t}_h-2\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-2)},$$

using Equation (33)

$$\hat{\sigma}^2 = (\mathbb{s}(\widehat{t}_h)^2\sigma^2 + \phi^2) - \frac{(t-\widehat{t}_h-2)\mathbb{s}(\widehat{t}_h)^2\sigma^2}{\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-2)}$$

$$= \frac{\phi^2\left(\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-1)\right)}{\frac{\phi^2}{\mathbb{s}(\widehat{t}_h)^2\sigma^2} + (t-\widehat{t}_h-2)}.$$