Thompson Sampling-like Algorithms for Stochastic Rising Rested Bandits

Marco Fiandri* **Alberto Maria Metelli*** **Francesco Trovò***

Abstract

Stochastic rising rested bandit (SRRB) is a specific bandit setting where the arms' expected rewards increase as they are pulled. They model scenarios in which the performances of the different options grow as an effect of an underlying learning process (e.g., online model selection). Even if the bandit literature provides specifically crafted algorithms based on upper-confidence bound approaches for such a setting, no study about Thompson sampling-like algorithms has been performed. Indeed, the specific trend and the strong regularity of the expected rewards given by the SRRB setting suggest that specific instances may be tackled effectively using classical Thompson sampling or some adapted versions. This work provides a novel theoretical analysis of the regret that such algorithms suffer in SRRB. Our results show that, under specific assumptions on the reward functions, even the Thompson sampling-like algorithms achieve the no-regret property.

1 Introduction

In the ever-evolving landscape of decision-making under uncertainty, the field of Multi-Armed Bandits (MAB) has witnessed a paradigm shift with the emergence of dynamic phenomena. Traditional bandit models (e.g., the one presented by [12, 26]) consider static environments, assuming arms with expected rewards that do not change during the learning process. However, many real-world applications present a more intricate scenario where the rewards associated with each arm dynamically evolve either over time or depending on the played actions. In particular, two significantly different scenarios have been analyzed in the literature: *restless* and *rested*. While the restless MAB setting assumes that the arms' expected reward changes as an effect of nature, in the *rested* MAB setting [43], the arm evolution is triggered by its pull. Many *restless* settings have been analyzed in the past, including the abruptly changing ones [19] and the smoothly changing ones [46]. Conversely, only recently, the *rested* setting has raised the attention of the bandit community [31].

This paper delves into this dynamic scenario, focusing specifically on the domain of *stochastic rising rested bandits* [SRRB, 31]. The SRRB scenario reflects situations where the arms' expected rewards *increase*, encapsulating the essence of growing trends in various applications (e.g., learning processes). Examples of such settings are represented by the so-called Combined Algorithm Selection and Hyperparameter optimization (CASH, [45, 28]), whose goal is to identify the best learning algorithm and the best hyperparameter configuration for a given machine learning task, one of the most fundamental problems in Automatic Machine Learning (AutoML). This setting of the concave rising bandit is also extremely useful when modeling satiation effects in recommendations, like the ones studied by [15, 48], and in particular, concave learning curves have been shown to arise in various laboratory environments (e.g., see [24, 5]) and is very natural and common in the context of human learning (e.g., see the work by [41]). Similar problems arise when we want to select an optimization algorithm among a predefined set to optimize a given function, a.k.a. online model selection problem [31]. Such problems can be tackled effectively using an SRRB modeling approach.

So far, the research has focused on designing algorithms capable of adapting to and exploiting these evolving trends, managing the delicate balance between exploration and exploitation. For instance, the seminal work by [31] approached the problem by designing a sliding-window algorithm

*Politecnico di Milano, Milan, Italy

17th European Workshop on Reinforcement Learning (EWRL 2024).

based on upper-confidence bounds for SRRB to provide a worst-case regret of the order of $\tilde{O}(T^{\frac{2}{3}} + \Upsilon(T))$, T being the learning horizon of the learning process and $\Upsilon(T)$ is a problem-dependent quantity which characterizes the growth rate of the arm expected rewards.[2] Similarly, [32] developed algorithms to handle the problem of best-arm identification in the SRRB setting. However, to the best of our knowledge, there has not been any analysis to show if Thompson sampling-like MAB algorithms [25, 3] can provide good performance over specific instances of the SRRB scenario, and which modifications should be introduced to better deal with a larger class of problems in the SRRB setting. Indeed, while in generic non-stationary bandits, they have been proven to perform poorly [46], the strong regularity provided by the classical SRRB assumptions (increasing and concave expected reward function) suggest that, in specific cases, they might have *sublinear* regret.

Original Contributions. In this paper, we analyze the regret guarantees of a set of algorithms based on the original idea of Thompson Sampling (TS) when applied to SRRBs, specifically we present:

- A regret analysis of the `Beta-TS` (Thompson Sampling with Beta priors) when it is applied to the SRRB rested setting, providing a distribution-dependent regret bound based on the total-variation distance between specifically defined Poisson-Binomial and Binomial distributions. Here, the proof relies on completely novel techniques (Lemma 4.1) that can be of independent interest (Section 4).
- A natural extension of Thompson Sampling with Gaussian priors, γ-`GTS`, with near-optimal instance-independent regret bounds for the general stationary subgaussian environments that allow the analysis of the SRB setting in more general settings. This enables the retrieval of a theoretically and empirically superior algorithm for the setting providing, in fact, under some weak assumptions, sublinear instance-independent regret upper bounds (Section 5).
- A comparison of the proposed methods with the `R-ed-UCB` [31], designed for SRRB settings, over two SRRB instances to highlight their advantages and disadvantages (Section 6);
- The application of a sliding window approach to the two above algorithms, resulting in `Beta-SWTS` and γ-`SWGTS`, to cope with cases having limited learning horizon T (Section 7);
- Some numerical simulations to compare the performances in terms of regret of the proposed algorithms w.r.t. the ones designed for the SRRB setting (Section 8).

The proofs of the results are reported in Appendix A.

2 Related Works

Restless Bandits. The seminal work by [6] proposed the `UCB1` algorithm, based on the optimism in the face of the uncertainty principle, and shows that it provides a $O(\log(T))$ regret in stochastic stationary MAB settings. Instead, `TS` was originally designed as a heuristics for sequential decision-making [44], while only in the past decade has it been analysed theoretically by [25, 3]. These works provided a finite time analysis for TS showing an asymptotic bound on the regret of order $O(\log(T))$ for stochastic stationary MAB. Even if they are order-optimal in the stationary case, it has been shown in multiple cases that their effectiveness in other restless [19, 46] or adversarial settings [13] they provide poor performances in terms of regret.

Lately, `UCB1` and `TS` algorithms inspired the development of techniques to tackle the rising complexities of restless MAB settings. The main idea behind these newly designed algorithms is to forget past observations, removing samples from the statistics of the arms reward. Two different approaches are available to do that: passive and active. The former iteratively discards the information coming from the far past, making decisions on the most recent samples coming from the arms pulled by the algorithms. Examples of such a family of algorithms are DUCB [19], Discounted TS [34, 35], SW-UCB [19], and SW-TS [46]. Instead, the latter class of algorithms uses change-detection techniques [7] to decide when it is the case to discard old samples. This occurs when a sufficiently large change affects the arms' expected rewards. Among the active approaches we mention CUSUM-UCB [29], REXP3 [8], GLR-klUCB [9], and BR-MAB [36]. The development of such algorithms was required since applying the classical ones has proven to fail when the environment is restless (for both abrupt and smoothly changing).

Rising Bandits. Rising Bandits are a specific instance of either the Rested Bandits or the Restless Bandits in which the expected reward of an arm increases respectively according to the number of times it has been pulled or according to the time a certain arm has been pulled. The problem of SRRB in its deterministic flavor has been proposed by [20]. They designed algorithms with provably optimal policy regret bounds, and, in the case in which the rewards are increasing and concave, they

[2] With the $\tilde{O}(\cdot)$ notation we disregard logarithmic factors w.r.t. the learning horizon T.

give an algorithm whose policy regret is sublinear. Instead, the stochastic version of SRRB has been studied from a regret minimization perspective by [31]. In this work, the authors provide worst-case bounds for the regret of the order of $\tilde{O}(T^{\frac{2}{3}} + \Upsilon(T))$, where $\Upsilon(T)$ is a problem-dependent quantity which characterizes the growth rate of the arm expected rewards, for specifically designed algorithms for the rested and restless cases, namely R-ed-UCB and R-less-UCB, respectively. Both algorithms are based on the combination of specifically crafted upper confidence bounds to consider the increase of the expected reward functions and the use of a sliding window over the available samples. Finally, the SRRB has also been studied in a Best Arm Identification framework by [32], where the authors propose the R-UCBE and R-SR algorithm, a UCB-inspired and successive elimination approaches, respectively, that provide guarantees for the fixed budget version of the Best-Arm Identification of SRRB. Finally, a specific instance of the non-stationary bandits closely related to SRRB is provided by the rotting bandits [40, 27]. Unlike the SRRB setting, the expected payoff for a given arm decreases over time. Even in this case, the authors propose specifically crafted algorithms to address the peculiar structure of the problem and derive theoretical guarantees on the regret. Howeverfor these algorithms there are no theoretical guarantees when applied in the SRRB setting(see [31]).

3 Problem Formulation

We consider an SRRB setting with stochastic rewards. Let $K \in \mathbb{N}$ be the number of arms. Every arm $i \in [\![K]\!] := \{1, \dots, K\}$ is associated with an expected reward $\mu_i : \mathbb{N} \to \mathbb{R}$, where $\mu_i(n)$ defines the expected reward of arm i when pulled for the n-th time with $n \in \mathbb{N}$. As common in the rising bandit literature, the expected reward function $\mu_i(n)$ is non-decreasing and concave, as follows:[3]

Assumption 3.1 (Rising)**.** *For every arm $i \in [\![K]\!]$ and number of pulls $n \in \mathbb{N}$, let $\gamma_i(n) := \mu_i(n+1) - \mu_i(n)$ be the increment function, it holds that:*

$$\textit{Non-decreasing}: \quad \gamma_i(n) \geqslant 0, \qquad\qquad \textit{Concave}: \quad \gamma_i(n+1) - \gamma_i(n) \leqslant 0. \tag{1}$$

The learning process occurs over $T \in \mathbb{N}$ rounds, where T is called the learning horizon. At every round $t \in [\![T]\!]$, the agent pulls an arm $I_t \in [\![K]\!]$ and observes a random reward $X_t \sim \nu_{I_t}(N_{I_t,t})$, where for every arm $i \in [\![K]\!]$, we have that $\nu_i(N_{i,t})$ is a probability distribution[4] depending on the current number of pulls up to round t $N_{i,t} := \sum_{l=1}^{t} \mathbb{1}\{I_l = i\}$ whose expected value is given by $\mu_i(N_{i,t}) \in [0,1]$. For every arm $i \in [\![K]\!]$ and round $t \in [\![T]\!]$, we define the *average expected reward* as:

$$\overline{\mu}_i(t) := \frac{1}{t} \sum_{l=1}^{t} \mu_i(l). \tag{2}$$

As previously shown by [20], the optimal policy constantly plays the arm with the maximum average expected reward computed at the end of the learning horizon T. We denote with $i^*(T) := \arg\max_{i \in [\![K]\!]} \overline{\mu}_i(T)$ the (unique) optimal arm, that depends on the value of the horizon T.

Suboptimality Gaps. For analysis purposes, we introduce, for every suboptimal arm $i \neq i^*(T)$ and every number of pulls $n, n' \in [\![T]\!]$, the following notions of suboptimality gaps of the expected reward $\Delta_i(n, n')$ and average expected reward $\overline{\Delta}_i(n, n')$, formally:

$$\Delta_i(n, n') := \max\{0, \mu_{i^*(T)}(n) - \mu_i(n')\}, \qquad \overline{\Delta}_i(n, n') := \max\{0, \overline{\mu}_{i^*(T)}(n) - \overline{\mu}_i(n')\}, \tag{3}$$

respectively. Notice that we might have that for specific rounds, the gaps w.r.t. the optimal arm $\overline{\mu}_{i^*(T)}(n) - \overline{\mu}_i(n')$ may be negative, however, at the end of the learning horizon T the assumption that there is a unique optimal arm implies that $\overline{\mu}_{i^*(T)}(T) > \overline{\mu}_i(n')$ for all $n' \in [\![T]\!]$. This property allows defining, for every suboptimal arm $i \neq i^*(T)$, the minimum number of pulls of the optimal arm needed so that the optimal arm average expected reward gap for arm i is positive, formally:

$$\sigma_i(T) := \min\{l \in [\![T]\!] : \overline{\mu}_{i^*(T)}(l) > \overline{\mu}_i(T)\}, \qquad \sigma(T) := \max_{i \neq i^*(T)} \sigma_i(T). \tag{4}$$

Regret. The goal of an algorithm $\mathfrak{A}$ in an SRRB is to minimize the *expected cumulative regret* (see also [20] and Theorem 4.1 in [31]):

$$R(\mathfrak{A}, T) := T\, \overline{\mu}_{i^*(T)}(T) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}(N_{I_t,t})\right], \tag{5}$$

where the expectation is w.r.t. the randomness of the rewards and the possible randomness of $\mathfrak{A}$.

[3] For the generic integers $a, b \in \mathbb{N}, a < b$, we denote with $[\![a]\!]$ the set $\{1, \dots, a\}$ and $[\![a, b]\!]$ the set $\{a, \dots, b\}$.

[4] In the following, we will analyze algorithms for Bernoulli and subgaussian distributions.

Algorithm 1 `Beta-TS` Algorithm

Input: Number of arms K, Time horizon T
Set $\alpha_{i,1} \leftarrow 1$ for each $i \in [\![K]\!]$
Set $\beta_{i,1} \leftarrow 1$ for each $i \in [\![K]\!]$
Set $\nu_{i,1} \leftarrow Beta(\alpha_{i,1}, \beta_{i,1})$ for each $i \in [\![K]\!]$
for $t \in [\![T]\!]$ **do**
  Sample $\theta_{i,t} \sim \nu_{i,t}$ for each $i \in [\![K]\!]$
  Select $I_t \in \arg\max_{i \in [\![K]\!]} \theta_{i,t}$
  Pull arm I_t
  Collect reward X_t
  Update $\nu_{I_t,t+1} \leftarrow \text{Beta}(\alpha_{i,t} + X_t, \beta_{i,t} + 1 - X_t)$
  Update $\nu_{i,t+1} \leftarrow \nu_{i,t}$ for each $i \in [\![K]\!] \setminus \{I_t\}$
end for

Algorithm 2 γ-`GTS` Algorithm

Input: Number of arms K, Time horizon T, exploration parameter γ
Play every arm once and collect reward X_t
Set $N_{i,t} \leftarrow 1$, $\hat{\mu}_{i,t} \leftarrow X_t$, $\widehat{\overline{\mu}}_{i,t} \leftarrow \hat{\mu}_{i,t}$ for each $i \in [\![K]\!]$
Set $\nu_{i,t} \leftarrow \mathcal{N}(\widehat{\overline{\mu}}_{i,t}, \frac{1}{\gamma})$ for each $i \in [\![K]\!]$
for $t \in [\![T]\!]$ **do**
  Sample $\theta_{i,t} \sim \nu_{i,t}$ for each $i \in [\![K]\!]$
  Select $I_t \in \arg\max_{i \in [\![K]\!]} \theta_{i,t}$
  Pull arm I_t
  Collect reward X_t
  $N_{I_t,t} \leftarrow N_{I_t,t} + 1$, $\hat{\mu}_{I_t,t} \leftarrow \hat{\mu}_{I_t,t} + X_t$, $\widehat{\overline{\mu}}_{I_t,t} \leftarrow \frac{\hat{\mu}_{I_t,t}}{N_{I_t,t}}$
  Update $\nu_{I_t,t+1} \leftarrow \mathcal{N}(\widehat{\overline{\mu}}_{I_t,t}, \frac{1}{\gamma N_{I_t,t}})$
  Update $\nu_{i,t+1} \leftarrow \nu_{i,t}$ for each $i \in [\![K]\!] \setminus \{I_t\}$
end for

Environment Assumptions. In the following, we provide analyses of TS-like algorithms for the SRRB class. However, we will provide more explicit regret guarantees under the following:

Assumption 3.2. *There exists finite $T^* < +\infty$, s.t. it exists an arm $i^* \in [\![K]\!]$ for which for each $i \in [\![K]\!] \setminus \{i^*\}$ we have $\overline{\mu}_{i^*}(T^*) > \overline{\mu}_i(+\infty)$.*

The intuition behind this assumption is that after T^* a single arm is optimal, and it remains optimal. In particular, we have that for time horizons $T > T^*$, the optimal arm is fixed, i.e., the function $i^*(T)$ becomes constant in T, precisely equal to i^*. This assumption also implies that the gaps $\Delta_i(T,T)$ are lower bounded as shown in Lemma A.2 (provided in the appendix); this will be crucial to derive in what follows the instance-independent bounds. This is a mild assumption since the environments that comply with this condition arise naturally. For instance, the condition also holds on those instances in which the rate of convergence does not depend on the time horizon, e.g., the decay rates in physics having $\mu_i(n) = c_i(1 - e^{-\lambda_i n})$, with $c_i > 0$, and $\lambda_i > 0$. Other examples in which the assumption holds are the environments generated for the experiments in the seminal paper by [31] and those by [32]. An example of such a setting is provided in Figure 1. Moreover, Assumption 3.2 does not prevent the expected reward of the arms from crossing each other but only for the mean average $\overline{\mu}_i$ of the suboptimal arms to be smaller than the optimal one $\overline{\mu}_{i^*}$.

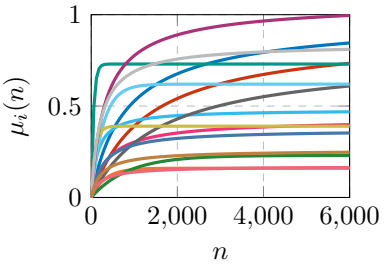


Figure 1: Example of the reward corresponding to the first $6,000$ pulls for a SRRB setting over 15 arms.

4 Analysis of Thompson Sampling with Beta priors (`Beta-TS`)

In this section, we provide the analysis of `Beta-TS`, i.e., Thompson Sampling instanced with Beta priors for the SRRB setting. The corresponding pseudocode is presented in Algorithm 1. For this algorithm, we will assume that the rewards are Bernoulli random variables.[5]

Assumption 4.1 (Bernoulli Rewards)**.** *For every $n \in [\![T]\!]$ and $i \in [\![K]\!]$, the reward $X \sim \nu_i(n)$ is Bernoulli distributed.*

Specifically, the `Beta-TS` algorithm initializes a beta distribution $\nu_{i,1} = \text{Beta}(\alpha_{i,1}, \beta_{i,1})$ with $\alpha_{i,1} = \beta_{i,1} = 1$ for each of the arms $i \in [\![K]\!]$ as prior. Then, for every round $t \in [\![T]\!]$ it collects one sample from each of the posterior distribution $\theta_{i,t} \sim \nu_{i,t}$ and plays the arm with the highest sample value, i.e., $I_t \in \arg\max_{i \in [\![K]\!]} \theta_{i,t}$. Then, based on the collected sample X_t, it updates the posterior distribution of the played arm $\nu_{I_t,t+1} = \text{Beta}(\alpha_{i,t} + X_t, \beta_{i,t} + 1 - X_t)$.

[5] If the reward X is not Bernoulli but X lies in $[0,1]$, we can sample a Bernoulli random variables with parameter X and use that sample for the update.

Regret Analysis. Compared to the standard case, the Beta-TS analysis in the SRRB setting poses additional challenges mainly because the samples collected from each of the arms are obtained from Bernoulli distributions with different parameters $\mu_i(n)$ since the arm expected reward changes as the arm is pulled. Consequently, unlike standard Beta-TS in the classical MABs setting, the sum of the rewards no longer represents a binomial distribution, but, due to the expected reward changes, the resulting random variable is a Poisson-Binomial [47]. This implies from a technical perspective that the standard analysis of Beta-TS [3] cannot be applied. Conversely, the following technical lemma was derived to deal with the complex structure of the cumulative reward distribution. We report it since it represents a significant technical novelty and can be of independent interest to the reader.

First, similarly to what has been provided by [3] in Definition 2.7, let us define $p_{i,t}$, for a random variable X whose number of successes is described by either a Poisson-Binomial or a Binomial distribution, for any $y_i \in (0,1)$, as:

$$p_{i,t} := \Pr\left(\text{Beta}\left(S_{1,t}+1, F_{1,t}+1\right) > y_i | \mathcal{F}_{t-1}\right), \tag{6}$$

where $S_{1,t}$, and $F_{1,t} = N_{1,t} - S_{1,t}$ are the number of successes, and failures of X, respectively, and $\mathcal{F}_{t-1}$ is the filtration of the history up to time $t-1$. In our framework, $\frac{1}{p_{i,t}}$ is related to the expected number of pulls of the suboptimal arm between two consecutive pulls of the best arm.

Lemma 4.1 (Technical Lemma)**.** *Let PB$(\underline{\mu}_1(j))$ be a Poisson-Binomial distribution with individual means $\underline{\mu}_1(j) = (\mu_1(1), \dots, \mu_1(j))$, and Bin$(j,x)$ be a binomial distribution with an arbitrary number j of trials and probability of success $x \leqslant \overline{\mu}_1(j)$. For any $N_{1,t} = j$ and $y_i \in (0,1)$, it holds that:*

$$\mathbb{E}_{S_{1,t} \sim PB(\underline{\mu}_1(j))}\left[\frac{1}{p_{i,t}} | N_{1,t} = j\right] \leqslant \mathbb{E}_{S_{1,t} \sim Bin(j, \overline{\mu}_1(j))}\left[\frac{1}{p_{i,t}} | N_{1,t} = j\right] \leqslant \mathbb{E}_{S_{1,t} \sim Bin(j,x)}\left[\frac{1}{p_{i,t}} | N_{1,t} = j\right].$$

The result is derived by showing the discrete log-convexity of the quantity $1/p_{i,t}$ and relying on order-statistics arguments to sort the expected value for different stochastic processes. This theorem states how $\mathbb{E}[\frac{1}{p_{i,t}}]$ in our nonstationary setting can be bounded by its value in a stationary setting. We are now ready to formalize the regret upper bound for Beta-TS.

Theorem 4.2 (Beta-TS - Regret Bound)**.** *Let $\sigma \in [\![\sigma(T), T]\!]$, with $\sigma(T)$ defined as in Equation 4. Under Assumption 4.1, for every $\epsilon \in (0,1)$, the Beta-TS algorithm suffers an expected cumulative regret bounded as:*

$$R(\textit{Beta-TS}, T) \leqslant O\Bigg(\sum_{i \neq i^*(T)} \Delta_i(T,0) \Big((1+\epsilon) \frac{\log(T)}{d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))} + \frac{1}{\epsilon^2} + \sum_{j=1}^{\sigma-1} \frac{1}{(1-\overline{\mu}_1(\sigma))^{j+1}} \delta_{TV}(PB(\underline{\mu}_1(j)), Bin(j, \overline{\mu}_1(\sigma_i))) \Big) \Bigg), \tag{7}$$

where $d(x,y) := x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$ for $x, y \in [0,1]$ is the Kullback-Leibler divergence between Bernoulli distributions, $\delta_{TV}(P,Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$ denotes the total variation divergence between distributions P and Q, PB$(\underline{\mu}_1(j))$ denotes the Poisson-Binomial distribution with individual means $\underline{\mu}_1(j) = (\mu_1(1), \dots, \mu_1(j))$, and Bin$(j,x)$ denotes the binomial with j trials and parameter x.

First, we observe that the regret bound reduces to that of standard Beta-TS of [3] when facing standard MABs. Indeed, in such a case, the Poisson-Binomial distribution reduces to the Binomial distribution, and consequently, the TV term vanishes:

$$R(\texttt{Beta-TS}, T) \leqslant O\left(\sum_{i \neq i^*} \Delta_i \left(\frac{(1+\epsilon)\log(T)}{d(\mu_i, \mu_1)} + \frac{1}{\epsilon^2} \right) \right).$$

Conversely, for the general case, the rested nature of the problem induces the presence of an additional term composed by the summation of the total variation distances $\delta_{\text{TV}}(\text{PB}(\underline{\mu}_1(j)), \text{Bin}(j, \overline{\mu}_1(\sigma_i))$. This term originates from a *change of measure* argument used in the analysis. Finally, let us state a corollary that shows that using Assumption 3.2, we have results on worst-case regret:

Corollary 4.3. *Under Assumption 3.2, the Beta-TS algorithm suffers an expected cumulative regret:*

$$R(\textit{Beta-TS}, T) \leqslant \begin{cases} O(\sqrt{KT\log(T)} + K\sigma(1-\overline{\mu}_1(T))^{-\sigma}) & \text{if } T \leqslant T^* \\ O(\sqrt{KT\log(T)}) & \text{if } T > T^* \end{cases}. \tag{8}$$

Notice that the above results do not require the knowledge of T^* by Beta-TS.

5 Analysis of γ-Thompson Sampling with Gaussian priors (γ-GTS)

In this section, we provide an analysis of the γ-Thompson Sampling with Gaussian priors (γ-GTS) algorithm, a modification of the classical TS with Gaussian priors, that provides instance-independent optimal regret bounds for the subgaussian standard MABs. In this case, we assume that the reward has subgaussian distribution with positive realizations.[6] Formally:

Assumption 5.1 (Non-negative Subgaussian rewards)**.** *For every $n \in [\![T]\!]$ and arm $i \in [\![K]\!]$, the reward $X \sim \nu_i(n)$ is non-negative almost surely, and σ^2_{var}-subgaussian with finite mean.*

The pseudocode for γ-GTS is presented in Algorithm 2. At first, the γ-GTS algorithm pulls each arm once. Using the collected reward, it initializes the priors $\nu_{i,t}$ for all the arms, setting all the variances equal to $1/\gamma$. Then, for every round $t \in [\![T]\!]$, it collects one sample from each of the posterior distribution $\theta_{i,t} \sim \nu_{i,t}$ and plays the arm with the highest sample value, i.e., $I_t \in \arg\max_{i \in [\![K]\!]} \theta_{i,t}$. Based on the collected sample X_t, it updates the posterior distribution of the played arm $\nu_{I_t,t+1}$.

Regret Analysis. The following theorem provides a bound on the γ-GTS algorithm regret.

Theorem 5.1 (γ-GTS - Regret Bound for Subgaussian SRB)**.** *Let $\sigma \in [\![\sigma(T), T]\!]$ with $\sigma(T)$ defined as in Equation 4. Under Assumption 5.1, setting $\gamma \leqslant \min\left\{\frac{1}{4\sigma^2_{var}}, 1\right\}$, the γ-GTS algorithm suffers an expected cumulative regret of:*

$$R(\gamma\text{-GTS}, T) \leqslant O\left(\sum_{i \neq i^*(T)} \Delta_i(T,0)\left(\frac{\log(T\overline{\Delta}_i(\sigma,T)^2 + e^6)}{\gamma\overline{\Delta}_i(\sigma,T)^2} + \frac{\sigma^2_{var}}{\overline{\Delta}_i(\sigma,T)^2} + \sum_{j=1}^{\sigma-1} \frac{\delta_{TV}(\mathbb{P}_j, \mathbb{Q}_j(\overline{\mu}_1(\sigma)))}{\textbf{\textit{erfc}}(\sqrt{\frac{\gamma j}{2}}(\overline{\mu}_1(\sigma)))}\right)\right),$$

where ***erfc***$(\cdot)$ *is the complementary error function, $\mathbb{P}_j$ is the distribution of the sample mean of the first j samples collected from arm 1, while $\mathbb{Q}_j(y)$ is the distribution of the sample mean of j samples collected from* any *σ^2_{var}-subgaussian distribution with mean y.*

Note that the structure of the regret resembles the one of the Beta-TS. Indeed, we have a summation of total variation divergences, where the place of the binomial distribution is taken by $\mathbb{Q}_j$, a distribution that is arbitrary as long as it is σ^2_{var}-subgaussian. Such a distance (proportional to σ and the total variation) between the real process and a stationary one that would suffer near-optimal regret encodes the additional complexity of the problem w.r.t. the classical stochastic MAB problem. Even in this case, we observe that also the regret bound of γ-GTS reduces to that of Thompson Sampling of [3] when facing standard MABs:

$$R(\gamma\text{-GTS}, T) \leqslant O\left(\sum_{i \neq i^*} \left(\frac{\log(T\Delta_i^2)}{\gamma\Delta_i} + \frac{\sigma^2_{\text{var}}}{\Delta_i}\right)\right),$$

with $\gamma = O(\min\{\sigma^2_{\text{var}}, 1\})$. Note that our result is more general than the one by [3]. Indeed, it also holds for arbitrary subgaussian rewards, while their results hold only if they have mean in $[0,1]$. Finally, similarly to what has been provided for Beta-TS, relying on Assumption 3.2, we have:

Corollary 5.2. *Under Assumption 3.2, the γ-GTS algorithm suffers an expected cumulative regret:*

$$R(\gamma\text{-GTS}, T) \leqslant \begin{cases} O(\sqrt{KT\gamma^{-1}\log(T)} + K\sigma e^{\gamma\sigma\overline{\mu}_1(\sigma)^2}) & \text{if } T \leqslant T^* \\ O(\sqrt{KT\gamma^{-1}\log(T)}) & \text{if } T > T^* \end{cases}. \tag{9}$$

Even in this case, the result is achieved by an algorithm not requiring the knowledge of T^*.

The following theorem provides an alternative way of setting the parameter γ if we have information about the time horizon T. Notice that the theoretical result in this case does not even require the expected reward to satisfy Assumption 3.2.

Theorem 5.3 (γ-GTS - Regret Bound for Subgaussian SRB γ-tuned)**.** *Let $\sigma \in [\![\sigma(T), T]\!]$ with $\sigma(T)$ defined as in Equation 4, let furthermore $\sigma \sim T^\beta$ and $\gamma \sim T^{-\alpha}$. Under Assumption 5.1, for every $\alpha \geqslant \beta$:*

$$R(\gamma\text{-GTS}, T) \leqslant O\left(\sum_{i \neq i^*(T)} \Delta_i(T,0)\left(\frac{T^\alpha \log(T\overline{\Delta}_i(\sigma,T)^2 + e^6)}{\overline{\Delta}_i(\sigma,T)^2} + \frac{\sigma^2_{var}}{\overline{\Delta}_i(\sigma,T)^2} + \sigma\right)\right). \tag{10}$$

[6] For the sake of presentation, we assume to have positive realizations. W.l.o.g. one might also consider realizations bounded from below by a given constant.

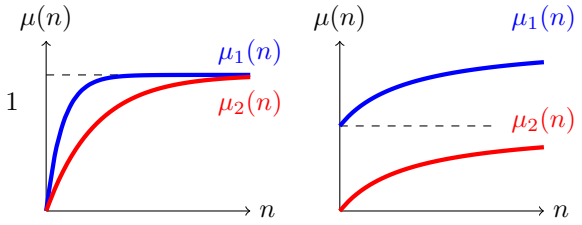


Figure 2: Different environments for the SRRB problem.

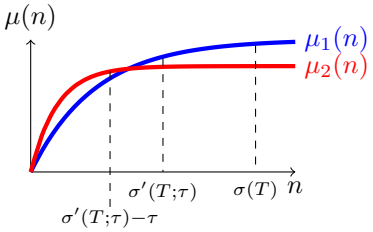


Figure 3: Visual representation of $\sigma'(T;\tau)$, a point in which $\overline{\mu}_{i*(T)}(\sigma'(T;\tau),\tau) > \mu_i(T)$.

Conversely, if Assumption 3.2, holds we have:

Corollary 5.4. *Under assumption 3.2 γ-GTS with γ tuned suffer an instance independent regret bound upper bounded by (for all T):*

$$R(\gamma\text{-GTS},T) \leqslant O\left(T^{\frac{1+\alpha}{2}}\sqrt{K\log(T)} + KT^{\alpha}\right). \tag{11}$$

6 Comparison with the R-ed-UCB Algorithm

We now provide and analyze two SRRB instances to highlight the advantages and disadvantages of the proposed algorithms when compared with the optimistic algorithm R-ed-UCB [31] designed for SRRB settings. We recall the regret result provided by [31] for R-ed-UCB:

Theorem 6.1 (Theorem 4.4, [31])**.** *R-ed-UCB with a suitable exploration index (see [31]) $\alpha > 2$, and $\epsilon \in (0,1/2)$ suffers an expected regret for every $q \in [0,1]$ bounded as:*

$$R(\text{R-ed-UCB},T) \leqslant O\left(\frac{K}{\epsilon}T^{\frac{2}{3}}(\alpha\log T)^{\frac{1}{3}} + \frac{KT^q}{1-2\epsilon}\Upsilon\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil, q\right)\right), \tag{12}$$

where $\Upsilon(M,q) := \sum_{l=1}^{M-1}\max_{i\in[\![K]\!]}\{\gamma_i(l)^q\}$ is a complexity index depending on the expected rewards.

First Instance. We start with an instance in which R-ed-UCB succeeds in delivering a sublinear regret, while our algorithms may fail. We define the expected reward functions as follows: $\mu_1(n) = 1 - e^{-\lambda n}$ and $\mu_2(n) = 1 - e^{-2\lambda n}$, where $\lambda > 0$ is an arbitrary parameter (Figure 2a). Notice that 1 is the optimal arm. Thus, this instance violates Assumption 3.2 since there is no possibility to define a $T^* < +\infty$ so that $\overline{\mu}_1(T^*) > \overline{\mu}_2(+\infty)$. Therefore, we cannot guarantee that our algorithms provide a sublinear regret. Conversely, by using the definition of Υ, for $q \in [0,1]$, we have:

$$\Upsilon\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil, q\right) = \sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\max_{y\in\{1,2\}}\{e^{-y\lambda n} - e^{-y\lambda(n+1)}\}^q \leqslant \left(1+\frac{2}{q}\right)e^{-q\lambda},$$

which implies that Theorem 6.1 provides a regret of order $O(T^{2/3} + T^q/q)$ for the R-ed-UCB algorithm which is sublinear for every $q < 1$ and, selecting $q = 1/\log T$ we obtain the best rate $O(T^{2/3} + \log T)$.

Second Instance. The second instance is designed so that our algorithms provide sublinear regret, while R-ed-UCB fails. We define the expected reward functions as: $\mu_1(n) = 1 - \frac{2^{\lambda-1}}{(t+1)^{\lambda}}$, and $\mu_2(n) = \frac{1}{2} - \frac{2^{\lambda-1}}{(t+1)^{\lambda}}$, where $\lambda \in [0,1]$ is an arbitrary parameter (Figure 2b). Assumption 3.2 holds with $T^* = 1$, since the optimal arm has larger expected reward $\mu_1(1) > \overline{\mu}_2(T)$ starting from the initial pull, regardless of T. The Υ factor of this setting for every $q \in [0,1]$ is given by:

$$\Upsilon\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil, q\right) = \sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\left(\frac{2^{\lambda-1}}{(n+1)^{\lambda}} - \frac{2^{\lambda-1}}{(n+2)^{\lambda}}\right)^q \geqslant O\begin{cases}\lambda^q & \text{if } q(\lambda+1) > 1\\ \lambda^{\frac{1}{\lambda+1}}\log T & \text{if } q(\lambda+1) = 1\\ \lambda^q T^{1-q(\lambda+1)} & \text{otherwise}\end{cases}.$$

We prove in Appendix D, that the optimal choice of q is $1/(\lambda+1)$, according to Theorem 6.1, this leads to a regret of order $O(T^{2/3} + (\lambda T)^{\frac{1}{\lambda+1}})$ for R-ed-UCB. Thus, for instance, choosing $\lambda = 1/3$, R-ed-UCB attains a regret bound of order $O(T^{3/4})$ while our approaches succeed to achieve $O(\sqrt{T})$ regret.

7 Analysis of Sliding Window Thompson Sampling Approaches

The instance-dependent bounds provided above in Theorems 4.2 and 5.1 confirm that the classical algorithms require a large number of pulls to converge to pulling the optimal arm at T. This is due to the fact that if $\overline{\mu}_{i*(T)}(n)$ does not converge fast enough to $\overline{\mu}_{i*(T)}(T)$ they may suffer linear regret. However, if the arms' expected reward functions are regular enough and the time horizon is sufficient ($T \geqslant T^*$), as encoded Assumption 3.2, the natural exploration induced by TS can retrieve optimal regret bounds for the problem. The main drawback of the previously presented approach is that they use *all the samples from the beginning of learning* for estimating the average expected reward. Intuition suggests that, as already seen with γ-GTS algorithm, in some cases, it might be convenient to forget the past and focus on the most recent samples only. In this section, we make use of a *sliding window* approach that was already employed by R-ed-UCB [31] for SRRBs.

Preliminaries. We extend the definitions of Section 3 to account for a sliding window. For every arm $i \in [\![K]\!]$, round $t \in [\![T]\!]$, and window size $\tau \in [\![t]\!]$, we define the *windowed average expected reward* as $\overline{\mu}_i(t;\tau) := \frac{1}{\tau}\sum_{l=t-\tau+1}^{t} \mu_i(l)$. Furthermore, we define the minimum number of pulls needed so that the optimal arm $i^*(T)$ can be identified as optimal in a window of size τ:

$$\sigma'_i(T;\tau) := \min\{\{l \in [\![T]\!] : \overline{\mu}_{i*(T)}(l;\tau) > \mu_i(T)\} \cup \{+\infty\}\}, \qquad \sigma'(T;\tau) := \max_{i \neq i*(T)} \sigma'_i(T;\tau). \quad (13)$$

These definitions resemble those of Equation (4). However, the comparison here involves the windowed average expected reward of the optimal arm $\overline{\mu}_{i*(T)}(l;\tau)$ compared against the expected reward (not averaged) of the other arms at the end of the learning horizon $\mu_i(T)$. Thus, even for $\tau = T$, we have a stronger requirement since $\sigma'(T;T) \geqslant \sigma(T)$. Furthermore, for some values of τ, a value of the number of pulls l so that $\overline{\mu}_{i*(T)}(l;\tau) > \mu_i(T)$ might not exist. In such a case, we set $\sigma'_i(T;\tau)$ (and thus $\sigma'(T;\tau)$) to $+\infty$. Nevertheless, as visible in Figure 3, in some cases $\sigma'(T;\tau) \ll \sigma(T)$, making the sliding window-based approaches convenient. Finally, we introduce a new definition of suboptimality gaps: $\Delta'_i(T;\tau) := \overline{\mu}_{i*(T)}(\sigma'(T;\tau)) - \mu_i(T)$ for every arm $i \in [\![K]\!]$.

Algorithms. In this section, we analyze the sliding window versions of TS for Bernoulli, namely Beta-SWTS, and Gaussian, namely γ-SW-GTS. Their pseudo-codes are reported in the Appendix. The following results provide the regret upper bounds achieved by these algorithms as a function τ.

Theorem 7.1 (Beta-SWTS Regret Bound)**.** *Under Assumption 4.1, the Beta-SWTS algorithm suffers an expected cumulative regret bounded as:*

$$R(\textit{Beta-SWTS}, T) \leqslant O\left(\sum_{i \neq i*(T)} \Delta_i(T,0)\left(\frac{T\log(T)}{\tau(\Delta'_i(T;\tau))^3} + \frac{\sigma'(T;\tau)}{(1-\overline{\mu}_1(\sigma'(T;\tau),\tau))^{\tau+1}}\right)\right). \quad (14)$$

Theorem 7.2 (γ-SW-GTS Regret Bound)**.** *Under Assumption 5.1, setting $\gamma \leqslant \min\left\{\frac{1}{4\sigma^2_{var}}, 1\right\}$, the γ-GTS algorithm suffers an expected cumulative regret of:*

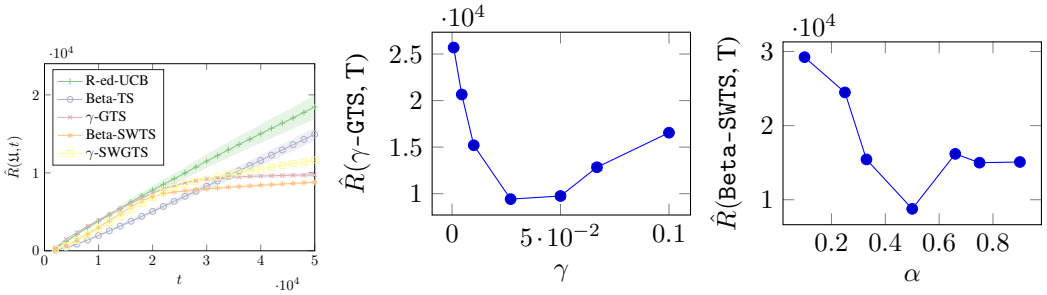
$$R(\gamma\textit{-SWGTS}, T) \leqslant O\left(\sum_{i \neq i*(T)} \Delta_i(T,0)\left(\frac{T\log(T(\Delta'_i(T;\tau))^2)}{\gamma\tau(\Delta'_i(T;\tau))^2} + \frac{T}{\tau} + \frac{\sigma'(T;\tau)}{\textbf{\textit{erfc}}\left(\sqrt{\frac{\gamma\tau}{2}}(\overline{\mu}_1(\sigma'(T;\tau),\tau))\right)}\right)\right). \quad (15)$$

Some comments are in order. First, the regret bounds are presented for a generic choice of the window size τ. The optimal choice depends on the instance-dependent quantities $\sigma'(T;\tau)$ and $\Delta'_i(T;\tau)$. Second, if we compare these regret bounds with those of the corresponding non-windowed versions, we observe that the last addendum derived from the change of distribution is of order $O(\sigma) = O(\sigma(T))$ in Theorems 4.2 and 5.1 and becomes $O(\sigma'(T;\tau))$ in Theorems 7.1 and 7.2. This quantifies the advantage of the sliding window algorithms in the cases in which $\sigma'(T;\tau) \ll \sigma(T)$. Finally, we remark that these regret bounds become vacuous when $\sigma'(T;\tau) = +\infty$.

8 Numerical Simulations

To back up the theoretical findings, we numerically test the Bayesian algorithms we developed against R-ed-UCB [31], which has been specifically designed for the rested setting.[7] We tested in the same

[7] We also tested for those baseline algorithms those considered in the experimental section of [31]. For the sake of presentation, we only report the figures with the full comparison in Appendix E.



(a) (b) (c)

Figure 4: Results 15-arms settings: (a) Average regret over the learning horizon T; (b) Regret for γ-GTS for different values of γ; (c) Regret for Beta-SWTS for different values of α, with $\tau = T^{\alpha}$.

15-arms experiments of [31]. A visual representation of the expected rewards is provided in Figure 1. The parameters selected for the experiments (that comply with the recommendation provided in the papers proposing them) and the parameters defining the setting are provided in Appendix E. We compare the algorithms in terms of empirical cumulative regret $\hat{R}(\cdot, t)$ averaged over 100 independent runs with the corresponding 95% confidence intervals over a time horizon of $T = 50,000$ rounds.

Results All the algorithms we presented outperform the baseline at the end of the time horizon T. In particular, while Beta-TS, γ-SWGTS, and Beta-SWTS are providing a lower regret than R-ed-UCB over the entire time horizon, while the γ-GTS is providing better results only for $t > 14,000$. However, there is statistical evidence for the superiority of our algorithms from that point on. Moreover, it seems that all the algorithms except Beta-TS can significantly slow down the increase in terms of regret in the rounds $t > 20,000$. This is due to the fact that the other algorithms have been modified to capture the properties of the setting, while Beta-TS is a general-purpose MAB algorithm.

We also run a sensitivity analysis on the γ parameter for γ-GTS and the sliding window τ for Beta-SWTS. We provide the value of their average regret $\hat{R}(\cdot, T)$ at the end of the time horizon T, where the average has been taken over 100 runs of the algorithms on the same 15-arms setting.[8]

Results This environment is the same 15-arm rested bandit generated for the first experiment. In both cases, the parameter providing the smallest regret is the one prescribed by the theoretical analysis, i.e., $\gamma \sim \log(T)T^{-\frac{1}{2}} =$ for γ-GTS and $\tau = T^{\frac{1}{2}}$ for SW-TS. The misspecification of the parameter leads to an increase in the regret by a factor of at most 3 over the analysed values of the parameters, especially for small values of the parameters. This suggests that the knowledge of the time horizon T significantly impacts the final performances of the algorithms, and, if the information about the time horizon is not known, one should set the parameter overestimating it.

9 Conclusions

In this paper, we investigated the properties of Thompson sampling-like algorithms for regret minimization in the setting of SRBs. We analyzed the TS algorithms with Beta and Gaussian priors. In both cases (Beta-TS and γ-GTS), we derived a general analysis that highlights the challenges of the setting, showing that the logarithmic regret is increased by a term that depends on the total variation distance between pairs of suitably defined distributions.We also analyzed what happens upon realistic assumptions on the suboptimality gaps that allow us to obtain order-optimal instance-independent regret guarantees. Furthermore, we derived regret bounds for the version of the two above-mentioned algorithms that use a sliding window, i.e., Beta-SWTS and γ-SWGTS. Since, to the best of our knowledge, the SRB setting currently lacks a lower bound analysis, future works should focus on this unavoidable step to understand the SRB setting fully.

[8] We reported the 95% interval bars which are not clearly visible due to their limited dimension.

References

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1968.

[2] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the Conference on learning theory (COLT)*, pages 39–1, 2012.

[3] Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM*, 64(5):1–24, 2017.

[4] R. Allesiardo, Raphaël Feraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3, 06 2017.

[5] John R Anderson and Lael J Schooler. Reflections of the environment in memory. *Psychological science*, 2(6):396–408, 1991.

[6] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[7] Michele Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. prentice Hall Englewood Cliffs, 1993.

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

[9] Lilian Besson and Emilie Kaufmann. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *HAL*, 2019(0), 2019.

[10] Philip J. Boland, Harshinder Singh, and Bojan Cukic. Stochastic orders in partition and random testing of software. *Journal of Applied Probability*, 39(3):555–565, 2002.

[11] Philip J. Boland, Harshinder Singh, and Bojan Cukic. The stochastic precedence ordering with applications in sampling and testing. *Journal of Applied Probability*, 41(1):73–82, 2004.

[12] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[13] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[14] Seok-Ho Chang, Pamela C. Cosman, and Laurence B. Milstein. Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

[15] Giulia Clerici, Pierre Laforgue, and Nicolo Cesa-Bianchi. Linear bandits with memory: from rotting to rising. *arXiv preprint arXiv:2302.08345*, 2023.

[16] Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 521–529, 2014.

[17] Werner Ehm. Binomial approximation to the poisson binomial distribution. *Statistics and Probability Letters*, 11(1):7–16, 1991.

[18] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the Conference On Learning Theory (COLT)*, pages 359–376, 2011.

[19] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the international conference on Algorithmic Learning Theory (ALT)*, pages 174–188, 2011.

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

[21] Theodore P Hill and Christian Houdré. *Advances in Stochastic Inequalities*, volume 234. American Mathematical Society, 1999.

[22] S.G Hoggar. Chromatic polynomials and logarithmic concavity. *Journal of Combinatorial Theory, Series B*, 16(3):248–254, 1974.

[23] Oliver Johnson and Christina Goldschmidt. Preservation of log-concavity on summation. *ESAIM: Probability and Statistics*, 10:206–215, April 2006.

[24] Boyan Jovanovic and Yaw Nyarko. A bayesian learning model fitted to a variety of empirical learning curves. *Brookings Papers on Economic Activity. Microeconomics*, 1995:247–305, 1995.

[25] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the international conference on Algorithmic Learning Theory (ALT)*, pages 199–213, 2012.

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

[27] Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

[28] Yang Li, Jiawei Jiang, Jinyang Gao, Yingxia Shao, Ce Zhang, and Bin Cui. Efficient automatic cash via rising bandits. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 34, pages 4763–4771, 2020.

[29] Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 32, 2018.

[30] Albert W Marshall, Ingram Olkin, and Barry C Arnold. Inequalities: Theory of majorization and its applications. *Springer Series in Statistics*, 2011.

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

[32] Marco Mussi, Alessandro Montenegro, Francesco Trovò, Marcello Restelli, and Alberto Maria Metelli. Best arm identification for stochastic rising bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[33] S-D Poisson. English translation of poisson's" recherches sur la probabilit\'e des jugements en mati\ere criminelle et en mati\ere civile"/" researches into the probabilities of judgements in criminal and civil cases". *arXiv preprint arXiv:1902.02782*, 2019.

[34] Han Qi, Yue Wang, and Li Zhu. Discounted thompson sampling for non-stationary bandit problems. *arXiv preprint arXiv:2305.10718*, 2023.

[35] Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

[36] Gerlando Re, Fabio Chiusano, Francesco Trovò, Diego Carrera, Giacomo Boracchi, and Marcello Restelli. Exploiting history data for nonstationary multi-armed bandit. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 51–66, 2021.

[37] Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

[38] Bero Roos. Binomial approximation to the poisson binomial distribution: The krawtchouk expansion. *Theory of Probability & Its Applications*, 45(2):258–272, 2001.

[39] S. M. Samuels. On the number of successes in independent trials. *The Annals of Mathematical Statistics*, 36(4):1272–1278, 1965.

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

[41] Lisa K Son and Rajiv Sethi. Metacognitive control and optimal learning. *Cognitive Science*, 30(4):759–774, 2006.

[42] Wenpin Tang and Fengmin Tang. The Poisson Binomial Distribution—Old and New. *Statistical Science*, 38(1):108 – 119, 2023.

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

[44] 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.

[45] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the ACM international conference on Knowledge discovery and data mining (SIGKDD)*, pages 847–855, 2013.

[46] Francesco Trovò, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

[47] Y. H. Wang. On the number of successes in independent trials. *Statistica Sinica*, 3(2):295–312, 1993.

[48] Jingxu Xu, Yuhang Wu, Yingfei Wang, Chu Wang, and Zeyu Zheng. Online experiments with diminishing marginal effects. *Available at SSRN 4640583*, 2023.

A Proofs and Derivations

In this appendix, we provide the complete proofs and derivations we have omitted in the main paper.

A.1 Further Definitions

Given the existence of T^* as defined in Assumption 3.2, it will be useful for the purpose of analysis to define (assuming $\max_{i\in[\![K]\!]}\{\Delta_i(+\infty,0)\}$ exists finite):

$$c=\frac{\max_{i\in[\![K]\!]}\{\Delta_i(+\infty,0)\}}{\min_{i\neq i^*(T'),T'\in[\![T^*-1]\!]}\{\overline{\Delta}_i(\sigma',T'),\overline{\Delta}_i(T^*,+\infty)\}}, \tag{16}$$

where $\sigma'\in[\![\sigma(T'),T']\!]$.

Lemma A.1 (Wald's Identity for Rising Bandits)**.** *For every algorithm $\mathfrak{A}$ and learning horizon $T\in\mathbb{N}$, it holds that:*

$$R(\mathfrak{A},T)\leqslant\sum_{i=2}^{K}\Delta_i(T,0)\,\mathbb{E}[N_{i,T}], \tag{17}$$

where $\Delta_i(T,0):=\mu_1(T)-\mu_i(0)$.

Proof. We start with the usual definition of regret and proceed as follows:

$$\begin{aligned}
R(\mathfrak{A},T)&=T\overline{\mu}_1-\mathbb{E}\left[\sum_{t=1}^{T}\mu_{I_t}(N_{I_t,t})\right] && (18)\\
&=\mathbb{E}\left[\sum_{t=1}^{T}(\mu_1(t)-\mu_{I_t}(N_{I_t,t}))\right] && (19)\\
&=\mathbb{E}\left[\sum_{t=1}^{T}\mu_1(t)-\sum_{j=0}^{N_{1,T}}\mu_1(j)-\sum_{i=2}^{K}\sum_{j=0}^{N_{i,T}}\mu_i(j)\right] && (20)\\
&=\mathbb{E}\left[\sum_{j=N_{1,T}+1}^{T}\mu_1(j)-\sum_{i=2}^{K}\sum_{j=0}^{N_{i,T}}\mu_i(j)\right] && (21)\\
&\leqslant\mathbb{E}\left[\sum_{i=2}^{K}\sum_{j=0}^{N_{i,T}}(\mu_1(T)-\mu_i(j))\right] && (22)\\
&\leqslant\sum_{i=2}^{K}(\mu_1(T)-\mu_i(0))\,\mathbb{E}\left[\sum_{j=0}^{N_{i,T}}1\right]. && (23)
\end{aligned}$$

□

Lemma A.2. *Assumption 3.2 entails the fact that $\overline{\Delta}_i(T,T)$ are lower bounded with:*

$$c=\frac{\max_{i\in[\![K]\!]}\{\Delta_i(+\infty,0)\}}{\min_{i\neq 1,T'\in[\![T^*-1]\!]}\{\overline{\Delta}_i(T',T'),\overline{\Delta}_i(T^*,+\infty)\}}.$$

Proof. Let us assume now it the arms dynamics are such that it exists a finite time horizon T^* defined as:

$$\overline{\mu}_1(T^*)>\overline{\mu}_i(+\infty),\forall i\neq 1, \tag{24}$$

where we used the fact that for $T>T^*$ the arm identified by the theorem is the optimal arm ($1=i^*$). Informally, there is a finite time over which the best arm will not change anymore, we can devise a finite grid of values for every T and every i of $\overline{\Delta}_i(T,T)$, up to T^*, for T^* we will consider $\overline{\Delta}_i(T^*,\infty)$. Indeed we notice that $\overline{\Delta}_i(T^*,\infty)$ is smaller than any possible permutation (as the arms' average is rising) for any horizon $T\geqslant T^*$, i.e by definition $\forall\ \epsilon,\ \eta\geqslant 0$:

$$\overline{\Delta}_i(T^*+\eta,T^*+\epsilon)\geqslant\overline{\Delta}_i(T^*,T^*+\epsilon)\geqslant\overline{\Delta}_i(T^*,\infty), \tag{25}$$

so any constant that were to bound $\frac{\max_{i\in[\![K]\!]}\{\Delta_i(+\infty,0)\}}{\overline{\Delta}_i(T^*,\infty)}$ would also bound all the infinite values $\frac{\max_{i\in[\![K]\!]}\{\Delta_i(+\infty,0)\}}{\overline{\Delta}_i(T^*+\eta,T^*+\epsilon)}$. Then c is well defined, and it is possible to define it as (assuming $\max_{i\in[\![K]\!]}\{\Delta_i(+\infty,0)\}$ exists finite):

$$c=\frac{\max_{i\in[\![K]\!]}\{\Delta_i(+\infty,0)\}}{\min_{\{i\neq 1,T'\in[\![T^*-1]\!]\}}\{\overline{\Delta}_i(T',T'),\overline{\Delta}_i(T^*,\infty)\}}, \tag{26}$$

□

A.2 Proofs of Section 4

Lemma 4.1 (Technical Lemma)**.** *Let PB($\underline{\mu}_1(j)$) be a Poisson-Binomial distribution with individual means $\underline{\mu}_1(j)=(\mu_1(1),\dots,\mu_1(j))$, and Bin($j,x$) be a binomial distribution with an arbitrary number j of trials and probability of success $x\leqslant\overline{\mu}_1(j)$. For any $N_{1,t}=j$ and $y_i\in(0,1)$, it holds that:*

$$\mathbb{E}_{S_{1,t}\sim PB(\underline{\mu}_1(j))}\left[\frac{1}{p_{i,t}}|N_{1,t}=j\right]\leqslant\mathbb{E}_{S_{1,t}\sim Bin(j,\overline{\mu}_1(j))}\left[\frac{1}{p_{i,t}}|N_{1,t}=j\right]\leqslant\mathbb{E}_{S_{1,t}\sim Bin(j,x)}\left[\frac{1}{p_{i,t}}|N_{1,t}=j\right].$$

Proof. Let $N_{1,t}=j$, $S_{1,t}=s$. Then, As shown by Agrawal et al. [2], $p_{i,t}$ can be written as:

$$p_{i,t}=\mathbb{P}(\theta_{1,t}>y_i)=F^B_{j+1,y_i}(s).$$

For ease of notation, let us denote $X'\sim\text{PB}(\underline{\mu}_1(j))$ and $X\sim\text{Bin}(j,x)$. We are now interested in finding if it does exist a number of trials j such that:

$$(**)=\mathbb{E}\left[\frac{1}{F^B_{j+1,y_i}(X')}\right]\leqslant\mathbb{E}\left[\frac{1}{F^B_{j+1,y_i}(X)}\right]=(*). \tag{27}$$

We notice that the PMF of a binomial distribution is discrete log-concave (see Lemma C.12), so that let Y be a binomial random variable, we will have:

$$p_Y(i+1)^2\geqslant p_Y(i)p_Y(i+2), \tag{28}$$

so by Lemma C.13 used with $\alpha=1$ and $r=\infty$, and q being the probability mass function of the binomial distribution (more in-depth $q(x)=p(-x)$ in Lemma C.13) we find that the CDF of the binomial is discrete log-concave on $\mathbb{Z}$ too (indeed the theorems cited above state that if the probability mass function of an integer-valued random variable is discrete log-concave as a function on $\mathbb{Z}$, then the corresponding CDF (F^B in our notation) is also discrete log-concave as a function on $\mathbb{Z}$) and so by definition, omitting superscripts and subscripts, $1/F$ is discrete log-convex (same inequality of 28 with different sign) on the set $S:=\{0,\dots j+1\}$ of all atoms of the distribution. So, $1/F$ is strictly discrete convex on S. Indeed by proving the discrete log-convexity of $\frac{1}{F}$ we have proved that:

$$\left(\frac{1}{F(x+1)}\right)^2\leqslant\frac{1}{F(x+2)}\frac{1}{F(x)}, \tag{29}$$

by applying the logarithm, we obtain the following:

$$2\log\left(\frac{1}{F(x+1)}\right)\leqslant\log\left(\frac{1}{F(x+2)}\frac{1}{F(x)}\right), \tag{30}$$

then, since the logarithm is monotonic, increasing:

$$\left(\frac{1}{F(x+1)}\right)\leqslant\left(\frac{1}{F(x+2)}\right)^{\frac{1}{2}}\left(\frac{1}{F(x)}\right)^{\frac{1}{2}}. \tag{31}$$

Using the AM-GM inequality, we obtain:

$$\left(\frac{1}{F(x+1)}\right)<\frac{1}{2}\left(\frac{1}{F(x+2)}\right)+\frac{1}{2}\left(\frac{1}{F(x)}\right) \tag{32}$$

$$2\left(\frac{1}{F(x+1)}\right)<\left(\frac{1}{F(x+2)}\right)+\left(\frac{1}{F(x)}\right), \tag{33}$$

where the inequality is strict since $\frac{1}{F(x+2)}<\frac{1}{F(x)}$ for $x\in\{0,\dots,j-1\}$, and, therefore, $\frac{1}{F(x+2)}\neq\frac{1}{F(x)}$. Using Lemma C.11 we obtain that $\forall\ j$, being the number of trials for both the Poisson-Binomial and the Binomial process, the expected value of the term of our interest for a Poisson-Binomial process with a certain average of the probabilities of the success at each trial, namely $\overline{\mu}_1(j)=\frac{\sum_{l=1}^{j}\mu_1(l)}{j}$, is always smaller than the one of a Binomial process where each Bernoulli trial has probability of

success equal to $\overline{\mu}_1(j)$. More formally:

$$\mathbb{E}_{\text{PB}(\underline{\mu}_1(j))}\left[\frac{1}{F_{j+1,y}}\right] \leqslant \mathbb{E}_{\text{Bin(j,}\overline{\mu}_1(j))}\left[\frac{1}{F_{j+1,y}}\right]. \tag{34}$$

To show that $(*) \leqslant (**)$ for any j such that $\overline{\mu}_1(j) \geqslant x$ we need to prove that the expected value of $\frac{1}{F}$ considered for a Binomial process with mean $\overline{\mu}_1(j)$ is smaller than the expected value of $\frac{1}{F}$ for a Binomial Process with mean x.

We apply Lemma C.3 stating that for a non-negative random variable (like ours $1/F^B_{j+1,y_i}$), the expected value can be computed as:

$$\mathbb{E}\left[\frac{1}{F^B_{j+1,y_i}}\right] = \int_0^{+\infty} \mathbb{P}\left(\frac{1}{F^B_{j+1,y_i}} > y\right) \mathrm{d}y. \tag{35}$$

Let $X'' \sim \text{Bin}(j, \overline{\mu}_1(j))$. Thus, we have:

$$\mathbb{P}\left(\frac{1}{F^B_{j+1,y_i}} > y\right) = \mathbb{P}(X''=0) + \mathbb{P}(X''=1) + \ldots + \mathbb{P}\left(X'' = \left(\frac{1}{F^B_{j+1,y_i}}\right)^{-1}(y) - 1\right) \tag{36}$$

$$= \mathbb{P}\left(X'' < \underbrace{\left(\frac{1}{F^B_{j+1,y_i}}\right)^{-1}(y)}_{=: k_j(y)}\right), \tag{37}$$

and the same goes for $X \sim \text{Bin}(j, x)$:

$$\mathbb{P}\left(\frac{1}{F^B_{j+1,y_i}} > y\right) = \mathbb{P}(X=0) + \mathbb{P}(X=1) + \ldots + \mathbb{P}\left(X = \left(\frac{1}{F^B_{j+1,y_i}}\right)^{-1}(y) - 1\right) \tag{38}$$

$$= \mathbb{P}\left(X < \underbrace{\left(\frac{1}{F^B_{j+1,y_i}}\right)^{-1}(y)}_{=: k_j(y)}\right), \tag{39}$$

where the inverse is formally defined as follows:

$$\left(\frac{1}{F^B_{j+1,y_i}}\right)^{-1}(y) := \min\left\{s \in \{0,\ldots,j\} : y \geqslant \frac{1}{F^B_{j+1,y_i}(s)}\right\}. \tag{40}$$

Thus, the above condition in Equation (27) becomes:

$$\int_0^{+\infty} \mathbb{P}(X'' < k_j(y))\mathrm{d}y \leqslant \int_0^{+\infty} \mathbb{P}(X < k_j(y))\mathrm{d}y. \tag{41}$$

A sufficient condition to ensure that the condition in Equation (41) is that:

$$\mathbb{P}(X'' \geqslant m) \geqslant \mathbb{P}(X \geqslant m), \forall m. \tag{42}$$

Where the latter condition follows by:

$$\int_0^{+\infty} \mathbb{P}(X'' < k_j(y))\mathrm{d}y \leqslant \int_0^{+\infty} \mathbb{P}(X < k_j(y))\mathrm{d}y \tag{43}$$

$$\int_0^{+\infty} (1 - \mathbb{P}(X'' \geqslant k_j(y)))\mathrm{d}y \leqslant \int_0^{+\infty} (1 - \mathbb{P}(X \geqslant k_j(y)))\mathrm{d}y \tag{44}$$

$$\int_0^{+\infty} \mathbb{P}(X'' \geqslant k_j(y))\mathrm{d}y \geqslant \int_0^{+\infty} \mathbb{P}(X \geqslant k_j(y))\mathrm{d}y \tag{45}$$

$$\int_0^{+\infty} \big(\mathbb{P}(X'' \geqslant k_j(y)) - \mathbb{P}(X \geqslant k_j(y))\big)\mathrm{d}y \geqslant 0, \tag{46}$$

that is guaranteed by condition (42). Let us recall the concept of stochastic order([10, 11, 30]) that is often useful in comparing random variables. For two random variables U and V, we say that U is greater than V in the usual stochastic order, and we denote it with $U \geqslant_{\text{st}} V$, when $\mathbb{P}(U \geqslant m) \geqslant \mathbb{P}(V \geqslant$

$m), \forall m$. Thus, if we have that $X'' \geqslant_{\text{st}} X$ we would have that also Equation (42) holds too. It has been shown by [10] (Lemma C.7) that the condition for that to happen when X'' and X are binomial distribution with mean μ'' and μ is that $\mu'' \geqslant \mu$. By doing that we showed that for any j such that $\overline{\mu}_1(j) \geqslant x$:

$$\mathbb{E}_{\text{PB}(\underline{\mu}_1(j))}\left[\frac{1}{F_{j+1,y}}\right] \leqslant \mathbb{E}_{\text{Bin(j,}\overline{\mu}_1(j))}\left[\frac{1}{F_{j+1,y}}\right] \leqslant \mathbb{E}_{\text{Bin}(j,x)}\left[\frac{1}{F_{j+1,y}}\right], \tag{47}$$

concludes the proof. □

Theorem 4.2 (Beta-TS - Regret Bound)**.** *Let $\sigma \in [\![\sigma(T), T]\!]$, with $\sigma(T)$ defined as in Equation 4. Under Assumption 4.1, for every $\epsilon \in (0,1)$, the Beta-TS algorithm suffers an expected cumulative regret bounded as:*

$$R(\texttt{Beta-TS}, T) \leqslant O\Bigg(\sum_{i \neq i^*(T)} \Delta_i(T,0) \Big((1+\epsilon) \frac{\log(T)}{d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))} + \frac{1}{\epsilon^2} + \sum_{j=1}^{\sigma-1} \frac{1}{(1-\overline{\mu}_1(\sigma))^{j+1}} \delta_{TV}(\mathit{PB}(\underline{\mu}_1(j)), \mathit{Bin}(j, \overline{\mu}_1(\sigma_i))) \Big) \Bigg), \tag{7}$$

where $d(x,y) := x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$ for $x, y \in [0,1]$ is the Kullback-Leibler divergence between Bernoulli distributions, $\delta_{TV}(P,Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$ denotes the total variation divergence between distributions P and Q, $\mathit{PB}(\underline{\mu}_1(j))$ denotes the Poisson-Binomial distribution with individual means $\underline{\mu}_1(j) = (\mu_1(1), \dots, \mu_1(j))$, and $\mathit{Bin}(j,x)$ denotes the binomial with j trials and parameter x.

Proof. For every suboptimal arm $i \in [\![2, K]\!]$, let us define the thresholds x_i and y_i s.t. $\overline{\mu}_i(T) < x_i < y_i < \overline{\mu}_1(\sigma)$. Thanks to the above thresholds, we can define the following events for every $t \in [\![T]\!]$:

- $E_i^{\mu}(t)$ as the event for which $\hat{\mu}_{i,t} \leqslant x_i$;
- $E_{i,t}^{\theta}$ as the event for which $\theta_{i,t} \leqslant y_i$, where $\theta_{i,t}$ denotes a sample generated for arm i from the posterior distribution at time t, i.e., Beta$(S_{i,t}+1, F_{i,t}+1)$, being $S_{i,t}$ and $F_{i,t}$ the number of successes and failures up to round t for arm i (note that $N_{i,t} = S_{i,t} + F_{i,t}$ and $\hat{\mu}_{i,t} = S_{i,t}/N_{i,t}$).

Moreover, let us denote with $E_i^{\mu}(t)^{\complement}$ and $E_i^{\theta}(t)^{\complement}$ the complementary event $E_i^{\mu}(t)$ and $E_i^{\theta}(t)$, respectively. Using Lemma A.1, we can rewrite the regret as:

$$R(\texttt{Beta-TS}, T) \leqslant \sum_{i=2}^{K} \Delta_i(T,0)\, \mathbb{E}[N_{i,T}] = \sum_{i=2}^{K} \Delta_i(T,0) \sum_{t=1}^{T} \mathbb{P}(I_t = i). \tag{48}$$

Let us focus on decomposing the probability term in the regret as follows:

$$\sum_{t=1}^{T} \mathbb{P}(I_t = i) = \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t)^{\complement})}_{=:P_A} + \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t)^{\complement})}_{=:P_B} \tag{49}$$

$$+ \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t))}_{=:P_C}. \tag{50}$$

The three terms correspond to the case of:

- (i) having a poor estimation of the mean for arm i (i.e., P_A);
- (ii) having a good estimation of the mean and having sampled a large value for the arm i posterior sample (i.e., P_B);
- (iii) having a good estimate for the mean of the reward and having sampled a small value for the posterior sample of arm i (i.e., P_C).

Let us analyze each term separately.

Term A Let $\tau_k \in [\![T]\!]$ denote the round at which we pull the arm i for the k-th time (we are omitting the dependence on the arm index i to avoid heaving the notation). In what follows, we let the sum run to times that can be greater than T. We have:

$$P_A = \sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^\mu(t)^\complement) \tag{51}$$

$$\leqslant \mathbb{E}\left[\sum_{k=1}^{T}\sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{1}\{I_t = i\}\,\mathbb{1}\left\{E_i^\mu(t)^\complement\right\}\right] \tag{52}$$

$$\leqslant \mathbb{E}\left[\sum_{k=0}^{T-1} \mathbb{1}\left\{E_i^\mu(\tau_k+1)^\complement\right\} \underbrace{\sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{1}\{I_t = i\}}_{=1}\right] \tag{53}$$

$$= \mathbb{E}\left[\sum_{k=0}^{T-1} \mathbb{1}\left\{E_i^\mu(\tau_k+1)^\complement\right\}\right] \tag{54}$$

$$\leqslant 1 + \mathbb{E}\left[\sum_{k=1}^{T-1} \mathbb{1}\left\{E_i^\mu(\tau_k+1)^\complement\right\}\right] = 1 + \sum_{k=1}^{T-1} \underbrace{\mathbb{P}(E_i^\mu(\tau_k+1)^\complement)}_{=:P_D}, \tag{55}$$

where Equation (53) follows from observing that the indicator function is 1 in a single round in the inner summation. Let us notice that thanks to the definition of the event $E_i^\mu(\tau_k+1)$, the term P_D corresponds to the probability that $\hat{\mu}_{i,\tau_k} > x_i$ after exactly k pulls (which is not a random variable). Thus, using Lemma C.1 with $\lambda = x_i - \overline{\mu}_i(k)$ and recalling that $\mathbb{E}[\hat{\mu}_{i,\tau_k}] = \overline{\mu}_i(k)$, we have:

$$P_D = \mathbb{P}(\hat{\mu}_{i,\tau_k} > x_i) = \mathbb{P}(\hat{\mu}_{i,\tau_k} > \overline{\mu}_i(k) - \overline{\mu}_i(k) + x_i) \tag{56}$$

$$\leqslant \exp\left(-k\; d(x_i, \overline{\mu}_i(k))\right) \leqslant \exp\left(-k\; d(x_i, \overline{\mu}_i(T))\right), \tag{57}$$

where $d(a,b) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$ is the Kullback-Leiber distance between two Bernoulli variables with expected value a and b, and the last inequality follows from the fact that $x_i > \mu_i(T)$. This implies that:

$$P_A \leqslant 1 + \sum_{k=1}^{T-1} \exp\left(-k\; d(x_i, \overline{\mu}_i(T))\right) \leqslant 1 + \frac{1}{d(x_i, \overline{\mu}_i(T))}, \tag{58}$$

where the last inequality follows from bounding the summation with the corresponding integral.

Term B Let us focus on the summands of the term P_B of the regret. To this end, let $(\mathbb{F}_{t-1})_{t\in[\![T]\!]}$ be the canonical filtration. We have:

$$\mathbb{P}(I_t = i, E_i^\theta(t)^\complement | E_i^\mu(t), \mathbb{F}_{t-1}) \leqslant \mathbb{P}(\theta_{i,t} > y_i | \hat{\mu}_{i,t} \leqslant x_i, \mathbb{F}_{t-1}) \tag{59}$$

$$= \mathbb{P}\left(\text{Beta}\left(\hat{\mu}_{i,t}N_{i,t}+1, (1-\hat{\mu}_{i,t})N_{i,t}+1\right) > y_i | \hat{\mu}_{i,t} \leqslant x_i\right) \tag{60}$$

$$\leqslant \mathbb{P}\left(\text{Beta}\left(x_i N_{i,t}+1, (1-x_i)N_{i,t}+1\right) > y_i\right) \tag{61}$$

$$\leqslant F^B_{N_{i,t},y_i}\left(x_i N_{i,t}\right) \leqslant \exp\left(-N_{i,t} d(x_i, y_i)\right), \tag{62}$$

where the last inequality follows from the generalized Chernoff-Hoeffding bounds (Lemma C.1) and the Beta-Binomial identity (Fact 3 of [3]). Equation (60) was derived by exploiting the fact that on the event $E_i^\mu(t)$ a sample from $\text{Beta}\left(x_i N_{i,t}+1, (1-x_i)N_{i,t}+1\right)$ is likely to be as large as a sample from $Beta(\hat{\mu}_{i,t}N_{i,t}(t)+1, (1-\hat{\mu}_{i,t})N_{i,t}+1)$, reported formally in Fact C.6. Therefore, for t such that $N_{i,t} > L_i(T)$, where $L_i(t) := \frac{\log T}{d(x_i,y_i)}$ we have:

$$\mathbb{P}(I_t = i, E_i^\theta(t)^\complement | E_i^\mu(t), \mathbb{F}_{t-1}) \leqslant \frac{1}{T}. \tag{63}$$

Let τ be the largest round until $N_{i,t} \leqslant L_i(T)$, then:

$$P_B = \sum_{t=1}^{T} \mathbb{P}\left(I_t = i, E_i^\mu(t), E_i^\theta(t)^\complement\right) \leqslant \sum_{t=1}^{T} \mathbb{P}\left(I_t = i, E_i^\theta(t)^\complement | E_i^\mu(t)\right) \tag{64}$$

$$= \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^\theta(t)^\complement | E_i^\mu(t), \mathbb{F}_{t-1})\Big] \tag{65}$$

$$= \mathbb{E}\Big[\sum_{t=1}^{\tau} \mathbb{P}(I_t = i, E_i^{\theta}(t)^{\complement} | E_i^{\mu}(t), \mathbb{F}_{t-1}) + \sum_{t=\tau+1}^{T} \mathbb{P}(I_t = i, E_i^{\theta}(t)^{\complement} | E_i^{\mu}(t), \mathbb{F}_{t-1})\Big] \tag{66}$$

$$\leqslant \mathbb{E}\Big[\sum_{t=1}^{\tau} Pr(I_t = i, E_i^{\theta}(t)^{\complement} | E_i^{\mu}(t), \mathbb{F}_{t-1})\Big] + \mathbb{E}\left[\sum_{t=\tau+1}^{T} \frac{1}{T}\right] \tag{67}$$

$$\leqslant \mathbb{E}\Big[\sum_{t=1}^{\tau} \mathbb{P}(I_t = i, E_i^{\theta}(t)^{\complement} | E_i^{\mu}(t), \mathbb{F}_{t-1})\Big] + 1 \tag{68}$$

$$= \mathbb{E}\Big[\sum_{t=1}^{\tau} \mathbb{1}(I_t = i)\Big] + 1 \tag{69}$$

$$\leqslant L_i(T) + 1. \tag{70}$$

Term C For this term, we shall use Lemma 1 by [3]. Let us define $p_{i,t} = \mathbb{P}(\theta_{1,t} > y_i | \mathbb{F}_{t-1})$. We have:

$$\mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t) | \mathbb{F}_{t-1}) \leqslant \frac{1 - p_{i,t}}{p_{i,t}} \mathbb{P}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t) | \mathbb{F}_{t-1}). \tag{71}$$

Thus, we can rewrite the term P_C as follows:

$$P_C = \sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t)) \tag{72}$$

$$= \sum_{t=1}^{T} \mathbb{E}[\mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t) | \mathbb{F}_{t-1})] \tag{73}$$

$$\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t)) \middle| \mathbb{F}_{t-1}\right]\right] \tag{74}$$

$$\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t))\right]. \tag{75}$$

$$\tag{76}$$

Let τ_k denote the time step at which arm 1 is played for the k-th time (notice we allow the sum to run trough times bigger than the learning horizon T), and let $\tau_0 = 0$:

$$P_C \leqslant \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1 - p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{1}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t))\right] \tag{77}$$

$$\leqslant \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1 - p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\right], \tag{78}$$

where the inequality in Equation (78) uses the fact that $p_{i,t}$ is fixed, given $\mathbb{F}_{t-1}$. Then, we observe that $p_{i,t} = \mathbb{P}(\theta_{1,t} > y_i | \mathbb{F}_{t-1})$ changes only when the distribution of $\theta_{1,t}$ changes, that is, only on the time step after each play of the first arm. Thus, $p_{i,t}$ is the same at all time steps $t \in \{\tau_k + 1, \ldots, \tau_{k+1}\}$, for every k. Finally, bounding the probability of selecting the optimal arm by 1 we have:

$$P_C \leqslant \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}} - 1\right]. \tag{79}$$

Let $N_{1,t} = j$, $S_{1,t} = s$. Then,

$$p_{i,t} = \mathbb{P}(\theta_{1,t} > y_i) = F^B_{j+1,y_i}(s)$$

due to the relation that links the Beta and the Binomial distributions (Fact 3 of [3]). Let $\tau_j + 1$ denote the time step after the j-th play of the optimal arm. Then, $N_{1,\tau_j+1} = j$. We do notice a sensible difference with respect to the stationary case. Indeed, the number of successes after j trial is not distributed anymore as a binomial distribution. Instead, it can be described by a Poisson-Binomial distribution $\text{PB}(\underline{\mu}_1(j))$ where the vector $\underline{\mu}_1(j) = (\mu_1(1), \ldots, \mu_1(j))$, and $\mu_1(m)$ represents the probability of success of the best arm at the m-th trial. The probability of having s successful

trials out of a total of j trials can be written as follows [47, 33]:

$$f_{j,\underline{\mu}_1(j)}(s) = \sum_{A \in F_s} \prod_{m \in A} \mu_1(m) \prod_{m' \in A^c} (1-\mu_1(m')), \tag{80}$$

where F_s is the set of all subsets of s integers that can be selected from $[\![j]\!]$. F_s by definition will contain $\frac{j!}{(j-s)!s!}$ elements, the sum over which is infeasible to compute in practice unless the number of trials j is small. A useful property of f is that it is invariant to the order of the elements in $\underline{\mu}_1(j)$. Moreover, we define density function of the binomial of j trials and mean $\overline{\mu}_1(j)$, i.e., $\text{Bin}(j, \overline{\mu}_1(j))$, as:

$$f_{j,\overline{\mu}_1(j)}(s) = \binom{j}{s} \overline{\mu}_1(j)^s (1-\overline{\mu}_1(j))^{j-s}. \tag{81}$$

By applying the change of measure argument of Lemma C.2, we have:

$$\mathbb{E}\left[\frac{1}{F^B_{j+1,y_i}}\right] = \underbrace{\sum_{s=0}^{j} \frac{f_{j,\underline{\mu}_1(j)}(s)}{F^B_{j+1,y_i}(s)}}_{(**)} \leqslant \left(\frac{1}{(1-y_i)^{j+1}} - 1\right) \delta_{\text{TV}}\left(\text{PB}(\underline{\mu}_1(j)), \text{Bin}(j, \overline{\mu}_1(\sigma))\right) + \\ + \underbrace{\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F^B_{j+1,y_i}(s)}}_{(*)}, \tag{82}$$

where $\delta_{TV}(P,Q) := \sup_{A \in \mathcal{F}} \left|P(A) - Q(A)\right|$ is the total variation between the probability measures P and Q (assuming they are defined over a measurable space $(\Omega, \mathcal{F})$, having observed that, using the notation of Lemma C.2:

$$b = \max_{s \in [\![0,j]\!]} \frac{1}{F^B_{j+1,y_i}(s)} = \frac{1}{F^B_{j+1,y_i}(0)} = \frac{1}{\mathbb{P}(\text{Bin}(j+1, y_i) = 0)} = \frac{1}{(1-y_i)^{j+1}}, \tag{83}$$

$$a = \min_{s \in [\![0,j]\!]} \frac{1}{F^B_{j+1,y_i}(s)} = \frac{1}{F^B_{j+1,y_i}(j)} = \frac{1}{\mathbb{P}(\text{Bin}(j+1, y_i) \leqslant j)} \geqslant \frac{1}{\mathbb{P}(\text{Bin}(j+1, y_i) \leqslant j+1)} = 1. \tag{84}$$

For ease of notation, let us denote $X' \sim \text{PB}(\underline{\mu}_1(j))$ and $X \sim \text{Bin}(j, \overline{\mu}(\sigma))$. We are now interested in finding if it does exist a minimum number of trials $j \in [\![0, T]\!]$ such that:

$$(**) = \mathbb{E}\left[\frac{1}{F^B_{j+1,y_i}(X')}\right] \leqslant \mathbb{E}\left[\frac{1}{F^B_{j+1,y_i}(X)}\right] = (*). \tag{85}$$

Thus, using Lemma 4.1, we conclude that:

$$\mathbb{E}\left[\frac{1}{F_{j+1,y}}\right] \leqslant \begin{cases} \left(\frac{1}{(1-y_i)^{j+1}} - 1\right) \delta_{\text{TV}}(\text{PB}(\underline{\mu}_1(j)), \text{Bin}(j, \overline{\mu}_1(\sigma))) + \sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F_{j+1,y_i^B}(s)} & \text{if } 0 \leqslant j < \sigma \\ \sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F_{j+1,y_i}(s)} & \text{if } j \geqslant \sigma \end{cases} \tag{86}$$

Where the total variation terms can be bounded by Lemma C.4 (imposing $s=0$) and Lemma C.5 in the auxiliary lemmas. From Lemma 2.9 by [3], we have that:

$$\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma_i)}(s)}{F_{j+1,y_i}(s)} - 1 \leqslant \begin{cases} \frac{3}{\Delta'_i} & \text{if } j < \frac{8}{\Delta'_i} \\ \Theta\left(e^{-\frac{\Delta'^2_i j}{2}} + \frac{e^{-D_i j}}{(j+1)\Delta'^2_i} + \frac{1}{e^{\Delta'^2_i \frac{j}{4}} - 1}\right) & \text{if } j \geqslant \frac{8}{\Delta'_i} \end{cases}, \tag{87}$$

where $\Delta'_i = \overline{\mu}_1(\sigma) - y_i$ and $D_i = y_i \log \frac{y_i}{\overline{\mu}_1(\sigma)} + (1-y_i) \log \frac{1-y_i}{1-\overline{\mu}_1(\sigma)}$. Thus, summing over all js and using the big-Oh notation to hide all functions of the μ_i's and Δ'_is, we obtain:

$$\sum_{j=0}^{T-1} \left(\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F_{j+1,y_i}(s)} - 1\right) \leqslant \frac{24}{\Delta'^2_i} + \sum_{j \geqslant \frac{8}{\Delta'_i}} \Theta\left(e^{-\frac{\Delta'^2_i j}{2}} + \frac{e^{-D_i j}}{(j+1)\Delta'^2_i} + \frac{1}{e^{\Delta'^2_i \frac{j}{4}} - 1}\right) \tag{88}$$

$$\leqslant \frac{24}{\Delta'^2_i} + \Theta\left(\frac{2}{\Delta'^2_i} + \frac{1}{\Delta'^2_i D_i} + \frac{1}{\Delta'^4_i}\right) = O(1). \tag{89}$$

which, summing all the contributions to the regret, provides the final result. □

Corollary 4.3. *Under Assumption 3.2, the* `Beta-TS` *algorithm suffers an expected cumulative regret:*

$$R(\texttt{Beta-TS},T)\leqslant\begin{cases}O(\sqrt{KT\log(T)}+K\sigma(1-\overline{\mu}_1(T))^{-\sigma}) & \text{if } T\leqslant T^*\\ O(\sqrt{KT\log(T)}) & \text{if } T>T^*\end{cases}. \tag{8}$$

Proof. If the arms dynamics is such that exists a finite time horizon T^* defined as:

$$\overline{\mu}_1(T^*)>\overline{\mu}_i(+\infty),\ \forall i\neq 1, \tag{90}$$

i.e., there exists a finite time over which the best arm will not change anymore, we can devise a finite grid of values for every T and every i of $\overline{\Delta}_i(\sigma(T),T)$ (we have taken $\sigma(T)$ for the sake of argument, notice that for every T we could choose any $\sigma\in[\![\sigma(T),T]\!]$ up to T^*, for T^* we will consider $\overline{\Delta}_i(T^*,\infty)$. Then, it is possible to define a constant c as in (16). Indeed, notice that for all $T\geqslant T^*$, taking in what we have proved earlier $\sigma=T^*$ for every time horizon $T\geqslant T^*$, the sum of the total variation distances becomes a constant with respect to the time horizon T and substituting in the result for the online regret we obtain for $T\geqslant T^*$, $\Delta_i(T^*,T)$ with $\Delta_i(T^*,\infty)$:

$$R(\texttt{Beta-TS},T)\leqslant O\Bigg(\sum_{i=2}^{K}\Delta_i(T,0)\Big((1+\epsilon)\frac{\log(T)}{d(\overline{\mu}_i(\infty),\overline{\mu}_1(T^*))}+\overbrace{\frac{1}{d(x_i,\overline{\mu}_i(\infty))}}^{(*)}+\underbrace{+\sum_{j=1}^{T^*-1}\frac{\delta_{\text{TV}}(\text{PB}(\underline{\mu}_1(j)),\text{Bin}(j,\overline{\mu}_1(T^*))}{(1-\overline{\mu}_1(T^*))^{j+1}}\Big)}_{(**)}\Bigg), \tag{91}$$

Notice that for all $T\geqslant T^*$ we have that both $(*)$ and $(**)$ are constant with T. So, neglecting these terms, we obtain:

$$R(\texttt{Beta-TS},T)\leqslant O\Bigg(\sum_{i=2}^{K}\Delta_i(T,0)\Big((1+\epsilon)\frac{\log(T)}{d(\overline{\mu}_i(\infty),\overline{\mu}_1(T^*))}\Big)\Bigg) \tag{92}$$

Using Pinsker's inequality and by definition of c, we obtain:

$$R(\texttt{Beta-TS},T)\leqslant O\Bigg(\sum_{i=2}^{K}c\overline{\Delta}_i(T^*,\infty)\Big((1+\epsilon)\frac{\log(T)}{2(\overline{\mu}_i(\infty)-\overline{\mu}_1(T^*))^2}\Big)\Bigg) \tag{93}$$

from which we can retrieve the classical instance-independent bound for Thompson Sampling. Let's now consider $T\leqslant T^*$. Rewriting all then we obtain (neglecting the constants):

$$R(\texttt{Beta-TS},T)\leqslant O\Bigg(\sum_{i=2}^{K}\Delta_i(T,0)\Big(\frac{\log(T)}{d(x_i,y_i)}+\overbrace{\frac{1}{d(x_i,\overline{\mu}_i(T))}}^{(*)}+\\ +\sum_{j=1}^{\sigma(T)-1}\frac{\delta_{\text{TV}}(\text{PB}(\underline{\mu}_1(j)),\text{Bin}(j,\overline{\mu}_1(\sigma(T)))}{(1-\overline{\mu}_1(\sigma(T)))^{j+1}}\Big)\Bigg), \tag{94}$$

$(*)$ can be bounded by a constant as by assumption we have a lower bound for the distances. So we can write thanks to the definition of c in (16), and using again Pinsker's inequality:

$$R(\texttt{Beta-TS},T)\leqslant O\Bigg(\sum_{i=2}^{K}c\overline{\Delta}_i(\sigma(T),T)\Big(\frac{\log(T)}{\overline{\Delta}_i(\sigma(T),T)^2}+\sum_{j=1}^{\sigma(T)-1}\frac{\delta_{\text{TV}}(\text{PB}(\underline{\mu}_1(j)),\text{Bin}(j,\overline{\mu}_1(\sigma(T)))}{(1-\overline{\mu}_1(\sigma(T)))^{j+1}}\Big)\Bigg), \tag{95}$$

By loosely bounding the total variation by 1 we can write neglecting the constants:

$$R(\texttt{Beta-TS},T)\leqslant O\Bigg(\sum_{i=2}^{K}\overline{\Delta}_i(\sigma(T),T)\Big(\frac{\log(T)}{\overline{\Delta}_i(\sigma(T),T)^2}+\sigma(T)\left(\frac{1}{1-\overline{\mu}_1(\sigma(T))}\right)^{\sigma(T)}\Big)\Bigg), \tag{96}$$

Similarly to what has been done in [2], by analyzing the two cases:

$$\overline{\Delta}_i(\sigma(T),T)\geqslant\sqrt{K\frac{\log(T)}{T}}, \tag{97}$$

$$\overline{\Delta}_i(\sigma(T), T) \leqslant \sqrt{K \frac{\log(T)}{T}}, \tag{98}$$

we retrieve the final result. □

A.3 Proofs of Section 5

Theorem 5.1 (γ-GTS - Regret Bound for Subgaussian SRB)**.** *Let $\sigma \in [\![\sigma(T), T]\!]$ with $\sigma(T)$ defined as in Equation 4. Under Assumption 5.1, setting $\gamma \leqslant \min\left\{\frac{1}{4\sigma_{var}^2}, 1\right\}$, the γ-GTS algorithm suffers an expected cumulative regret of:*

$$R(\gamma\text{-GTS}, T) \leqslant O\left(\sum_{i \neq i^*(T)} \Delta_i(T,0)\left(\frac{\log(T\overline{\Delta}_i(\sigma,T)^2 + e^6)}{\gamma\overline{\Delta}_i(\sigma,T)^2} + \frac{\sigma_{var}^2}{\overline{\Delta}_i(\sigma,T)^2} + \sum_{j=1}^{\sigma-1} \frac{\delta_{TV}(\mathbb{P}_j, \mathbb{Q}_j(\overline{\mu}_1(\sigma)))}{\textbf{\textit{erfc}}(\sqrt{\frac{\gamma j}{2}}(\overline{\mu}_1(\sigma)))}\right)\right),$$

where ***erfc***$(\cdot)$ *is the complementary error function, $\mathbb{P}_j$ is the distribution of the sample mean of the first j samples collected from arm 1, while $\mathbb{Q}_j(y)$ is the distribution of the sample mean of j samples collected from* any *σ_{var}^2-subgaussian distribution with mean y.*

Proof. For every suboptimal arm $i \in [\![2, K]\!]$, let us define the thresholds x_i and y_i s.t. $\overline{\mu}_i(T) < x_i < y_i < \overline{\mu}_1(\sigma)$. Thanks to the above thresholds, we can define the following events for every $t \in [\![T]\!]$:

- $E_i^{\mu}(t)$ as the event for which $\widehat{\overline{\mu}}_{i,t} \leqslant x_i$;
- $E_{i,t}^{\theta}$ as the event for which $\theta_{i,t} \leqslant y_i$, where $\theta_{i,t}$ denotes a sample generated for arm i from the posterior distribution at time t, i.e., $\mathcal{N}(\widehat{\overline{\mu}}_{i,t}, \frac{1}{\gamma N_{i_t,t}})$, being $N_{i_t,t}$ of trials at time t for arm i_t.

Moreover, let us denote with $E_i^{\mu}(t)^{\complement}$ and $E_i^{\theta}(t)^{\complement}$ the complementary event $E_i^{\mu}(t)$ and $E_i^{\theta}(t)$, respectively. Using Lemma A.1, we can rewrite the regret as:

$$R(\gamma\text{-GTS}, T) \leqslant \sum_{i=2}^{K} \Delta_i(T,0)\,\mathbb{E}[N_{i,T}] = \sum_{i=2}^{K} \Delta_i(T,0) \sum_{t=1}^{T} \mathbb{P}(I_t = i). \tag{99}$$

Let us focus on decomposing the probability term in the regret as follows:

$$\sum_{t=1}^{T} \mathbb{P}(I_t = i) = \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t)^{\complement})}_{=:P_A} + \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t)^{\complement})}_{=:P_B} \tag{100}$$

$$+ \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t))}_{=:P_C}. \tag{101}$$

The three terms correspond to the case of:

- (i) having a poor estimation of the mean for arm i (i.e., P_A);
- (ii) having a good estimation of the mean and having sampled a large value for the arm i posterior sample (i.e., P_B);
- (iii) having a good estimate for the mean of the reward and having sampled a small value for the posterior sample of arm i (i.e., P_C).

Let us analyze each term separately. We will neglect the error due to the round robin that will sum up to a constant w.r.t the time.

Term A Let $\tau_k \in [\![T]\!]$ denote the round at which we pull the arm i for the k-th time (we are omitting the dependence on the arm index i to avoid heaving the notation). In what follows, we let the sum run to times that can be greater than T. We have:

$$P_A = \sum_{t=K+1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t)^{\complement}) \tag{102}$$

$$\leqslant \mathbb{E}\left[\sum_{k=1}^{T} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{1}\{I_t = i\}\,\mathbb{1}\{E_i^{\mu}(t)^{\complement}\}\right] \tag{103}$$

$$\leqslant \mathbb{E}\left[\sum_{k=1}^{T-1} \mathbb{1}\left\{E_i^{\mu}(\tau_k+1)^{\complement}\right\} \underbrace{\sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{1}\left\{I_t=i\right\}}_{=1}\right] \tag{104}$$

$$= \mathbb{E}\left[\sum_{k=1}^{T-1} \mathbb{1}\left\{E_i^{\mu}(\tau_k+1)^{\complement})\right\}\right] \tag{105}$$

$$\leqslant 1+\mathbb{E}\left[\sum_{k=1}^{T-1} \mathbb{1}\left\{E_i^{\mu}(\tau_k+1)^{\complement}\right\}\right] = 1+\sum_{k=1}^{T-1} \underbrace{\mathbb{P}(E_i^{\mu}(\tau_k+1)^{\complement})}_{=:P_D}, \tag{106}$$

where Equation (104) follows from observing that the indicator function is 1 in a single round in the inner summation. Let us notice that thanks to the definition of the event $E_i^{\mu}(\tau_k+1)$, the term P_D corresponds to the probability that $\overline{\hat{\mu}}_{i,\tau_k} > x_i$ after exactly k pulls (which is not a random variable). Thus, using Lemma C.10 and recalling that $\mathbb{E}[\overline{\hat{\mu}}_{i,\tau_k}] = \overline{\mu}_i(k)$, we have:

$$P_D = \mathbb{P}(\overline{\hat{\mu}}_{i,\tau_k} > x_i) = \mathbb{P}(\overline{\hat{\mu}}_{i,\tau_k} > \overline{\mu}_i(k) - \overline{\mu}_i(k) + x_i) \tag{107}$$

$$\leqslant \exp\left(-k\frac{(x_i-\overline{\mu}_i(k))^2}{2\sigma_{var}^2}\right) \leqslant \exp\left(-k\frac{(x_i-\overline{\mu}_i(T))^2}{2\sigma_{var}^2}\right), \tag{108}$$

This implies that:

$$P_A \leqslant 1+\sum_{k=1}^{T-1} \exp\left(-k\frac{(x_i-\overline{\mu}_i(T))^2}{2\sigma_{var}^2}\right) \leqslant 1+\frac{2\sigma_{var}^2}{(x_i-\overline{\mu}_i(T))^2}, \tag{109}$$

where the last inequality follows by bounding the summation with the corresponding integral.

Term B Defining $L_i(T) = \frac{288\log(T\overline{\Delta}_i(\sigma,T)^2+e^6)}{\gamma\overline{\Delta}_i(\sigma,T)^2}$, we decompose each summand into two parts:

$$P_B = \sum_{t=K+1}^{T} \mathbb{P}\left(I_t=i, E_i^{\mu}(t), E_i^{\theta}(t)^{\complement}\right) \tag{110}$$

$$= \sum_{t=K+1}^{T} \mathbb{P}\left(I_t=i, k_i(t) \leqslant L_i(T), E_i^{\mu}(t), E_i^{\theta}(t)^{\complement}\right) + \mathbb{P}\left(I_t=i, k_i(t) > L_i(T), E_i^{\mu}(t), E_i^{\theta}(t)^{\complement}\right). \tag{111}$$

The first term is bounded by $L_i(T)$. For the second term:

$$\sum_{t=K+1}^{T} \mathbb{P}\left(i(t)=i, k_i(t) > L_i(T), E_i^{\theta}(t)^{\complement}, E_i^{\mu}(t)\right) \leqslant$$

$$\leqslant \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{P}\left(i(t)=i, E_i^{\theta}(t)^{\complement} \,|\, k_i(t) > L_i(T), E_i^{\mu}(t), \mathbb{F}_{t-1}\right)\right] \tag{112}$$

$$\leqslant \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{P}\left(\theta_i(t) > y_i \,|\, k_i(t) > L_i(T), \overline{\hat{\mu}}_i(t) \leqslant x_i, \mathbb{F}_{t-1}\right)\right]. \tag{113}$$

Now, $\theta_i(t)$ is a $\mathcal{N}\left(\overline{\hat{\mu}}_i(t), \frac{1}{\gamma k_i(t)}\right)$ distributed Gaussian random variable. An $\mathcal{N}\left(m,\sigma^2\right)$ distributed r.v. (i.e., a Gaussian random variable with mean m and variance σ^2) is stochastically dominated by $\mathcal{N}\left(m',\sigma^2\right)$ distributed r.v. if $m' \geqslant m$. Therefore, given $\overline{\hat{\mu}}_i(t) \leqslant x_i$, the distribution of $\theta_i(t)$ is stochastically dominated by $\mathcal{N}\left(\overline{\hat{\mu}}_i(t), \frac{1}{\gamma k_i(t)}\right)$. That is,

$$\mathbb{P}\left(\theta_i(t) > y_i \,|\, k_i(t) > L_i(T), \overline{\hat{\mu}}_i(t) \leqslant x_i, \mathbb{F}_{t-1}\right) \leqslant \mathbb{P}\left(\mathcal{N}\left(x_i, \frac{1}{\gamma k_i(t)}\right) > y_i \,\middle|\, \mathbb{F}_{t-1}, k_i(t) > L_i(T)\right).$$

Here, we a slight abuse of notation we say that $\mathbb{P}\left(\mathcal{N}\left(m,\sigma^2\right)>y_i\right)$ represents the probability that a random variable distributed as $\mathcal{N}\left(m,\sigma^2\right)$ takes value greater than y_i. We have:

$$\mathbb{P}\left(\theta_i(t)>y_i \,|\, k_i(t)>L_i(T), \overline{\hat{\mu}}_i(t)\leqslant x_i, \mathbb{F}_{t-1}\right)\leqslant \mathbb{P}\left(\mathcal{N}\left(x_i,\frac{1}{\gamma k_i(t)}\right)>y_i \,\middle|\, \mathbb{F}_{t-1}, k_i(t)>L_i(T)\right). \tag{114}$$

Using Lemma C.9 we have:

$$\mathbb{P}\left(\mathcal{N}\left(x_i,\frac{1}{\gamma k_i(t)}\right)>y_i\right)\leqslant \frac{1}{2}e^{-\frac{(\gamma k_i(t))(y_i-x_i)^2}{2}} \tag{115}$$

$$\leqslant \frac{1}{2}e^{-\frac{(\gamma L_i(T))(y_i-x_i)^2}{2}}, \tag{116}$$

which is smaller than $\frac{1}{T\overline{\Delta}_i(\sigma,T)^2}$ because $L_i(T)\geqslant \frac{2\ln\left(T\overline{\Delta}_i(\sigma,T)^2\right)}{\gamma(y_i-x_i)^2}$. Substituting, we get,

$$\mathbb{P}\left(\theta_i(t)>y_i \,|\, k_i(t)>L_i(T), \hat{\mu}_i(t)\leqslant x_i, \mathcal{F}_{t-1}\right)\leqslant \frac{1}{T\overline{\Delta}_i(\sigma,T)^2}. \tag{117}$$

Summing over $t=1,\dots,T$, we get a bound of $\frac{1}{\overline{\Delta}_i(\sigma,T)^2}$.

Term C For this term, we shall use Lemma 1 by [3]. Let us define $p_{i,t}=\mathbb{P}(\theta_{1,t}>y_i|\mathbb{F}_{t-1})$. We have:

$$\mathbb{P}(I_t=i,E_i^\mu(t),E_i^\theta(t)|\mathbb{F}_{t-1})\leqslant \frac{1-p_{i,t}}{p_{i,t}}\,\mathbb{P}(I_t=1,E_i^\mu(t),E_i^\theta(t)|\mathbb{F}_{t-1}). \tag{118}$$

Thus, we can rewrite the term P_C as follows:

$$P_C=\sum_{t=K+1}^{T}\mathbb{P}(I_t=i,E_i^\mu(t),E_i^\theta(t)) \tag{119}$$

$$=\sum_{t=K+1}^{T}\mathbb{E}[\mathbb{P}(I_t=i,E_i^\mu(t),E_i^\theta(t)|\mathbb{F}_{t-1})] \tag{120}$$

$$\leqslant \sum_{t=K+1}^{T}\mathbb{E}\left[\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\mathbb{1}(I_t=1,E_i^\mu(t),E_i^\theta(t))\middle|\mathbb{F}_{t-1}\right]\right] \tag{121}$$

$$\leqslant \sum_{t=K+1}^{T}\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\mathbb{1}(I_t=1,E_i^\mu(t),E_i^\theta(t))\right]. \tag{122}$$

$$\tag{123}$$

Let τ_k denote the time step at which arm 1 is played for the k-th time (notice we allow the sum to run trough times bigger than the learning horizon T), and let $\tau_0=0$:

$$P_C\leqslant \sum_{k=1}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\sum_{t=\tau_k+1}^{\tau_{k+1}}\mathbb{1}(I_t=1,E_i^\mu(t),E_i^\theta(t))\right] \tag{124}$$

$$\leqslant \sum_{k=1}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\right], \tag{125}$$

where the inequality in Equation (125) uses the fact that $p_{i,t}$ is fixed, given $\mathbb{F}_{t-1}$. Then, we observe that $p_{i,t}=\mathbb{P}(\theta_{1,t}>y_i|\mathbb{F}_{t-1})$ changes only when the distribution of $\theta_{1,t}$ changes, that is, only on the time step after each play of the first arm. Thus, $p_{i,t}$ is the same at all time steps $t\in\{\tau_k+1,\dots,\tau_{k+1}\}$, for every k. Finally, bounding the probability of selecting the optimal arm by 1 we have:

$$P_C\leqslant \sum_{k=1}^{T-1}\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}-1\right]. \tag{126}$$

Now in order to face this term, let's consider the arbitrary $j-th$ trial, $\forall\, j$ thanks to Lemma C.2 we can bound the difference between the real process and an analogous (same number of trials) virtual process with mean $\overline{\mu}_1(\sigma)$ (where by stationary we mean that all the trials of the virtual process will

have a fixed mean):

$$\underbrace{\mathbb{E}_{\overline{\mu}_1(j)}\left[\frac{1}{p_{i,\tau_j+1}}\right]}_{(*)} \leqslant \frac{2\delta_{TV}(\mathbb{P}_j, \mathbb{Q}_j(\overline{\mu}_1(\sigma_i)))}{\mathbf{erfc}(\sqrt{\frac{\gamma j}{2}}\overline{\mu}_1(\sigma_i))} + \underbrace{\mathbb{E}_{\overline{\mu}_1(\sigma)}\left[\frac{1}{p_{i,\tau_j+1}}\right]}_{(**)}. \tag{127}$$

where $\delta_{TV}(P,Q) := \sup_{A\in\mathcal{F}} \left|P(A) - Q(A)\right|$ is the total variation between the probability measures P and Q (assuming they are defined over a measurable space $(\Omega, \mathcal{F})$, having observed that, using the notation of Lemma C.2 (as the environment can't produce rewards smaller than zero):

$$b = \max_s \frac{1}{\mathbb{P}\left(\mathcal{N}\left(s, \frac{1}{\gamma k_i(t)}\right) > y_i\right)} \leqslant \frac{1}{\mathbb{P}\left(\mathcal{N}\left(0, \frac{1}{\gamma k_i(t)}\right) \geqslant \overline{\mu}_1(\sigma)\right)} = \frac{2}{\mathbf{erfc}(\sqrt{\frac{\gamma j}{2}}\overline{\mu}_1(\sigma))}, \tag{128}$$

$$a = 1. \tag{129}$$

For example, as both binomial and poisson-binomial process are subgaussian, when we have a poisson-binomial process the analogous one shall be a binomial with fixed mean, $\overline{\mu}_1(\sigma)$ and σ^2_{var} will be $\frac{1}{4}$, in general for bounded random variables between $[a,b]$, i.e the samples can be sampled only within $[a,b]$, we will have in what follows $\sigma^2_{var} = \frac{(b-a)^2}{4}$, then for this process the analogous will be a stationary process in which the samples can be sampled within interval $[a,b]$ centered in $\overline{\mu}_1(\sigma)$ considered for the same number of trials (like for the example can be the sum of uniform random variables). When the interval changes at every trials, without loss of generality in what follows we can take σ^2_{var} as the maximum of these variances, i.e. the maximum variance a sample can have in the setting. For random variables explicitly written in term of a mean and a variance term (like the Gaussian) holds the same. Our interest is to find if there is a minimum number of trials j such that we will have $(*) \geqslant (**)$ without any adding term. Given $\mathbb{F}_{\tau_j}$, let Θ_j denote a $\mathcal{N}\left(\overline{\hat{\mu}}_1(\tau_j+1), \frac{1}{\gamma j}\right)$ distributed Gaussian random variable. Let G_j be the geometric random variable denoting the number of consecutive independent trials until and including the trial where a sample of Θ_j becomes greater than y_i. Then observe that $p_{i,\tau_j+1} = \Pr\left(\Theta_j > y_i \,|\, \mathbb{F}_{\tau_j}\right)$ and

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}}\right] = \mathbb{E}\left[\mathbb{E}\left[G_j \,|\, \mathbb{F}_{\tau_j}\right]\right] = \mathbb{E}\left[G_j\right] \tag{130}$$

We compute first the expected value for the real process. We will consider first j such that $\overline{\mu}_1(j) \geqslant \overline{\mu}_1(\sigma)$, we will bound the expected value of G_j by a constant for all j defined as earlier. Consider any integer $r \geqslant 1$. Let $z = \sqrt{\ln r}$ and let random variable MAX_r denote the maximum of r independent samples of Θ_j. We abbreviate $\overline{\hat{\mu}}_1(\tau_j+1)$ to $\overline{\hat{\mu}}_1$ and we will abbreviate $\overline{\mu}_1(j)$ as μ_1 and $\overline{\Delta}_i(j,T)$ as Δ_i in the following. Then for any integer $r \geqslant 1$:

$$\mathbb{P}(G_j \leqslant r) \geqslant \mathbb{P}(\mathrm{MAX}_r > y_i) \tag{131}$$

$$\geqslant \mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) \tag{132}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) \middle| \mathbb{F}_{\tau_j}\right]\right] \tag{133}$$

$$= \mathbb{E}\left[\mathbb{1}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) \mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \middle| \mathbb{F}_{\tau_j}\right)\right] \tag{134}$$

For any instantiation F_{τ_j} of $\mathbb{F}_{\tau_j}$, since Θ_j is Gaussian $\mathcal{N}\left(\hat{\mu}_1, \frac{1}{\gamma j}\right)$ distributed r.v., this gives using C.8:

$$\mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \middle| \mathbb{F}_{\tau_j} = F_{\tau_j}\right) \geqslant 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{z}{(z^2+1)} e^{-z^2/2}\right)^r \tag{135}$$

$$= 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\ln r}}{(\ln r + 1)} \frac{1}{\sqrt{r}}\right)^r \tag{136}$$

$$\geqslant 1 - e^{-\frac{r}{\sqrt{4\pi r \ln r}}}. \tag{137}$$

For $r \geqslant e^{12}$:

$$\mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \middle| \mathbb{F}_{\tau_j} = F_{\tau_j}\right) \geqslant 1 - \frac{1}{r^2}. \tag{138}$$

Substituting we obtain:

$$\mathbb{P}(G_j \leqslant r) \geqslant \mathbb{E}\left[\mathbb{1}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right)\left(1 - \frac{1}{r^2}\right)\right] \tag{139}$$

$$= \left(1 - \frac{1}{r^2}\right)\mathbb{P}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right). \tag{140}$$

Applying Lemma C.10 to the second term, we can write:

$$\mathbb{P}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant \mu_1\right) \geqslant 1 - e^{-\frac{z^2}{2\gamma\sigma_{var}^2}} \geqslant 1 - \frac{1}{r^2}, \tag{141}$$

being $\gamma \leqslant \frac{1}{4\sigma_{var}^2}$. Using, $y_i \leqslant \mu_1$, this gives

$$\mathbb{P}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) \geqslant 1 - \frac{1}{r^2}. \tag{142}$$

Substituting all back we obtain:

$$\mathbb{E}[G_j] = \sum_{r=0}^{\infty} \mathbb{P}(G_j \geqslant r) \tag{143}$$

$$= 1 + \sum_{r=1}^{\infty} \mathbb{P}(G_j \geqslant r) \tag{144}$$

$$\leqslant 1 + e^{12} + \sum_{r \geqslant 1}\left(\frac{1}{r^2} + \frac{1}{r^2}\right) \tag{145}$$

$$\leqslant 1 + e^{12} + 2 + 2. \tag{146}$$

This shows a constant bound of $\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}} - 1\right] = \mathbb{E}[G_j] - 1 \leqslant e^{12} + 5$ for all $j \geqslant \sigma$. We derive a bound for large j. Consider $j > L_i(T)$ (and still $j \geqslant \sigma$). Given any $r \geqslant 1$, define G_j, MAX_r, and $z = \sqrt{\ln r}$ as defined earlier. Then,

$$\mathbb{P}(G_j \leqslant r) \geqslant \mathbb{P}(\mathrm{MAX}_r > y_i) \tag{147}$$

$$\geqslant \mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \geqslant y_i\right) \tag{148}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \geqslant y_i\right)\middle|\, \mathbb{F}_{\tau_j}\right]\right] \tag{149}$$

$$= \mathbb{E}\left[\mathbb{1}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} + \frac{\Delta_i}{6} \geqslant \mu_1\right)\mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6}\middle|\, \mathbb{F}_{\tau_j}\right)\right]. \tag{150}$$

where we used that $y_i = \mu_1 - \frac{\Delta_i}{3}$. Now, since $j \geqslant L_i(T) = \frac{288\ln\left(T\Delta_i^2 + e^6\right)}{\gamma\Delta_i^2}$,

$$2\frac{\sqrt{2\ln\left(T\Delta_i^2 + e^6\right)}}{\sqrt{\gamma j}} \leqslant \frac{\Delta_i}{6}. \tag{151}$$

Therefore, for $r \leqslant \left(T\Delta_i^2 + e^6\right)^2$,

$$\frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} = \frac{\sqrt{\ln(r)}}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \leqslant -\frac{\Delta_i}{12}. \tag{152}$$

Then, since Θ_j is $\mathcal{N}\left(\overline{\hat{\mu}}_1(\tau_j + 1), \frac{1}{\gamma j}\right)$ distributed random variable, using the upper bound in Lemma C.9, we obtain for any instantiation F_{τ_j} of history $\mathbb{F}_{\tau_j}$,

$$\mathbb{P}\left(\Theta_j > \overline{\hat{\mu}}_1(\tau_j + 1) - \frac{\Delta_i}{12}\middle|\, \mathbb{F}_{\tau_j} = F_{\tau_j}\right) \geqslant 1 - \frac{1}{2}e^{-\gamma j\frac{\Delta_i^2}{288}} \geqslant 1 - \frac{1}{2\left(T\Delta_i^2 + e^6\right)}. \tag{153}$$

being $j \geqslant L_i(T)$. This implies:

$$\mathbb{P}\left(\mathrm{MAX}_r > \hat{\mu}_1(\tau_j + 1) + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6}\middle|\, \mathbb{F}_{\tau_j} = F_{\tau_j}\right) \geqslant 1 - \frac{1}{2^r\left(T\Delta_i^2 + e^6\right)^r}. \tag{154}$$

Also, for any $t \geqslant \tau_j + 1$, we have $k_1(t) \geqslant j$, and using Lemma C.10, we get:

$$\mathbb{P}\left(\widehat{\overline{\mu}}_1(t) + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \geqslant y_i\right) \geqslant \mathbb{P}\left(\widehat{\overline{\mu}}_1(t) \geqslant \mu_1 - \frac{\Delta_i}{6}\right) \geqslant 1 - e^{-k_1(t)\Delta_i^2/72\sigma_{var}^2} \geqslant 1 - \frac{1}{\left(T\Delta_i^2 + e^6\right)^{16}}. \tag{155}$$

Let $T' = \left(T\Delta_i^2 + e^6\right)^2$. Therefore, for $1 \leqslant r \leqslant T'$, we have:

$$\mathbb{P}(G_j \leqslant r) \geqslant 1 - \frac{1}{2^r \left(T'\right)^{r/2}} - \frac{1}{\left(T'\right)^8}. \tag{156}$$

When $r \geqslant T' \geqslant e^{12}$, we obtain:

$$\mathbb{P}(G_j \leqslant r) \geqslant 1 - \frac{1}{r^2} - \frac{1}{r^2}. \tag{157}$$

Combining all the bounds we have derived:

$$\begin{aligned}
\mathbb{E}[G_j] &\leqslant \sum_{r=0}^{\infty} \mathbb{P}(G_j \geqslant r) && (158)\\
&\leqslant 1 + \sum_{r=1}^{T'} \mathbb{P}(G_j \geqslant r) + \sum_{r=T'}^{\infty} \mathbb{P}(G_j \geqslant r) && (159)\\
&\leqslant 1 + \sum_{r=1}^{T'} \frac{1}{\left(2\sqrt{T'}\right)^r} + \frac{1}{\left(T'\right)^7} + \sum_{r=T'}^{\infty} \frac{1}{r^2} + \frac{1}{r^{1.5}} && (160)\\
&\leqslant 1 + \frac{1}{\sqrt{T'}} + \frac{1}{\left(T'\right)^7} + \frac{2}{T'} + \frac{3}{\sqrt{T'}} && (161)\\
&\leqslant 1 + \frac{5}{T\Delta_i^2 + e^6}. && (162)
\end{aligned}$$

So we have proved that:

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}}\right] \leqslant \begin{cases} \frac{2\delta_{TV}(\mathbb{P}_j(\overline{\mu}_1(j)), \mathbb{P}_j(\overline{\mu}_1(\sigma)))}{\textbf{erfc}(\sqrt{\frac{\gamma j}{2}}\overline{\mu}_1(\sigma_i))} + \mathbb{E}_{\overline{\mu}_1(\sigma)}\left[\frac{1}{p_{i,\tau_j+1}}\right]. & \text{if } 0 \leqslant j < \sigma \\ (e^{12}+5) & \text{if } j \geqslant \sigma \\ \frac{5}{T\overline{\Delta}_i(j,T)^2} & \textit{if } j \geqslant L_i(T,j) \textit{ and } j \geqslant \sigma \end{cases} \tag{163}$$

Notice that $L_i(T,j) = \frac{288\log(T\overline{\Delta}_i(j,T)^2 + e^6)}{\gamma\overline{\Delta}_i^2(j,T)}$ is decreasing w.r.t. $\overline{\Delta}_i(j,T)$, so we can write:

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}}\right] \leqslant \begin{cases} \frac{2\delta_{TV}(\mathbb{P}_j(\overline{\mu}_1(j)), \mathbb{P}_j(\overline{\mu}_1(\sigma)))}{\textbf{erfc}(\sqrt{\frac{\gamma j}{2}}\overline{\mu}_1(\sigma))} + \mathbb{E}_{\overline{\mu}_1(\sigma)}\left[\frac{1}{p_{i,\tau_j+1}}\right]. & \text{if } 0 \leqslant j < \sigma \\ (e^{12}+5) & \text{if } j \geqslant \sigma \\ \frac{5}{T\overline{\Delta}_i(j,T)^2} & \textit{if } j \geqslant \frac{288\log(T\overline{\Delta}_i^2(\sigma,T)+e^6)}{\gamma\overline{\Delta}_i^2(j,T)} \textit{ and } j \geqslant \sigma \end{cases} \tag{164}$$

By definition, we have:

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}}\right] \leqslant \begin{cases} \frac{2\delta_{TV}(\mathbb{P}_j(\overline{\mu}_1(j)), \mathbb{P}_j(\overline{\mu}_1(\sigma)))}{\textbf{erfc}(\sqrt{\frac{\gamma j}{2}}\overline{\mu}_1(\sigma))} + \mathbb{E}_{\overline{\mu}_1(\sigma)}\left[\frac{1}{p_{i,\tau_j+1}}\right]. & \text{if } 0 \leqslant j < \sigma \\ (e^{12}+5) & \text{if } j \geqslant \sigma \\ \frac{5}{T\overline{\Delta}_i(\sigma,T)^2} & \textit{if } j \geqslant \frac{288\log(T\overline{\Delta}_i^2(\sigma,T)+e^6)}{\gamma\overline{\Delta}_i^2(\sigma,T)} \textit{ and } j \geqslant \sigma \end{cases} \tag{165}$$

We can end up in two scenarios:

First Case It may happen that $\sigma \geqslant \frac{288\log(T\overline{\Delta}_i^2(\sigma,T)+e^6)}{\gamma\overline{\Delta}_i^2(\sigma,T)}$, then, in this case, we already are in a situation in which we will sum $T - \sigma$ times the term $\frac{5}{T\overline{\Delta}_i(\sigma,T)^2}$.

Second Case The second case is the scenario in which we have $\sigma_i \leqslant \frac{288\log(T\overline{\Delta}_i^2(\sigma,T)+e^6)}{\gamma\overline{\Delta}_i^2(\sigma,T)}$. In this situation we will sum $\frac{288\log(T\overline{\Delta}_i^2(\sigma,T)+e^6)}{\gamma\overline{\Delta}_i^2(\sigma,T)} - \sigma$ times the constant bound $(e^{12}+5)$ and $T - \frac{288\log(T\overline{\Delta}_i^2(\sigma,T)+e^6)}{\gamma\overline{\Delta}_i^2(\sigma,T)}$ times the term $\frac{5}{T\overline{\Delta}_i(\sigma,T)^2}$.

Notice that what we have found is the same bound we would find doing the exact same passages for $\mathbb{E}_{\overline{\mu}_1(\sigma_i)}\left[\frac{1}{p_{i,\tau_j+1}}\right]$ for $j \geqslant \sigma$, furthermore the inequality for $\mathbb{E}_{\overline{\mu}_1(\sigma_i)}\left[\frac{1}{p_{i,\tau_j+1}}\right]$ holds true for any j by definition, i.e. it's easy to show that:

$$\mathbb{E}_{\overline{\mu}_1(\sigma)}\left[\frac{1}{p_{i,\tau_j+1}}\right] \leqslant \begin{cases} (e^{12}+5) & \forall j \\ \frac{1}{T\overline{\Delta}_i(\sigma)^2} & \textit{if } j \geqslant \frac{288\log(T\overline{\Delta}_i^2(\sigma,T)+e^6)}{\gamma\overline{\Delta}_i^2(\sigma,T)} \end{cases} \tag{166}$$

So that summing all the terms:

$$\begin{aligned} P_C \leqslant \sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}-1\right] \leqslant (e^{12}+5)\frac{288\log(T\overline{\Delta}_i^2(\sigma,T)+e^6)}{\gamma\overline{\Delta}_i^2(\sigma,T)} + \\ + \frac{5}{\overline{\Delta}_i^2(\sigma,T)} + \sum_{j=1}^{\sigma-1}\frac{2\delta_{\text{TV}}(\mathbb{P}_j(\overline{\mu}_1(j)),\mathbb{P}_j(\overline{\mu}_1(\sigma))}{\textbf{erfc}(\sqrt{\frac{\gamma j}{2}}(\overline{\mu}_1(\sigma)))}. \end{aligned} \tag{167}$$

Summing all the other term follows the statement. Notice furthermore that as a corollary we've proven in this way the optimality of γ-GTS for the generic subgaussian stationary environment. □

Corollary 5.2. *Under Assumption 3.2, the γ-GTS algorithm suffers an expected cumulative regret:*

$$R(\gamma\text{-GTS},T) \leqslant \begin{cases} O(\sqrt{KT\gamma^{-1}\log(T)} + K\sigma e^{\gamma\sigma\overline{\mu}_1(\sigma)^2}) & \textit{if } T \leqslant T^* \\ O(\sqrt{KT\gamma^{-1}\log(T)}) & \textit{if } T > T^* \end{cases}. \tag{9}$$

Proof. If the arms' dynamics is such that exists a finite time horizon T^* defined as:

$$\overline{\mu}_1(T^*) > \overline{\mu}_i(+\infty),\ \forall i \neq 1, \tag{168}$$

i.e., informally, there's a finite time over which the best arm won't change anymore, we can devise a finite grid of values for every T and every i of $\overline{\Delta}_i(\sigma(T),T)$ (we have taken $\sigma(T)$ for the sake if the argument, notice however that for every T we could choose any $\sigma \in [\![\sigma(T),T]\!]$) up to T^*, for T^* we will consider $\overline{\Delta}_i(T^*,\infty)$. Then it is possible to define a constant c as in 16. In fact notice that for all $T \geqslant T^*$, taking in what we've proved earlier $\sigma = T^*$ for every time horizon $T \geqslant T^*$, the sum of the total variation distances becomes a constant with respect to the time and substituting in all the terms for $T \geqslant T^*$, $\Delta_i(T^*,T)$ with $\Delta_i(T^*,\infty)$, we obtain (neglecting the constant terms with respect to the time), since all the terms are increasing for decreasing $\overline{\Delta}_i(\sigma(T),T)$, we find:

$$R(\gamma\text{-GTS},T) \leqslant O\left(\sum_{i=2}^{K}\Delta_i(T,0)\Big(C_1\frac{\log(T\overline{\Delta}_i(T^*,T)^2+e^6)}{\gamma\overline{\Delta}_i(T^*,T)^2} + \frac{18\sigma_{var}^2+6}{\overline{\Delta}_i(T^*,T)^2}\Big)\right), \tag{169}$$

Then, by definition:

$$R(\gamma\text{-GTS},T) \leqslant O\left(\sum_{i=2}^{K}\Delta_i(T,0)\Big(C_1\frac{\log(T\overline{\Delta}_i(T^*,\infty)^2+e^6)}{\gamma\overline{\Delta}_i(T^*,\infty)^2} + \frac{18\sigma_{var}^2+6}{\overline{\Delta}_i(T^*,\infty)^2}\Big)\right). \tag{170}$$

Notice that also the second term in the above inequality is time-independent for $T \geqslant T^*$. Using the definition of c (Equation (16)), we can rewrite the regret as follows:

$$R(\gamma\text{-GTS},T) \leqslant O\left(\sum_{i=2}^{K}c\Delta_i(\sigma(T^*),\infty)\Big(C_1\frac{\log(T\overline{\Delta}_i(T^*,\infty)^2+e^6)}{\gamma\overline{\Delta}_i(T^*,\infty)^2}\Big)\right). \tag{171}$$

Disregarding the constant terms w.r.t. time T:

$$R(\gamma\text{-GTS},T) \leqslant O\left(\sum_{i=2}^{K}\Delta_i(T^*,\infty)\Big(\frac{\log(T\overline{\Delta}_i(T^*,\infty)^2+e^6)}{\overline{\Delta}_i(T^*,\infty)^2}\Big)\right), \tag{172}$$

that is equivalent to the bound provided for the classical instance-independent regret bound by [2] for the stationary sunbgaussian bandit.

Now consider $T \leqslant T^*$, we can write, using the definition of c in Equation (26):

$$R(\gamma\text{-GTS}, T) \leqslant O\Bigg(\sum_{i=2}^{K} c\overline{\Delta}_i(\sigma(T), T)\Big(C_1 \frac{\log(T\overline{\Delta}_i(\sigma(T), \infty)^2 + e^6)}{\gamma \overline{\Delta}_i(\sigma(T), T)^2} + \\ + \frac{18\sigma_{var}^2 + 6}{\overline{\Delta}_i(\sigma(T), T)^2} + \sum_{j=1}^{\sigma(T)-1} \frac{2\delta_{\text{TV}}(\mathbb{P}_j, \mathbb{Q}_j(\overline{\mu}_1(\sigma(T))))}{\textbf{erfc}(\sqrt{\frac{\gamma j}{2}}(\overline{\mu}_1(\sigma(T))))}\Big)\Bigg), \tag{173}$$

We notice that thanks to the definition of c in 16, $\frac{18\sigma_{var}^2+6}{\overline{\Delta}_i(\sigma(T),T)^2}$ is bounded with a constant term with respect to the time horizon T. Then, we have:

$$R(\gamma\text{-GTS}, T) \leqslant O\Bigg(\sum_{i=2}^{K} c\overline{\Delta}_i(\sigma(T), T)\Big(C_1 \frac{\log(T\overline{\Delta}_i(\sigma(T), \infty)^2 + e^6)}{\gamma \overline{\Delta}_i(\sigma(T), T)^2} + \sum_{j=1}^{\sigma(T)-1} \frac{2\delta_{\text{TV}}(\mathbb{P}_j, \mathbb{Q}_j(\overline{\mu}_1(\sigma(T))))}{\textbf{erfc}(\sqrt{\frac{\gamma j}{2}}(\overline{\mu}_1(\sigma(T))))}\Big)\Bigg), \tag{174}$$

[14] proved that the complementary error function for $x \geqslant 0$ can be bounded lower-bounded as:

$$\textbf{ercf}(x) \geqslant \sqrt{\frac{e}{\pi}} e^{-2x^2} \tag{175}$$

so by loosely bounding the total variation distances with 1 we get:

$$R(\gamma\text{-GTS}, T) \leqslant O\Bigg(\sum_{i=2}^{K} c\overline{\Delta}_i(\sigma(T), T)\Big(C_1 \frac{\log(T\overline{\Delta}_i(\sigma(T), T)^2 + e^6)}{\gamma \overline{\Delta}_i(\sigma(T), T)^2} + C_2\sigma(T) e^{\gamma\sigma(T)\overline{\mu}_1(\sigma(T))^2}\Big)\Bigg), \tag{176}$$

so that:

$$R(\gamma\text{-GTS}, T) \leqslant O\Bigg(\sum_{i=2}^{K} c\overline{\Delta}_i(\sigma(T), T)\Big(\frac{\log(T\overline{\Delta}_i(\sigma(T), T)^2 + e^6)}{\gamma \overline{\Delta}_i(\sigma(T), T)^2} + \sigma(T) e^{\gamma\sigma(T)\overline{\mu}_1(\sigma(T))^2}\Big)\Bigg), \tag{177}$$

considering then the two cases:

$$\overline{\Delta}_i(\sigma(T), T) \leqslant e\sqrt{K\frac{1}{\gamma T}}, \tag{178}$$

$$\overline{\Delta}_i(\sigma(T), T) \geqslant e\sqrt{K\frac{1}{\gamma T}}. \tag{179}$$

Substituting the above cases in Equation (177), concludes the proof, noticing that as Assumption 5.1 holds true for any number off pulls it does exist a time independent constant M such that $\overline{\Delta}_i(\sigma(T), T) < M$. □

Theorem 5.3 (γ-GTS - Regret Bound for Subgaussian SRB γ-tuned)**.** *Let $\sigma \in [\![\sigma(T), T]\!]$ with $\sigma(T)$ defined as in Equation 4, let furthermore $\sigma \sim T^\beta$ and $\gamma \sim T^{-\alpha}$. Under Assumption 5.1, for every $\alpha \geqslant \beta$:*

$$R(\gamma\text{-}\textit{GTS}, T) \leqslant O\Bigg(\sum_{i \neq i^*(T)} \Delta_i(T, 0)\Big(\frac{T^\alpha \log(T\overline{\Delta}_i(\sigma, T)^2 + e^6)}{\overline{\Delta}_i(\sigma, T)^2} + \frac{\sigma_{var}^2}{\overline{\Delta}_i(\sigma, T)^2} + \sigma\Big)\Bigg). \tag{10}$$

Proof. The proof follows from the proof of Theorem 5.1 setting $\gamma = T^{-\alpha}$, with α within the bounds given in the statement, noticing that as by Assumption 5.1 the rewards can be bounded by a time independent constant M and so also $\frac{1}{\textbf{erfc}(M)}$. □

Corollary 5.4. *Under assumption 3.2 γ-GTS with γ tuned suffer an instance independent regret bound upper bounded by (for all T):*

$$R(\gamma\text{-}\textit{GTS}, T) \leqslant O\left(T^{\frac{1+\alpha}{2}} \sqrt{K \log(T)} + KT^\alpha \right). \tag{11}$$

Proof. The corollary follows by substituing $\gamma = T^{-\alpha}$ in Corollary 5.2 and considering the worst case scenario. □

Algorithm 3 `Beta-SWTS` Algorithm

Input: Number of arms K, Time horizon T, time window τ
Set $X_{i,t,\tau} \leftarrow 0$ for each $i \in [\![K]\!]$
Set $\alpha_{i,1} \leftarrow 1 + X_{i,t,\tau}$ and $\beta_{i,1} \leftarrow 1 + (1 - X_{i,t,\tau})$ for each $i \in [\![K]\!]$
Set $\nu_{i,1} \leftarrow Beta(\alpha_{i,1}, \beta_{i,1})$ for each $i \in [\![K]\!]$
for $t \in [\![T]\!]$ **do**
  Sample $\theta_{i,t,\tau} \sim \nu_{i,t}$ for each $i \in [\![K]\!]$
  Select $I_t \in \arg\max_{i \in [\![K]\!]} \theta_{i,t,\tau}$
  Pull arm I_t
  Collect reward X_t
  Update $X_{i,t,\tau}$ and $T_{i,t,\tau}$, respectively the sum of collected rewards within t and $t-\tau+1$ for arm i and the number arm i has been pulled within t and $t-\tau+1$
  Update for each $i \in [\![K]\!]$ $\nu_{i,t+1} \leftarrow$ Beta$(1 + X_{i,t,\tau}, 1 + (T_{i,t,\tau} - X_{i,t,\tau}))$
end for

Algorithm 4 γ-`SWGTS` Algorithm

Input: Number of arms K, Time horizon T, exploration parameter γ, time window τ
Play every arm once and collect reward X_t
Set $T_{i,t,\tau} \leftarrow 1$, $\hat{\mu}_{i,t,\tau} \leftarrow X_t$, $\overline{\hat{\mu}}_{i,t,\tau} \leftarrow \hat{\mu}_{i,t,\tau}$ for each $i \in [\![K]\!]$
Set $\nu_{i,t} \leftarrow \mathcal{N}(\overline{\hat{\mu}}_{i,t,\tau}, \frac{1}{\gamma})$ for each $i \in [\![K]\!]$
for $t \in [\![T]\!]$ **do**
  Sample $\theta_{i,t,\tau} \sim \nu_{i,t}$ for each $i \in [\![K]\!]$
  Select $I_t \in \arg\max_{i \in [\![K]\!]} \theta_{i,t,\tau}$
  Pull arm I_t
  Collect reward X_t
  Update the sum of the collected rewards within t and $t-\tau+1$, namely $\hat{\mu}_{i,t,\tau}$, $T_{i,t,\tau}$ the number of pulls within t and $t-\tau+1$, and $\overline{\hat{\mu}}_{i,t,\tau} = \frac{\hat{\mu}_{i,t,\tau}}{T_{i,t,\tau}}$
  Update $\nu_{i,t+1} \leftarrow \mathcal{N}(\overline{\hat{\mu}}_{i,t,\tau}, \frac{1}{\gamma T_{i,t,\tau}})$ for each $i \in [\![K]\!]$
  Every τ times play every arm once to ensure $T_{i,t,\tau} > 0$
end for

B Proofs of Section 7

In this section, we report the proof of the sliding window approach version of the algorithms we proposed. We also present the pseudocode for the `Beta-SWTS` and γ-`SWGTS` algorithms in Algorithm 3 and 4, respectively.

Theorem 7.1 (`Beta-SWTS` Regret Bound)**.** *Under Assumption 4.1, the `Beta-SWTS` algorithm suffers an expected cumulative regret bounded as:*

$$R(\texttt{Beta-SWTS}, T) \leqslant O\left(\sum_{i \neq i^*(T)} \Delta_i(T,0) \left(\frac{T \log(T)}{\tau (\Delta_i'(T;\tau))^3} + \frac{\sigma'(T;\tau)}{(1 - \overline{\mu}_1(\sigma'(T;\tau),\tau))^{\tau+1}} \right) \right). \tag{14}$$

Proof. For ease of notation we set $\sigma'(T;\tau) = \sigma'(\tau), \overline{\mu}_1(\sigma'(T;\tau);\tau) = \overline{\mu}_1(\sigma'(\tau))$ and $\Delta_i(T,\tau)' = \Delta_i$. For every suboptimal arm $i \in \{2, K\}$, let us define the thresholds x_i and y_i s.t. $\mu_i(T) < x_i < y_i < \overline{\mu}_1(\sigma'(\tau))$. Thanks to the above thresholds, we can define the following events for every $t \in T$:

- $E_i^{\mu}(t)$ as the event for which $\hat{\mu}_{i,t,\tau} \leqslant x_i$;
- $E_{i,t}^{\theta}$ as the event for which $\theta_{i,t,\tau} \leqslant y_i$, where $\theta_{i,t,\tau}$ denotes a sample generated for arm i from the posterior distribution at time t from the sample collected in the last τ pulls, i.e., Beta$(S_{i,t,\tau} + 1, F_{i,t,\tau} + 1)$, being $S_{i,t,\tau}$ and $F_{i,t,\tau}$ the number of successes and failures from $t - \tau$ up to round t for arm i (note that $T_{i,t,\tau} = S_{i,t,\tau} + F_{i,t,\tau}$ and $\hat{\mu}_{i,t,\tau} = S_{i,t,\tau}/T_{i,t,\tau}$).

In the current framework we will define $p_{i,t}$ as follows:

$$p_{i,t} = \Pr(\theta_{1,t,\tau} \geqslant y_i \,|\, \mathbb{F}_{t-1}).$$

Moreover, let us denote with $E_i^{\mu}(t)^{\complement}$ and $E_i^{\theta}(t)^{\complement}$ the complementary event $E_i^{\mu}(t)$ and $E_i^{\theta}(t)$, respectively. Let us decompose the probability term in the regret as follows:

$$\sum_{t=1}^{T} \mathbb{P}(I_t = i) = \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t)^{\complement})}_{=:P_A} + \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t)^{\complement})}_{=:P_B} \tag{180}$$

$$+ \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t))}_{=:P_C}. \tag{181}$$

The three terms correspond to the case of:

- (i) having a poor estimation of the mean for arm i (i.e., P_A);
- (ii) having a good estimation of the mean and having sampled a large value for the arm i posterior sample (i.e., P_B);
- (iii) having a good estimate for the mean of the reward and having sampled a small value for the posterior sample of arm i (i.e., P_C).

Let us analyze each term separately.

Term A We have:

$$P_A = \sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t)^{\complement}) \tag{182}$$

$$\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{I_t = i, E_i^{\mu}(t)^{\complement}\right\}\right] \tag{183}$$

$$\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{I_t = i, E_i^{\mu}(t)^{\complement}, T_{i,t,\tau} \leqslant \frac{\ln(T)}{(x_i - \mu_i(T)^2}\right\}\right] +$$

$$+\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{I_t = i, E_i^{\mu}(t)^{\complement}, T_{i,t,\tau} \geqslant \frac{\ln(T)}{(x_i - \mu_i(T)))^2}\right\}\right] \tag{184}$$

$$\leqslant \frac{T\ln(T)}{\tau(x_i - \mu_i(T))^2} + \sum_{t=1}^{T} \Pr\left(E_i^{\mu}(t)^{\complement} \,|\, T_{i,t,\tau} \geqslant \frac{\ln(T)}{(x_i - \mu_i(T))^2}\right) \tag{185}$$

$$\leqslant \frac{T\ln(T)}{\tau(x_i - \mu_i(T))^2} + \sum_{t=1}^{T} \frac{1}{T}, \tag{186}$$

where we used the Chernoff-Hoeffding bound for the second term in Equation (185) and Lemma C.14 for the first term.

Term B Let us focus on the summands of the term P_B of the regret. To this end, let $(\mathbb{F}_{t-1})_{t\in[\![T]\!]}$ be the canonical filtration. We have:

$$\mathbb{P}(I_t = i, E_i^{\theta}(t)^{\complement} | E_i^{\mu}(t), \mathbb{F}_{t-1}) \leqslant \mathbb{P}(\theta_{i,t,\tau} > y_i | \hat{\mu}_{i,t,\tau} \leqslant x_i, \mathbb{F}_{t-1}) \tag{187}$$

$$= \mathbb{P}\left(\text{Beta}\left(\hat{\mu}_{i,t,\tau} T_{i,t,\tau} + 1, (1 - \hat{\mu}_{i,t,\tau}) T_{i,t,\tau} + 1\right) > y_i | \hat{\mu}_{i,t,\tau} \leqslant x_i\right) \tag{188}$$

$$\leqslant \mathbb{P}\left(\text{Beta}\left(x_i T_{i,t,\tau} + 1, (1 - x_i) T_{i,t,\tau} + 1\right) > y_i\right) \tag{189}$$

$$\leqslant F^{B}_{T_{i,t,\tau}, y_i}\left(x_i T_{i,t,\tau}\right) \leqslant \exp\left(-T_{i,t,\tau} d(x_i, y_i)\right), \tag{190}$$

where the last inequality follows from the generalized Chernoff-Hoeffding bounds (Lemma C.1) and the Beta-Binomial identity (Fact 3 of [3]). Equation (188) was derived by exploiting the fact that on the event $E_i^{\mu}(t)$ a sample from $\text{Beta}\left(x_i T_{i,t,\tau} + 1, (1 - x_i) T_{i,t,\tau} + 1\right)$ is likely to be as large as a sample from $Beta(\hat{\mu}_{i,t} T_{i,t,\tau}(t) + 1, (1 - \hat{\mu}_{i,t,\tau}) T_{i,t,\tau} + 1)$, reported formally in Lemma C.6. Therefore, for t such that $T_{i,t,\tau} > L_i(T)$, where $L_i(T) := \frac{\log T}{d(x_i, y_i)}$ we have:

$$\mathbb{P}(I_t = i, E_i^{\theta}(t)^{\complement} | E_i^{\mu}(t), \mathbb{F}_{t-1}) \leqslant \frac{1}{T}. \tag{191}$$

We decompose P_B in two events, when $T_{i,t,\tau} \leqslant L_i(T)$ and when $T_{i,t,\tau} \geqslant L_i(T)$, then:

$$P_B = \sum_{t=1}^{T} \mathbb{P}\left(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t)^{\complement}\right) \leqslant \sum_{t=1}^{T} \mathbb{P}\left(I_t = i, E_i^{\theta}(t)^{\complement} | E_i^{\mu}(t)\right) \tag{192}$$

$$= \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\theta}(t)^{\complement} | E_i^{\mu}(t), \mathbb{F}_{t-1})\Big] \tag{193}$$

$$= \mathbb{E}\left[\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}(I_t = i, E_i^\theta(t)^\complement, T_{i,t,\tau} \leqslant L_i(T) | E_i^\mu(t), \mathbb{F}_{t-1}) + \sum_{t=1}^{T} \mathbb{1}(I_t = i, E_i^\theta(t)^\complement, T_{i,t,\tau} \geqslant L_i(T) | E_i^\mu(t), \mathbb{F}_{t-1})\Big]\right] \tag{194}$$

$$\leqslant L_i(T)\frac{T}{\tau} + \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{T}\right] \tag{195}$$

$$\leqslant L_i(T)\frac{T}{\tau} + 1, \tag{196}$$

where for the first term in Equation (195) we used Lemma C.14.

Term C For this term, we use Lemma 1 by [3]. Let us define $p_{i,t} = \mathbb{P}(\theta_{1,t,\tau} > y_i | \mathbb{F}_{t-1})$. We have:

$$\mathbb{P}(I_t = i, E_i^\mu(t), E_i^\theta(t) | \mathbb{F}_{t-1}) \leqslant \frac{1-p_{i,t}}{p_{i,t}} \mathbb{P}(I_t = 1, E_i^\mu(t), E_i^\theta(t) | \mathbb{F}_{t-1}). \tag{197}$$

Thus, we can rewrite the term P_C as follows:

$$P_C = \sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^\mu(t), E_i^\theta(t)) \tag{198}$$

$$= \sum_{t=1}^{T} \mathbb{E}[\mathbb{P}(I_t = i, E_i^\mu(t), E_i^\theta(t) | \mathbb{F}_{t-1})] \tag{199}$$

$$\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^\mu(t), E_i^\theta(t)) \middle| \mathbb{F}_{t-1}\right]\right] \tag{200}$$

$$\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^\mu(t), E_i^\theta(t))\right]. \tag{201}$$

$$\tag{202}$$

We rewrite the last inequality as the sum of two contributions: when the total pulls of the best arm at time t $T_{1,t} > \sigma'(\tau)$ and when $T_{1,t} \leqslant \sigma'(\tau)$. This way, we obtain the following:

$$P_C \leqslant \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^\mu(t), E_i^\theta(t), T_{1,t} \leqslant \sigma'(\tau))\right]}_{A} +$$

$$+ \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^\mu(t), E_i^\theta(t), T_{1,t} \geqslant \sigma'(\tau))\right]}_{B}. \tag{203}$$

We further decompose the term A in other terms:

$$A \leqslant \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}\left(\overbrace{I_t = 1, E_i^\mu(t), E_i^\theta(t), T_{1,t} \leqslant \sigma'(\tau), T_{1,t,\tau} \leqslant \frac{8\ln(T)}{(\overline{\mu}_1(\sigma'(\tau)) - y_i)^2}}^{\mathcal{C}1}\right)\right]}_{(A1)}$$

$$+ \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}\left(\overbrace{I_t = 1, E_i^\mu(t), E_i^\theta(t), T_{1,t} \leqslant \sigma'(\tau), T_{1,t,\tau} \geqslant \frac{8\ln(T)}{(\overline{\mu}_1(\sigma'(\tau)) - y_i)^2}}^{\mathcal{C}2}\right)\right]}_{(A2)}. \tag{204}$$

As $\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y \mid X]]$, we can bound the term $A1$ as follows:

$$A1 = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{C}1)\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mid \mathbb{1}(\mathcal{C}1)\right]\right] \tag{205}$$

$$\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{C}1)\left(\frac{\delta_{TV}(P_{t|\mathcal{C}1}, Q_{t|\mathcal{C}1})}{(1-\overline{\mu}_1(\sigma'(\tau)))^{\tau+1}} + \underbrace{\mathbb{E}_{\overline{\mu}_1(\sigma'(\tau))}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(\mathcal{C}1)\right]}_{(*)}\right)\right]. \tag{206}$$

Now consider an arbitrary instantiation $T'_{1,t,\tau}$ of $T_{1,t,\tau}$ (i.e., an arbitrary number of pulls of the optimal arm within the time window τ) in which $\mathcal{C}1$ holds true, we can rewrite $(*)$ as:

$$(*) = \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(C1)\right] = \mathbb{E}_{T'_{1,t,\tau}}\left[\underbrace{\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(C1), T_{1,t,\tau} = T'_{1,t,\tau}\right]}_{(*')}\right]. \tag{207}$$

We can bound $(*')$ using Lemma 4 by Agrawal et al. [2]:

$$\begin{aligned}(*') &= \sum_{s=0}^{T'_{1,t,\tau}} \frac{f_{T'_{1,t,\tau},\overline{\mu}_1(\sigma'(\tau))}(s)}{F_{T'_{1,t,\tau}+1,y_i}(s)} - 1 \\ &\leqslant \begin{cases} \frac{3}{\Delta'_i} & \text{if } T'_{1,t,\tau} < \frac{8}{\Delta'_i} \\ \mathcal{O}\left(e^{-\frac{\Delta'^2_i T'_{1,t,\tau}}{2}} + \frac{e^{-DT'_{1,t,\tau}}}{T'_{1,t,\tau}\Delta'^2_i} + \frac{1}{e^{\Delta'^2_i \frac{T'_{1,t,\tau}}{4}} - 1}\right) & \text{if } \frac{8}{\Delta'} \leqslant T'_{1,t,\tau} \leqslant \frac{8\ln(T)}{\Delta'^2_i} \end{cases}. \end{aligned} \tag{208}$$

We notice that the worst case scenario we can have is for $T'_{i,t,\tau} \leqslant \frac{8}{\Delta'_i}$ so that every possible instantiation in which condition $\mathcal{C}1$ holds true the expectation value of $\frac{1-p_{i,t}}{p_{i,t}}$ can be upper bounded by substituting in the latter inequalities the worst case scenario for $T'_{i,t,\tau}$ we obtain a term which is independent from the pulls:

$$(*) \leqslant \mathcal{O}\left(\frac{1}{(\overline{\mu}_1(\sigma'(\tau)) - y_i)}\right), \tag{209}$$

so that the inequality for $A1$ can be rewritten as:

$$\mathcal{O}\left(\sum_{t=1}^{T} \mathbb{1}(C1)(*)\right) \leqslant \mathcal{O}\left(\frac{T\ln(T)}{(\overline{\mu}_1(\sigma'(\tau)) - y_i)^3}\right), \tag{210}$$

where we have exploited the fact that for Lemma C.14 we have:

$$\sum_{t=1}^{T} \mathbb{1}(\mathcal{C}1) \leqslant \frac{8T\ln(T)}{\tau(\overline{\mu}_1(\sigma'(\tau)) - y_i)^2}. \tag{211}$$

Finally, we obtain:

$$A1 \leqslant \mathcal{O}\left(\frac{\sigma'(\tau)}{(1-\overline{\mu}_1(\sigma'(\tau)))^{\tau+1}} + \frac{T\ln(T)}{\tau(\overline{\mu}_1(\sigma'(\tau)) - y_i)^3}\right), \tag{212}$$

Where the last inequality is a consequence of the fact that both inequalities hold:

$$\sum_{t=1}^{T} \mathbb{1}(\mathcal{C}1) \leqslant \begin{cases} \sigma'(\tau) \\ \frac{8T\ln(T)}{\tau(\overline{\mu}_1(\sigma'(\tau)) - y_i)^2} \end{cases}. \tag{213}$$

Let us upper bound $A2$:

$$A2 = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{C}2)\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(\mathcal{C}2)\right]\right] \tag{214}$$

$$\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{C}2)\left(\frac{\delta_{TV}(P_{t|\mathcal{C}2}, Q_{t|\mathcal{C}2})}{(1-\overline{\mu}_1(\sigma'(\tau)))^{\tau+1}} + \underbrace{\mathbb{E}_{\overline{\mu}_1(\sigma'(\tau))}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(\mathcal{C}2)\right]}_{(**)}\right)\right]. \tag{215}$$

Let us consider an arbitrary instantiation $T'_{1,t,\tau}$ of $T_{1,t,\tau}$ in which $\mathcal{C}2$ holds true, i.e., an arbitrary number of pulls of the optimal arm within the time window τ. We have:

$$(**) = \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(C2)\right] = \mathbb{E}_{T'_{1,t,\tau}}\left[\underbrace{\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(C2), T_{1,t,\tau} = T'_{1,t,\tau}\right]}_{(**')}\right], \tag{216}$$

where we bound the term $(**')$ using Lemma 4 by Agrawal et al. [2]:

$$\begin{aligned}(**') &= \sum_{s=0}^{T'_{1,t,\tau}} \frac{f_{T'_{1,t,\tau},\overline{\mu}_1(\sigma'(\tau))}(s)}{F_{T'_{1,t,\tau}+1,y_i}(s)} - 1 \\ &\leqslant \mathcal{O}\left(e^{-\frac{\Delta_i'^2 T'_{1,t,\tau}}{2}} + \frac{e^{-DT'_{1,t,\tau}}}{T'_{1,t,\tau}\Delta_i'^2} + \frac{1}{e^{\Delta_i'^2\frac{T'_{1,t,\tau}}{4}} - 1}\right) \textit{ for } T'_{1,t,\tau} \geqslant \frac{8\ln(T)}{\Delta_i'^2}.\end{aligned} \tag{217}$$

We see that the worst case scenario when $\mathcal{C}2$ holds true is when $T'_{i,t,\tau} = \frac{8\ln(T)}{\Delta_i'^2}$, so considering the worst case scenario for the case $\mathcal{C}2$ holds true we can bound the expected value for $\frac{1-p_{i,t}}{p_{i,t}}$ for every possible realization of $\mathcal{C}2$ independently from $T'_{1,t,\tau}$ as:

$$(**) \leqslant \mathcal{O}\left(\frac{1}{T-1}\right) \leqslant \mathcal{O}\left(\frac{1}{T}\right), \tag{218}$$

so that:

$$A2 \leqslant \mathcal{O}\left(\frac{\sigma'(\tau)}{(1-\overline{\mu}_1(\sigma'(\tau)))^{\tau+1}}\right), \tag{219}$$

where the latter inequality is a consequence of the fact that:

$$\sum_{t=1}^{T} \mathbb{1}(\mathcal{C}2) \leqslant \sigma'(\tau). \tag{220}$$

Let us bound term B. We decompose this term in two contributions:

$$B = \sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t), T_{1,t} \geqslant \sigma'(\tau))\right], \tag{221}$$

so that, similarly to what we have done earlier, we have:

$$\begin{aligned}B = &\underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}\left(\overbrace{I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t), T_{1,t} \geqslant \sigma'(\tau), T_{1,t,\tau} \leqslant \frac{8\ln(T)}{(\overline{\mu}_1(\sigma'(\tau)) - y_i)^2}}^{\mathcal{C}1'}\right)\right]}_{B1} + \\ &\underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}\left(\overbrace{I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t), T_{1,t} \geqslant \sigma'(\tau), T_{1,t,\tau} \geqslant \frac{8\ln(T)}{(\overline{\mu}_1(\sigma'(\tau)) - y_i)^2}}^{\mathcal{C}2'}\right)\right]}_{B_2}.\end{aligned} \tag{222}$$

Let us deal with $B1$ first. We have:

$$B1 = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{C}1') \underbrace{\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(\mathcal{C}1')\right]}_{(*)}\right]. \tag{223}$$

Let us analyse $(*)$ first.

$$(*) \leqslant \mathbb{E}_{T'_{1,t,\tau}} \left[\underbrace{\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(\mathcal{C}1'), T_{1,t,\tau} = T'_{1,t,\tau} \right]}_{(**)} \right] \tag{224}$$

Lemma 4.1 applied to $(**)$, states that a bound for a $\text{Bin}(T'_{1,t,\tau}, \mu_1(\sigma'(\tau)))$, i.e., binomial process with parameters $T'_{1,t,\tau}$ and $\mu_1(\sigma'(\tau))$ holds also for $(**)$, since such a Poisson-binomial has a mean equal or larger than $\overline{\mu}_1(\sigma'(\tau)))$. It follows, applying Lemma 4 by [2] to $(**)$, we have that:

$$(**) \leqslant \mathcal{O}\left(\frac{1}{(\overline{\mu}_1(\sigma'(\tau)) - y_i)} \right).$$

Therefore we have by Lemma C.14:

$$\sum_{t=1}^{T} \mathbb{1}(\mathcal{C}1') \leqslant \mathcal{O}\left(\frac{T \ln(T)}{\tau(\overline{\mu}_1(\sigma'(\tau)) - y_i)^2} \right). \tag{225}$$

Finally, we obtain that:

$$B1 \leqslant \mathcal{O}\left(\frac{T \ln(T)}{\tau(\overline{\mu}_1(\sigma'(\tau)) - y_i)^3} \right), \tag{226}$$

where the above inequality follows from Lemma C.14.

Let us analyse $B2$:

$$B2 = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{C}2') \underbrace{\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \,|\, \mathbb{1}(\mathcal{C}2') \right]}_{(*')} \right]. \tag{227}$$

Similarly to what has been done for term B1, applying Lemma 4.1 to $(*')$, we have that that term can be bounded by the same bound we would have for a process governed by a Binomial distribution $\text{Bin}(\cdot, \overline{\mu}_1(\sigma'(\tau)))$. Thus, applying Lemma 4 by [2] to such a distribution :

$$(*') \leqslant \mathcal{O}\left(\frac{1}{T} \right),$$

and, finally:

$$B2 \leqslant \mathcal{O}(1). \tag{228}$$

Choosing $x_i = \mu_i(T) + \frac{\Delta_i}{3}$ and $y_i = \overline{\mu}_1(\sigma'(\tau)) - \frac{\Delta_i}{3}$ and summing all the term concludes the proof. □

Theorem 7.2 (γ-SW-GTS Regret Bound)**.** *Under Assumption 5.1, setting* $\gamma \leqslant \min\left\{ \frac{1}{4\sigma^2_{var}}, 1 \right\}$*, the* γ*-GTS algorithm suffers an expected cumulative regret of:*

$$R(\gamma\text{-SWGTS}, T) \leqslant O\left(\sum_{i \neq i^*(T)} \Delta_i(T,0) \left(\frac{T \log(T(\Delta'_i(T;\tau))^2)}{\gamma\tau(\Delta'_i(T;\tau))^2} + \frac{T}{\tau} + \frac{\sigma'(T;\tau)}{\textbf{erfc}\left(\sqrt{\frac{\gamma\tau}{2}}(\overline{\mu}_1(\sigma'(T;\tau),\tau))\right)} \right) \right). \tag{15}$$

Proof. For ease of notation we set $\sigma'_i(T;\tau) = \sigma'_i(\tau), \overline{\mu}_1(\sigma'_i(T;\tau);\tau) = \overline{\mu}_1(\sigma'_i(\tau))$ and $\Delta'_i = \Delta_i$. For every suboptimal arm $i \in \{2, K\}$, let us define the thresholds x_i and y_i s.t. $\mu_i(T) < x_i < y_i < \overline{\mu}_1(\sigma'_i(\tau))$. Thanks to the above thresholds, we can define the following events for every $t \in T$:

- $E^{\mu}_i(t)$ as the event for which $\widehat{\overline{\mu}}_{i,t,\tau} \leqslant x_i$;
- $E^{\theta}_{i,t}$ as the event for which $\theta_{i,t,\tau} \leqslant y_i$, where $\theta_{i,t,\tau}$ denotes a sample generated for arm i from the posterior distribution at time t , i.e., $\mathcal{N}(\widehat{\overline{\mu}}_{i,t,\tau}, \frac{1}{\gamma T_{i_t,t,\tau}})$, being $T_{i_t,t}$ of trials at time t in the temporal window τ for arm i_t.

In such a framework $p_{i,t}$ is defined as $p_{i,t} = \Pr(\theta_{1,t,\tau} \geqslant y_i \,|\, \mathbb{F}_{t-1})$ Moreover, let us denote with $E_i^\mu(t)^\complement$ and $E_i^\theta(t)^\complement$ the complementary event $E_i^\mu(t)$ and $E_i^\theta(t)$, respectively. Let us focus on decomposing the probability term in the regret as follows:

$$\sum_{t=1}^{T} \mathbb{P}(I_t = i) \leqslant \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^\mu(t)^\complement)}_{=:P_A} + \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^\mu(t), E_i^\theta(t)^\complement)}_{=:P_B} + \underbrace{\sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^\mu(t), E_i^\theta(t))}_{=:P_C} + \underbrace{\frac{T}{\tau}}_{\text{Term due to the round robin every } \tau \text{ times}} . \tag{229}$$

Let us analyze each term separately.

Term A We have:

$$\begin{aligned}
P_A &= \sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^\mu(t)^\complement) && (230)\\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{I_t = i, E_i^\mu(t)^\complement\right\}\right] && (231)\\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{I_t = i, E_i^\mu(t)^\complement, T_{i,t,\tau} \leqslant \frac{\ln(T\Delta_i^2 + e)}{\gamma(x_i - \mu_i(T))^2}\right\}\right] + \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{I_t = i, E_i^\mu(t)^\complement, T_{i,t,\tau} \geqslant \frac{\ln(T\Delta_i^2 + e)}{\gamma(x_i - \mu_i(T))^2}\right\}\right] && (232)\\
&\leqslant \frac{T\ln(T\Delta_i^2 + e)}{\gamma\tau(x_i - \mu_i(T))^2} + \sum_{t=1}^{T} \Pr\left(E_i^\mu(t)^\complement \,|\, T_{i,t,\tau} \geqslant \frac{\ln(T\Delta_i^2 + e)}{\gamma(x_i - \mu_i(T))^2}\right) && (233)\\
&\leqslant \frac{T\ln(T\Delta_i^2 + e)}{\gamma\tau(x_i - \mu_i(T))^2} + \sum_{t=1}^{T} \frac{1}{T\Delta_i^2}, && (234)
\end{aligned}$$

Where we used Lemma C.14 and Lemma C.10 as we did in the proof of Theorem 5.1.

Term B Defining $L_i(T) = \frac{288\log(T\Delta_i^2 + e^6)}{\gamma\Delta_i^2}$, we decompose each summand into two parts:

$$\begin{aligned}
P_B &= \sum_{t=1}^{T} \mathbb{P}\left(I_t = i, E_i^\mu(t), E_i^\theta(t)^\complement\right) && (235)\\
&= \sum_{t=1}^{T} \mathbb{P}\left(I_t = i, T_{i,t,\tau} \leqslant L_i(T), E_i^\mu(t), E_i^\theta(t)^\complement\right) + \mathbb{P}\left(I_t = i, T_{i,t,\tau} > L_i(T), E_i^\mu(t), E_i^\theta(t)^\complement\right). && (236)
\end{aligned}$$

The first term is bounded by $L_i(T)\frac{T}{\tau}$ using Lemma C.14. Instead, regarding the second term:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, T_{i,t,\tau} > L_i(\tau), E_i^\theta(t)^\complement, E_i^\mu(t)\right) && (237)\\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^\theta(t)^\complement \,|\, T_{i,t,\tau} > L_i(T), E_i^\mu(t), \mathbb{F}_{t-1}\right)\right] && (238)\\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{P}\left(\theta_{i,t,\tau} > y_i \,|\, T_{i,t,\tau} > L_i(T), \overline{\widehat{\mu}}_{i,t,\tau} \leqslant x_i, \mathbb{F}_{t-1}\right)\right]. && (239)
\end{aligned}$$

In this setting, $\theta_{i,t,\tau}$ is a Gaussian random variable distributed as $\mathcal{N}\left(\overline{\widehat{\mu}}_{i,t,\tau}, \frac{1}{\gamma T_{i,t,\tau}}\right)$. We recall that an $\mathcal{N}\left(m, \sigma^2\right)$ distributed r.v. (i.e., a Gaussian random variable with mean m and variance σ^2) is stochastically dominated by $\mathcal{N}\left(m', \sigma^2\right)$ distributed r.v. if $m' \geqslant m$. Therefore, given $\overline{\widehat{\mu}}_{i,t,\tau} \leqslant x_i$, the

distribution of $\theta_{i,t,\tau}$ is stochastically dominated by $\mathcal{N}\left(x_i, \frac{1}{\gamma T_{i,t,\tau}}\right)$. Formally:

$$\mathbb{P}\left(\theta_{i,t,\tau} > y_i \,|\, T_{i,t,\tau} > L_i(T), \widehat{\overline{\mu}}_{i,t,\tau} \leqslant x_i, \mathbb{F}_{t-1}\right) \leqslant \mathbb{P}\left(\mathcal{N}\left(x_i, \frac{1}{\gamma T_{i,t,\tau}}\right) > y_i \,\middle|\, \mathbb{F}_{t-1}, T_{i,t,\tau} > L_i(T)\right). \tag{240}$$

Using Lemma C.9 we have:

$$\mathbb{P}\left(\mathcal{N}\left(x_i, \frac{1}{\gamma T_{i,t,\tau}}\right) > y_i\right) \leqslant \frac{1}{2} e^{-\frac{(\gamma T_{i,t,\tau})(y_i - x_i)^2}{2}} \tag{241}$$

$$\leqslant \frac{1}{2} e^{-\frac{(\gamma L_i(T))(y_i - x_i)^2}{2}} \tag{242}$$

which is smaller than $\frac{1}{T\Delta_i^2}$ because $L_i(T) \geqslant \frac{2\ln\left(T\Delta_i^2\right)}{\gamma(y_i - x_i)^2}$. Substituting into Equation (240), we get:

$$\mathbb{P}\left(\theta_{i,t,\tau} > y_i \,|\, T_{i,t,\tau} > L_i(T), \widehat{\overline{\mu}}_{i,t,\tau} \leqslant x_i, \mathbb{F}_{t-1}\right) \leqslant \frac{1}{T\Delta_i^2}. \tag{243}$$

Summing over t follows that $P_B \leqslant O\left(\frac{T}{\tau} L_i(T) + \frac{1}{\Delta_i^2}\right)$.

Term C For this term, we use Lemma 1 by [3]. Let us define $p_{i,t} := \mathbb{P}(\theta_{1,t,\tau} > y_i | \mathbb{F}_{t-1})$. We have:

$$\mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t) | \mathbb{F}_{t-1}) \leqslant \frac{1 - p_{i,t}}{p_{i,t}} \mathbb{P}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t) | \mathbb{F}_{t-1}). \tag{244}$$

Thus, we can rewrite the term P_C as follows:

$$P_C = \sum_{t=1}^{T} \mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t)) \tag{245}$$

$$= \sum_{t=1}^{T} \mathbb{E}[\mathbb{P}(I_t = i, E_i^{\mu}(t), E_i^{\theta}(t) | \mathbb{F}_{t-1})] \tag{246}$$

$$\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t)) \,\middle|\, \mathbb{F}_{t-1}\right]\right] \tag{247}$$

$$\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t))\right]. \tag{248}$$

We decompose Equation (248) into two contributions:

$$P_C \leqslant \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t), T_{1,t} \leqslant \sigma'(\tau))\right]}_{B1} +$$

$$+ \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{1}(I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t), T_{1,t} \geqslant \sigma'(\tau))\right]}_{A1}. \tag{249}$$

Analyzing term $A1$:

$$A1 \leqslant \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{1}\left(\overbrace{I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t), T_{1,t,\tau} \leqslant L_i(T), T_{1,t} \geqslant \sigma'(\tau)}^{\mathcal{C}1}\right)\right]}_{A} +$$

$$+ \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{1}\left(\overbrace{I_t = 1, E_i^{\mu}(t), E_i^{\theta}(t), T_{1,t,\tau} \geqslant L_i(T), T_{1,t} \geqslant \sigma'(\tau)}^{\mathcal{C}2}\right)\right]}_{B} \tag{250}$$

Let us tackle the term A by exploiting the fact that $\mathbb{E}[XY]=\mathbb{E}[X\mathbb{E}[Y\,|\,X]]$. This way, we can rewrite it as:

$$A=\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}(\mathcal{C}1)\underbrace{\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\,|\,\mathbb{1}(\mathcal{C}1)\right]}_{(*)}\right]. \tag{251}$$

In the following, we show that whenever condition $\mathcal{C}1$ holds $(*)$ is bounded by a constant. Let Θ_j denote a $\mathcal{N}\left(\overline{\widehat{\mu}}_{1,j},\frac{1}{\gamma j}\right)$ distributed Gaussian random variable, where $\overline{\widehat{\mu}}_{1,j}$ is the sample mean of the optimal arm's rewards played j times within a time window τ. Let G_j be a geometric random variable denoting the number of consecutive independent trials up to j included where a sample of Θ_j is greater than y_i. We will show that for any realization of the number of pulls within a time window τ such that condition $\mathcal{C}1$ holds, the expected value of G_j is bounded by a constant for all j.

Consider an arbitrary realization of $T_{1,t,\tau}=j$ that satisfies condition $\mathcal{C}1$. Observe that $p_{i,t}=\Pr\left(\Theta_j>y_i\,|\,\mathbb{F}_{\tau_j}\right)$ and:

$$\mathbb{E}\left[\frac{1}{p_{i,t}}\,|\,\mathbb{1}(\mathcal{C}1)\right]=\mathbb{E}_j\left[\mathbb{E}\left[\frac{1}{p_{i,t}}\,|\,\mathbb{1}(\mathcal{C}1),T_{1,t,\tau}=j\right]\right]=\mathbb{E}_{j_{|\mathcal{C}1}}\left[\mathbb{E}\left[\mathbb{E}\left[G_j\,|\,\mathbb{F}_{\tau_j}\right]\right]\right]=\mathbb{E}_{j_{|\mathcal{C}1}}\left[\mathbb{E}\left[G_j\right]\right]. \tag{252}$$

Notice that the term $\mathbb{E}[G_j]$ in Equation (252) is the same as the one we had in Equation (130) to derive bounds for the γ-GTS algortihtm. Relying on the same mathematical steps we bound it as follows:

$$\mathbb{E}[G_j]\leqslant e^{12}+5.$$

This shows a constant bound independent from j of $\mathbb{E}\left[\frac{1}{p_{i,t}}-1\right]$ for any j such that condition $\mathcal{C}1$ holds. Then, using Lemma C.14, A can be rewritten as:

$$A\leqslant(e^{12}+5)\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}(\mathcal{C}1)\right] \tag{253}$$

$$\leqslant(e^{12}+5)\frac{288T\ln(T\Delta_i^2+e^6)}{\gamma\tau\Delta_i^2}. \tag{254}$$

Let us tackle B by exploiting the fact that $\mathbb{E}[XY]=\mathbb{E}[X\mathbb{E}[Y\,|\,X]]$:

$$B=\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}(\mathcal{C}2)\underbrace{\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\,|\,\mathbb{1}(\mathcal{C}2)\right]}_{(**)}\right]. \tag{255}$$

We derive a bound for $(**)$ for large j as imposed by condition $\mathcal{C}2$. Consider then an arbitrary case in which $T_{i,t,\tau}=j\geqslant L_i(T)$ (as dictated by $\mathcal{C}2$), we have:

$$\mathbb{E}\left[\frac{1}{p_{i,t}}\,|\,\mathbb{1}(\mathcal{C}2)\right]=\mathbb{E}_j\left[\mathbb{E}\left[\frac{1}{p_{i,t}}\,|\,\mathbb{1}(\mathcal{C}2),T_{1,t,\tau}=j\right]\right]=\mathbb{E}_{j_{|\mathcal{C}2}}\left[\mathbb{E}\left[\mathbb{E}\left[G_j\,|\,\mathbb{F}_{\tau_j}\right]\right]\right]=\mathbb{E}_{j_{|\mathcal{C}2}}\left[\mathbb{E}\left[G_j\right]\right]. \tag{256}$$

Notice that the term $\mathbb{E}[G_j]$ in the last equation is the same that bounded in Theorem 5.1 for the regret of γ-GTS. Therefore, using the same proof line it is bounded by $\mathbb{E}[G_j]\leqslant\frac{1}{T\Delta_i^2}$.

For term $B1$, we made the same passages that we did for Equations (251) and (255), adding the $\delta_{TV}(\cdot,\cdot)$ term, yielding to:

$$B1\leqslant O\left(\frac{\sigma'(T;\tau)}{\textbf{erfc}(\sqrt{\frac{\gamma\tau}{2}}(\overline{\mu}_1(\sigma'(T;\tau),\tau))}+(e^{12}+5)\frac{288T\ln(T\Delta_i^2+e^6)}{\gamma\tau\Delta_i^2}+\frac{1}{\Delta_i^2}\right), \tag{257}$$

and summing up the terms concludes the proof. □

C Auxiliary Lemmas

In this section, we report some results that already exist in the bandit literature and have been used to demonstrate our results.

Lemma C.1 (Generalized Chernoff-Hoeffding bound from [3])**.** *Let $X_1,\dots,X_n$ be independent Bernoulli random variables with $\mathbb{E}[X_i]=p_i$, consider the random variable $X=\frac{1}{n}\sum_{i=1}^{n}X_i$, with $\mu=\mathbb{E}[X]$. For any $0<\lambda<1-\mu$ we have:*

$$\mathbb{P}(X\geqslant\mu+\lambda)\leqslant\exp\big(-nd(\mu+\lambda,\mu)\big),$$

and for any $0<\lambda<\mu$

$$\mathbb{P}(X\leqslant\mu-\lambda)\leqslant\exp\big(-nd(\mu-\lambda,\mu)\big),$$

where $d(a,b):=a\ln\frac{a}{b}+(1-a)\ln\frac{1-a}{1-b}$.

Lemma C.2 (Change of Measure Argument from [26])**.** *Let $(\Omega,\mathcal{F})$ be a measurable space, and $P,Q:\mathcal{F}\to[0,1]$. Let $a<b$ and $X\to[a,b]$ be a $\mathcal{F}$-measurable random variable, we have:*

$$\left|\int_\Omega X(\omega)\,dP(\omega)\;-\int_\Omega X(\omega)\,dQ(\omega)\right|\leqslant(b-a)\delta_{TV}(P,Q). \tag{258}$$

Lemma C.3 ([26], proposition 2.8)**.** *For a nonnegative random variable X, the expected value $\mathbb{E}[X]$ can be computed as:*

$$\mathbb{E}[X]=\int_0^\infty Pr(X>y)dy.$$

Lemma C.4 ([38], Theorem 2)**.** *Let us define $\underline{\mu}_n:=(\mu_1,\dots,\mu_n)$, $s\in(0,\dots,n)$ and $\mu\in(0,1)$. We have that the total variation distance between two variables $PB(\underline{\mu}_n)$ and $B_s(n,\mu)$ is:*

$$\delta_{TV}(PB(\underline{\mu}_n),B_s(n,\mu))\leqslant\begin{cases}C_1(s)\theta(\mu,\underline{\mu}_n)^{\frac{s+1}{2}}\frac{\left(1-\frac{s}{s+1}\sqrt{\theta(\mu,\underline{\mu}_n)}\right)}{(1-\sqrt{\theta(\mu,\underline{\mu}_n)})^2} & \text{if } \theta(\mu,\underline{\mu}_n)<1\\ C_2(s)\eta(\mu,\underline{\mu}_n)^{\frac{s+1}{2}}(1+\sqrt{2\eta(\mu,\underline{\mu}_n)})\exp(2\eta(\mu,\underline{\mu}_n)) & \text{otherwise}\end{cases}, \tag{259}$$

where $\theta(\mu,\underline{\mu}_n):=\frac{\eta(\mu,\underline{\mu}_n)}{2n\mu(1-\mu)}$, $\eta(\mu,\underline{\mu}_n):=2\gamma_2(\mu,\underline{\mu}_n)+\gamma_1(\mu,\underline{\mu}_n)^2$, $\gamma_k(\mu,\underline{\mu}_n):=\sum_{n'=1}^{n}(\mu-\mu_{n'})^k$, $C_1(s):=\frac{\sqrt{e}(s+1)^{\frac{1}{4}}}{2}$, $C_2(s):=\frac{(2\pi)^{\frac{1}{4}}\exp(\frac{1}{24(s+1)})2^{\frac{s-1}{2}}}{\sqrt{s!}(s+1)^{\frac{1}{4}}}$.

Lemma C.5 ([17], Theorem 1, Lemma 2)**.** *Using the quantities defined in the Lemma C.4,*

$$\frac{\theta(\overline{\mu},\underline{\mu}_n)}{124}\min\{1,n\overline{\mu}(1-\overline{\mu})\}\leqslant\delta_{TV}(PB(\underline{\mu}_n),Bin(n,\overline{\mu}))\leqslant\frac{1-\overline{\mu}^{n+1}-(1-\overline{\mu})^{n+1}}{(n+1)\overline{\mu}(1-\overline{\mu})}\gamma_2(\overline{\mu},\underline{\mu}_n) \tag{260}$$

where $\overline{\mu}$ is the mean of the components of the means' vector $\underline{\mu}_n$, i.e. $\overline{\mu}=\frac{\sum_{n'=1}^{n}\mu_{n'}}{n}$

Lemma C.6 (Beta-Binomial identity)**.** *For all positive integers $\alpha,\beta\in\mathbb{N}$, the following equality holds:*

$$F^{beta}_{\alpha,\beta}(y)=1-F^{B}_{\alpha+\beta-1,y}(\alpha-1), \tag{261}$$

where $F^{beta}_{\alpha,\beta}(y)$ is the cumulative distribution function of a beta with parameters α and β, and $F^{B}_{\alpha+\beta-1,y}(\alpha-1)$ is the cumulative distribution function of a binomial variable with $\alpha+\beta-1$ trials having each probability y.

Lemma C.7 ([10], Theorem 1 (iii))**.** *Let $Y\sim Bin(n,\lambda)$ and $X=\sum X_i$ where the $X_i\sim Bin(n_i,\lambda_i)$ are independent random variables for $i=1,\dots,k$ then:*

$$X\geqslant_{st}Y \text{ if and only if } \lambda\leqslant\overline{\lambda}_g, \tag{262}$$

$$X\leqslant_{st}Y \text{ if and only if } \lambda\geqslant\overline{\lambda}_{cg}, \tag{263}$$

where $X\geqslant_{st}Y$ means that X is greater than Y in the stochastic order, i.e. $Pr(X\geqslant m)\geqslant Pr(Y\geqslant m)\ \forall m$, and:

$$\overline{\lambda}_g=\left(\prod_{i=1}^{k}\lambda_i^{n_i}\right)^{\frac{1}{n}}, \tag{264}$$

$$\overline{\lambda}_{cg} = 1 - \left(\prod_{i=1}^{k} (1-\lambda_i)^{n_i} \right)^{\frac{1}{n}}. \tag{265}$$

Lemma C.8 ([1] Formula 7.1.13)**.** *Let Z be a Gaussian random variable with mean μ and standard deviation σ, then:*

$$\mathbb{P}(Z > \mu + x\sigma) \geqslant \frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} e^{-\frac{x^2}{2}} \tag{266}$$

Lemma C.9 ([1])**.** *Let Z be a Gaussian r.v. with mean m and standard deviation σ, then:*

$$\frac{1}{4\sqrt{\pi}} e^{-7z^2/2} < \mathbb{P}(|Z-m| > z\sigma) \leqslant \frac{1}{2} e^{-z^2/2}. \tag{267}$$

Lemma C.10 ([37] Corollary 1.7)**.** *Let $X_1, \ldots, X_n$ be n independent random variables such that $X_i \sim \text{SUBG}(\sigma^2)$, then for any $a \in \mathbb{R}^n$, we have*

$$\mathbb{P}\left[\sum_{i=1}^{n} a_i X_i > t \right] \leqslant \exp\left(-\frac{t^2}{2\sigma^2 |a|_2^2} \right), \tag{268}$$

and

$$\mathbb{P}\left[\sum_{i=1}^{n} a_i X_i < -t \right] \leqslant \exp\left(-\frac{t^2}{2\sigma^2 |a|_2^2} \right) \tag{269}$$

Of special interest is the case where $a_i = 1/n$ for all i we get that the average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, satisfies

$$\mathbb{P}(\bar{X} > t) \leqslant e^{-\frac{nt^2}{2\sigma^2}} \quad \text{and} \quad \mathbb{P}(\bar{X} < -t) \leqslant e^{-\frac{nt^2}{2\sigma^2}}$$

Lemma C.11 ([39],[42], Theorem 2.1 (2))**.** *Let $X \sim \text{PB}(p_1, \ldots, p_n)$, and $\bar{X} \sim \text{Bin}(n, \bar{p})$, for any convex function $g : [n] \to \mathbb{R}$ in the sense that $g(k+2) - 2g(k+1) + g(k) > 0$, $0 \leqslant k \leqslant n-2$, we have*

$$\mathbb{E}g(X) \leqslant \mathbb{E}g(\bar{X}), \tag{270}$$

where the equality holds if and only if $p_1 = \cdots = p_n$ of the poisson-binomial distribution are all equal to $\bar{p}$ of the binomial.

Fact C.1 (Bretagnolle-Hubner inequality)**.** *The Bretagnolle-Huber inequality states:*

$$\delta_{\text{TV}}(P,Q) \leqslant \sqrt{1 - \exp\left(-D_{\text{KL}}(P\|Q)\right)} \leqslant 1 - \frac{1}{2}\exp\left(-D_{\text{KL}}(P\|Q)\right) \tag{271}$$

Lemma C.12 ([23] Definition 1.2, [22])**.** *A random variable V taking values in $\mathbb{Z}_+$ is discrete log-concave if its probability mass function $p_V(i) = P(V = i)$ forms a log-concave sequence. That is, V is log-concave if for all $i \geqslant 1$:*

$$p_V(i)^2 \geqslant p_V(i-1)p_V(i+1) \tag{272}$$

Any Bernoulli random variable (that is, only taking values in $\{0,1\}$) is discrete log-concave. Further, any binomial distribution is discrete log-concave. In fact any random variable $S = \sum_{i=1}^{n} X_i$, where X_i are independent (not necessarily identical) Bernoulli variables, is discrete log-concave. Notice then that by definition $\frac{1}{p_V(i)}$ is discrete log-convex

Lemma C.13 ([21], Theorem 2 p.152, Remark 13 p.153, Remark 1 p.150)**.** *Let $1 \leqslant \alpha < r \leqslant \infty$ and let $q : \mathbb{Z} \to [0, \infty]$ be r-concave (Definition 1 p.150 [21], furthermore we highlight that for Remark 1 p.150 [21] to be ∞-concave is equivalent to be discrete log-concave). Then $\mathcal{J}^{\alpha} q$ is $(r-\alpha)$-concave, we assume $r - \alpha = \infty$ when $r = \infty$ and $r > \alpha$. Where the α-fractional (tail) sum of a function $q : \mathbb{Z} \to [0, \infty]$ is defined for every $\alpha > 0$ by the formula:*

$$\mathcal{J}^{\alpha} q(n) = \sum_{k=0}^{\infty} \binom{\alpha + k - 1}{k} q(n+k), \tag{273}$$

so that for a binomial pdf p_{bin}, being p_{bin} discrete log-concave (see C.12), follows that $\mathcal{J}^{\alpha} p_{bin}$ is ∞-concave on $\mathbb{Z}$ for $\alpha \geqslant 1$.

Lemma C.14 ([16], Lemma D.1)**.** *Let $A \subset \mathbb{N}$, and $\tau \in \mathbb{N}$ fixed. Define $a(n) = \sum_{t=n-\tau}^{n-1} \mathbb{1}(t \in A)$. Then for all $T \in \mathbb{N}$ and $s \in \mathbb{N}$ we have the inequality:*

$$\sum_{n=1}^{T} \mathbb{1}(n \in A, a(n) \leqslant s) \leqslant s\lceil T/\tau \rceil. \tag{274}$$

D Detailed Computations for the Instances of Section 6

First Instance. Let us upper bound the complexity index:

$$\begin{aligned}\Upsilon\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil,q\right)&=\sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\max_{y\in\{1,2\}}\{e^{-y\lambda n}-e^{-y\lambda(n+1)}\}^q\\&\leqslant e^{-q\lambda}+\sum_{n=2}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\max_{y\in\{1,2\}}\{e^{-y\lambda n}-e^{-y\lambda(n+1)}\}^q\\&\leqslant e^{-q\lambda}+\sum_{n=2}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\max_{y\in\{1,2\}}\{y\lambda e^{-y\lambda n}\}^q\\&=e^{-q\lambda}+2\lambda\sum_{n=2}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\lambda e^{-q\lambda n}\\&\leqslant e^{-q\lambda}+2\lambda\int_{n=1}^{+\infty}e^{-q\lambda n}\mathrm{d}n\leqslant\left(1+\frac{2}{q}\right)e^{-q\lambda},\end{aligned}$$

where we used $e^{-y\lambda n}-e^{-y\lambda(n+1)}\leqslant\max_{x\in[n,n+1]}\frac{\partial}{\partial x}(1-e^{-y\lambda x})=y\lambda e^{-y\lambda n}$ and bounded the summation with the integral.

Second Instance. Let us lower bound the complexity index:

$$\begin{aligned}\Upsilon\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil,q\right)&=\sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\left(\frac{2^{\lambda-1}}{(n+1)^\lambda}-\frac{2^{\lambda-1}}{(n+2)^\lambda}\right)^q\\&\geqslant\sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\left(\frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}}\right)^q,\end{aligned}$$

where we used $\frac{2^{\lambda-1}}{(n+1)^\lambda}-\frac{2^{\lambda-1}}{(n+2)^\lambda}\geqslant\min_{x\in[n,n+1]}\frac{\partial}{\partial x}\left(1-\frac{2^{\lambda-1}}{(x+1)^\lambda}\right)=\frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}}$. For $q(\lambda+1)>1$, we proceed as follows:[9]

$$\sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\left(\frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}}\right)^q\geqslant 2^{q(\lambda-1)}3^{-q(\lambda+1)}\lambda^q=O(\lambda^q).$$

Instead, for $q(\lambda+1)=1$, we bound the summation with the integral:

$$\begin{aligned}\sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\left(\frac{2^{q(\lambda-1)}\lambda^q}{n+2}\right)^q&\geqslant\int_{n=1}^{(1-2\epsilon)\frac{T}{K}}\left(\frac{2^{q(\lambda-1)}\lambda^q}{n+2}\right)\mathrm{d}n\\&\geqslant 2^{q(\lambda-1)}\lambda^q\log\left((1-2\epsilon)\frac{T}{K}-\frac{2}{3}\right)=O(\lambda^q\log T).\end{aligned}$$

Finally, for $q(\lambda+1)<1$, we still bound the summation with the integral:

$$\begin{aligned}\sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil}\left(\frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}}\right)^q&\geqslant\int_{n=1}^{(1-2\epsilon)\frac{T}{K}}\left(\frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}}\right)^q\mathrm{d}n\\&\geqslant\frac{2^{q(\lambda-1)}\lambda^q}{1-q(\lambda+1)}\left(\left((1-2\epsilon)\frac{T}{K}\right)^{1-q(\lambda+1)}-3^{1-q(\lambda+1)}\right)=O(\lambda^qT^{1-q(\lambda+1)}).\end{aligned}$$

Now, recalling that the instance-dependent component of the regret of Theorem 6.1 is in the order of $T^q\Upsilon\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil,q\right)$, we have for the three cases the optimal choice of q that minimizes the regret:

$$q(\lambda+1)>1\Longrightarrow q\downarrow\frac{1}{\lambda+1}\Longrightarrow O\left(\lambda^{\frac{1}{\lambda+1}}T^{\frac{1}{\lambda+1}}\right);$$

[9] We use big-O notation to highlight the dependences on $\lambda\to 0$ and $T\to+\infty$.

$$q(\lambda+1)=1 \Longrightarrow q=\frac{1}{\lambda+1} \Longrightarrow O\left(\lambda^{\frac{1}{\lambda+1}} T^{\frac{\lambda}{\lambda+1}} \log T\right);$$
$$q(\lambda+1)<1 \Longrightarrow q \uparrow \frac{1}{\lambda+1} \Longrightarrow O\left(\lambda^{\frac{1}{\lambda+1}} T^{\frac{1}{\lambda+1}}\right).$$

Thus, we have that the bound of the instance-dependent component of the regret is at least $O\left(T^{\frac{1}{\lambda+1}}\right)$.

E Numerical Simulations Parameters and Reproducibility Details

E.1 Parameters

The choices of the parameters are those suggested by the authors:

- Rexp3: $V_T = K$ as we've considered bounded rewards within zero and the maximum global variation possible is equal to the number of arms of the bandit; $\gamma = \min\left\{1, \sqrt{\frac{K\log K}{(e-1)\Delta_T}}\right\}, \Delta_T = \left\lceil (K\log K)^{1/3}\left(T/V_T\right)^{2/3}\right\rceil$ ([8]);
- KL-UCB: $c = 3$ as required by the theoretical results on the regret provided by [18];
- Ser4: according to what suggested by [4] we selected $\delta = 1/T, \epsilon = \frac{1}{KT}$, and $\phi = \sqrt{\frac{N}{TK\log(KT)}}$;
- SW-UCB: as suggested by [18] we selected the sliding-window $\tau = 4\sqrt{T\log T}$ and the constant $\xi = 0.6$;
- SW-KL-UCB as suggested by Garivier & Moulines ([19] 2011) we selected the sliding-window $\tau = \sigma^{-4/5}$;
- SW-TS: as suggested by [46] for the smoothly changing environment we set $\beta = 1/2$ and sliding-window $\tau = T^{1-\beta} = \sqrt{T}$.
- R-ed-UCB: the window is set as $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ as suggested by the authors ([31]), $\epsilon \in (0, \frac{1}{2})$, being $N_{i,t-1}$ the numbers of plays of the $i-th$ arm up to time t.

E.2 Environment

To evaluate the algorithms in the rested setting with $K = 15$ arms over a time horizon of $T = 50,000$ rounds. The payoff functions $\mu_i(\cdot)$ have been chosen in these families:

$$F_{\text{exp}} = \left\{ f(n) \mid f(n) = c\left(1 - e^{-an}\right)\right\}, \tag{275}$$

$$F_{\text{poly}} = \left\{ f(n) \mid f(n) = c\left(1 - b\left(n + b^{1/\rho}\right)^{-\rho}\right)\right\}, \tag{276}$$

where $a, c, \rho \in (0, 1]$ and $b \in \mathbb{R}_{\geqslant 0}$ are parameters, whose values have been selected randomly. The complete settings and function selection method, in compliance with what has been presented by [31], have been provided in the attached code.

E.3 Experimental Infrastructure

In this section, we provide additional information for the full reproducibility of the experiments provided in the main paper.

The code has been run on an AMD Ryzen 7 4800H CPU with 16 GiB of system memory. The operating system was Windows 11, and the experiments have been run on Python 3.8. The libraries used in the experiments, with the corresponding versions, were:

- `matplotlib==3.3.4`
- `tikzplotlib==0.10.1`
- `numpy==1.20.1`

On this architecture, the average execution time of each algorithm takes an average of ≈ 30 sec for a time horizon of $T = 50,000$.

E.4 15-arms Numerical Simulation Results

The results of the numerical simulation presented in Section 8 are reported in Figure 5. The results show how the methods that have been designed for the restless case are performing worse than the one we presented in our paper. The only exception is the `Beta-SWTS` that we showed to have also nice theoretical properties in the SRRB setting. Overall, the comparison with such methods do not invalidate the conclusions we drew in the main paper.

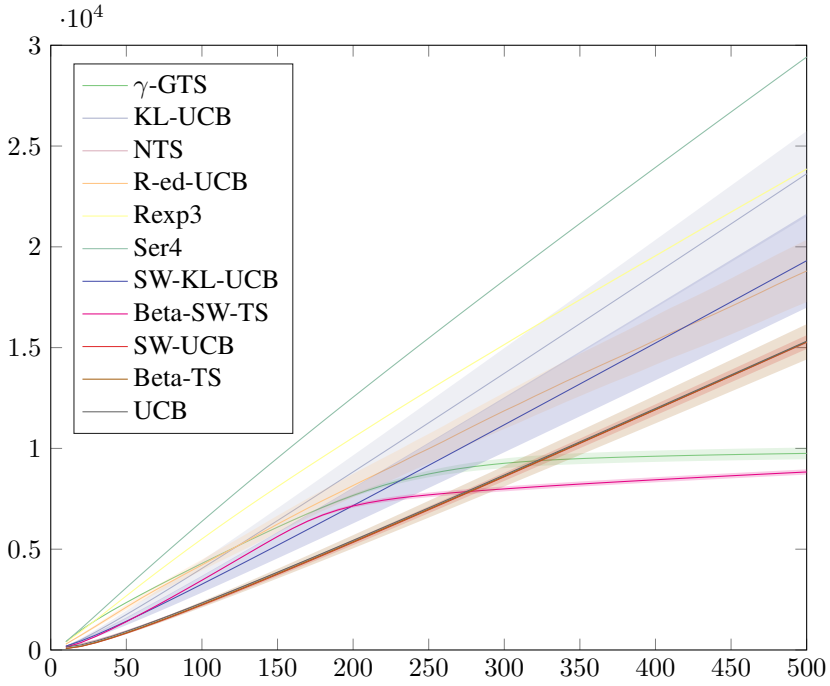


Figure 5: Regret in the 15-arm environment.