On Slowly-Varying Non-Stationary Bandits

Ramakrishnan Krishnamurthy , Aditya Gopalan

Keywords: Multi-armed bandits, Slowly-varying Non-stationary rewards, Change point detection.

Summary

We consider minimisation of dynamic regret in non-stationary multi-armed bandits with a slowly varying property. Namely, we assume that arms' rewards are stochastic and independent over time, but that the absolute difference between the expected rewards of any arm at any two consecutive time-steps is at most a drift limit $\delta > 0$. For this setting that has not received enough attention in the past, we give a new algorithm and establish the first instance-dependent regret upper bound for slowly varying non-stationary bandits. The analysis, in turn, relies on a novel characterization of the instance as a *detectable gap* profile that depends on the expected arm reward differences. We also provide the first minimax regret lower bound for this problem, enabling us to show that our algorithm is essentially minimax optimal. Also, this lower bound we obtain establishes that the seemingly easier slowly-varying bandits problem is at least as hard as the more general total variation-budgeted bandits problem in the minimax sense. We complement our theoretical results with experimental illustrations.

Contribution(s)

1. We design a new algorithm for the problem of slowly-varying non-stationary multi-armed bandits. We show an instance-dependent regret upper bound for this algorithm. For this, we come up with a novel instance-dependent quantity that we call 'detectable gap'.
   Context: To the best of our knowledge, this is the first instance-dependent regret bound for the slowly-varying settings. Instance-dependent bounds have so far been elusive in any continuously varying bandits (both total variation-budgeted setting and slowly-varying setting).
2. We show a minimax regret upper bound for our algorithm and establish that it is minimax optimal.
   Context: Besbes et al. (2014) already show a minimax optimal algorithm for the more general total variation-budgeted bandits problem, so ours is not the first/only minimax optimal algorithm.
3. We show a fundamental lower bound for slowly-varying non-stationary bandits problem. This bound matches our upper bound and also matches the known lower bound of the total variation-budgeted bandits problem. This establishes that the more constrained slowly-varying setting is at least as hard (in a worst case sense) as the more general total variation-budgeted setting.
   Context: To the best of our knowledge, this is the first lower bound for the slowly-varying non-stationary bandits problem.
4. We experimentally evaluate the performance of our new algorithm, and compare with existing approaches from the literature.
   Context: None.

On Slowly-Varying Non-Stationary Bandits

Ramakrishnan Krishnamurthy [1], **Aditya Gopalan** [2]
rk4312@nyu.edu, aditya@iisc.ac.in

[1]**Computer Science, Courant Institute of Mathematical Sciences, New York University**
[2]**Electrical Communication Engineering, Indian Institute of Science**

Abstract

We consider minimisation of dynamic regret in non-stationary multi-armed bandits with a slowly varying property. Namely, we assume that arms' rewards are stochastic and independent over time, but that the absolute difference between the expected rewards of any arm at any two consecutive time-steps is at most a drift limit $\delta > 0$. For this setting that has not received enough attention in the past, we give a new algorithm and establish the first instance-dependent regret upper bound for slowly varying non-stationary bandits. The analysis, in turn, relies on a novel characterization of the instance as a *detectable gap* profile that depends on the expected arm reward differences. We also provide the first minimax regret lower bound for this problem, enabling us to show that our algorithm is essentially minimax optimal. Also, this lower bound we obtain establishes that the seemingly easier slowly-varying bandits problem is at least as hard as the more general total variation-budgeted bandits problem in the minimax sense. We complement our theoretical results with experimental illustrations.

1 Introduction

Reinforcement learning, and specifically bandit optimization, in dynamically changing environments has remained an active topic of study in machine learning. A variety of non-stationary bandit settings have been studied incorporating a range of structural assumptions. At one end are classical stochastic models such as restless bandits (Whittle, 1988), where the state of the arms governs the bandit problem at any instant, but the transitions between these problems (states) follow probabilistic dynamics. At the other extreme are settings with non-stochastic and arbitrarily changing rewards such as prediction with expert advice (and the EXP3 algorithm)(Cesa-Bianchi & Lugosi, 2006; Auer et al., 2002). In between these extremes lie settings of changing environments where the adversary (environment) is assumed to be limited in its ability to change the rewards, i.e., a structural constraint is enforced on the amount of change in the rewards across time. These include the abrupt change (or switching experts) model (Garivier & Moulines, 2011), where at most M arbitrary changes to the reward distributions are allowed in the entire time horizon, and the variation-budgeted (drifting) change model (Besbes et al., 2014), in which the total magnitude of changes (of rewards) across successive time steps is constrained to be within an overall budget.

In this paper, we focus on slowly-varying bandits—a different and arguably more commonly encountered, yet less studied, model of non-stationary bandits. In this setting, the arms are allowed to change arbitrarily over time as long as the amount of change in their mean rewards between two successive time steps is bounded uniformly across the horizon. Many real-world settings naturally involve observables whose distributions are 'smooth' over time, in the sense that their instantaneous rate of change is not too large, e.g., slowly drifting distributions in natural language tasks (Lu et al., 2020), data from physical transducers (position, velocity, power, temperature, chemical concentration), and slowly fading wireless channels (Tse & Viswanath, 2005). Though the slowly-varying

bandit setting is subsumed by the total variation-budgeted model with an appropriate budget, the hope is that the local smoothness of the rewards with time can be exploited by the learner to perform better.

1.1 Our contributions

We make the following contributions in this regard.

1. We give a new algorithm, SNR, for slowly-varying multi-armed bandits. The algorithm is based on the principle of adaptive exploration followed by commitment in phases—a strategy that is known to be optimal for stationary stochastic bandits— but where the commitment phase is adjusted depending on the smoothness constraint on the rewards. The design of the algorithm also involves the estimation of a novel property of the bandit instance called the detectable gap profile, which essentially characterises the gap between the local averaged means of the arms that can be reliably detected, across time.
2. We derive a regret bound for this algorithm in terms of the detectable gap profile of the bandit instance (Theorem 1). The bound, to our knowledge, is the first instance-dependent regret guarantee for any algorithm for drifting stochastic bandits. Moreover, the worst case regret bound for a horizon T, over all instances for a given constraint δ on instantaneous reward change, is shown to be $O(TK^{1/3}\delta^{1/3})$ (Theorem 2).
3. We complement this worst case regret upper bound with a matching fundamental lower bound (Theorem 3) —the first of its kind for slowly-varying bandits— that shows that our algorithm is order-wise minimax-optimal. Interestingly, the minimax regret rate happens to be the same as that of the total variation-budgeted (with equivalent budget δT) setting, establishing that the more constrained slowly-varying setting is at least as hard (in a worst case sense) as the more general total variation-budgeted setting.
4. We evaluate the performance of our new algorithm and existing approaches, showing that it outperforms the other approaches in some synthetic experiments (Section 6).

1.2 Related Work

Non-stationarity in bandit optimization has been extensively studied and dates back to the seminal work of Whittle (1988) on restless bandits, in which arms' states (and thus their reward distributions) change according to Markovian dynamics (also see Slivkins & Upfal (2008)). The past few decades, however, have witnessed the growth of hybrid adversarial and stochastic bandit models for non-stationarity, where the distributions of arms across time can be set by an adversary ahead of time in an arbitrary manner.

Switching and total variation-budgeted non-stationary MAB. Among the many examples of non-stationary models are the following two: First, the abruptly changing (or switching) bandits setting (Garivier & Moulines, 2011; Mukherjee & Maillard, 2019; Auer et al., 2019), where the distributions are piecewise stationary and can arbitrarily change at unknown time steps. And second, the total variation-budgeted (or the 'drifting' or the 'continuously changing') bandits setting introduced by Besbes et al. (2014; 2019) where $V_T = \sum_{t=2}^{T} \|\mu_t - \mu_{t-1}\|_\infty$ is the total-variation budget that puts a constraint on the total amount of successive changes in arms' reward means. The authors show that it is possible to devise algorithms with guarantees on the dynamic regret that depend on the variation budget V_T and total time horizon T of order roughly[1] $\tilde{O}\left(V_T^{1/3}K^{1/3}T^{2/3}\right)$. Table 1 juxtaposes our results against these relevant results from literature.

Other non-stationary bandit settings. Subsequently, the notions of non-stationarity have been well studied under other bandit settings. (Kim & Tewari, 2020; Cheung et al., 2019; Russac et al., 2019) consider problems in the linear bandits (of d dimensions) with drifting non-stationarity

[1] $\tilde{O}$ (.) hides logarithmic factors.

Table 1: A comparison of different regret bounds from the literature for common structures/models of non-stationarity. We highlight some observation below.
(a) For the abruptyly changing (switching) model of non-stationarity that has sequential periods (or blocks of time) of stationarity, the instance-dependent regret of an arm is expressed as a function of the sub-optimality gap Δ_j in a block j summed over all M blocks. In their work, the term H_j is a measure of statistical distinguishability of block j from the preceding block $j-1$.
(b) Unlike in switching model above, however, in the continuously-varying models (total variation-budgeted or slowly-varying), these notions such as blocks and a single value $\Delta_.$ of gap therein don't apply. Thus, characterizing instance-dependent regret expression is non-obvious and has remained elusive. The $\lambda_{a,min}$ term we use to give such an instance-dependent regret bound above shall be formalized in Theorem 1.
(c) On the minimax regret bounds front, we discover that the slowly-varying bandits and total-variation budgeted bandits have the same minimax regret upper and lower bounds with a budget $V = T\delta$. In this lens, it is to be noted that the minimax regret we show is indeed sublinear in T when $\delta = V/T$ is a fraction that is $o(1)$ w.r.t. T.

Scenario	Instance-dependent upper bound - $\tilde{O}(.)$	Minimax upper bound - $\tilde{O}(.)$	Minimax lower bound - $\Omega(.)$
Abruptly switching. (M breakpoints)	$\sum_{\text{arms } a} \sum_{\text{blocks } j} \frac{\log T/H_j}{\Delta_{a,j}}$ (Mukherjee & Maillard, 2019)	$\sqrt{MKT}$ (Garivier & Moulines, 2011)	$\sqrt{MKT}$ (Wei et al., 2016)
Total variation-budgeted. (budget V)	—	$V^{1/3}K^{1/3}T^{2/3}$ Besbes et al. (2014)	$V^{1/3}K^{1/3}T^{2/3}$ Besbes et al. (2014)
Slowly Varying. (δ drift limit)	$\sum_{\text{arms } a} \sum_{\text{blocks } j} \frac{\log T}{\lambda_{a,min}(j)}$ **[Our Work]**	$\delta^{1/3}K^{1/3}T$ **[Our Work]** Besbes et al. (2014)	$\delta^{1/3}K^{1/3}T$ **[Our Work]**

regime, while Russac et al. (2020) look at generalized linear bandits with switching non-stationarity and both provide optimal minimax regret bounds Keskin & Zeevi (2017) work on the dynamic pricing problem where the demand over time is assumed to be drifting adhering to a variation metric, and Saha & Gupta (2022) investigate both switching and drifting rewards in a Duelling bandits setup. Most of these mentioned works use UCB-based approaches tailored to the non-stationary setup—such as only using samples from a sliding window of the recent past, or applying lesser (decaying) weightage to samples the further they are in the past—to impart more importance to newer samples over older ones. On that note, Zhao et al. (2020) handle non-stationarity by restarting algorithms to discard old information, and Wu et al. (2018) employ a master-slave paradigm, where the master bandit optimizes over many slave bandits that pull arms based on their knowledge. Manegueu et al. (2021) provide an algorithm and analysis that handles multiple closely related types of non-stationarity at once and provide an algorithm that uses restarting mechanism on top of change-point detection procedures to handle non-stationarity.

Parameter-free algorithms There is a different thread of work that has removed the assumption of knowing the non-stationary parameter upfront and has incrementally progressed (Karnin & Anava, 2016; Luo et al., 2018; Cheung et al., 2022) to get to optimal minimax regret upper bounds (Auer et al., 2019; Chen et al., 2019). Further, Wei & Luo (2021) provide a (non-stationary) parameterless black-box meta-algorithm that converts an minimax-optimal stationary bandit algorithm into a minimax optimal non-stationary bandit algorithm. However, an instance-dependent regret bound characterization has so far remained elusive under any sort of drifting reward constraint even when the non-stationary parameter is assumed to be known. Our results not only match the optimal regret

rates in a minimax sense, but we also give a more instance-dependent regret characterisation for our algorithm in the slowly-varying drifting bandits setting.

Lower Bounds. In terms of lower bounds, Besbes et al. (2014) provide a minimax lower bound of $\Omega(V_T^{1/3}K^{1/3}T^{2/3})$ that matched their algorithm's upper bound (upto logarithmic factors), thereby closing the door on any possible improvements to the total variation-budgeted setting. However, we show an identical lower bound for slowly-varying bandits setting as well, with equivalent per-round drift limit. This is a stronger result as our minimax lower bound applies to a more restricted class of problems instances.

Slowly/smoothly varying bandits. Looking at works whose setups are very close to the slowly-varying non-stationarity that we consider, the setting where rewards are Lipschitz continuous over time (with the Lipschitz constant being similar in function to our drift limit δ) is considered in Combes & Proutiere (2014) who provide asymptotical regret guarantees, and in Trovo et al. (2020) who provide a Thompson Sampling based approach to obtain an instance specific expression for the regret. However, both these works additionally assume that the seperation between two arms (in terms of reward means) is arbitrarily small only for a limited number of time-steps, which makes the problem instance more manageable from an algorithmic perspective of distinguishing sub-optimal arms. Recently, Jia et al. (2023) consider smoothly-varying rewards that are β-Hölder functions over time for the two-armed case, and specifically exploit additional smoothness when $\beta = 2$ to show improved upper bounds. Additionally, a lower bound expression that is $\Omega(T^{2/3})$ for 1-Hölder (or Lipschitz) reward function is shown without characterising dependence on the Lipschitz constant or the number of arms. Perhaps the only other work that studies the exact slowly-varying non-stationary bandit setting that we consider here is that of Wei & Srivatsva (2018). They modify the sliding-window-UCB algorithm of Garivier & Moulines (2011) to employ windows that grow in size with time to get the SW-UCB# algorithm and show minimax regret upper bounds (ignoring number of arms K) of $O(\delta^{1/4}T)$. In our work, we show an improved bound of $O(\delta^{1/3}T)$, where $\delta \in [0, 1]$ is the drift limit that shall be formally introduced in the next section.

2 Setting and Preliminaries

We consider bandits with arm/action set $\mathcal{A} = \{1, 2, \ldots, K\}$ and a time horizon of T. At time $t \in [T] := \{1, 2, \ldots, T\}$, when arm $a \in \mathcal{A}$ is played by a bandit algorithm depending on only past history, a stochastic reward $\widehat{\mu}_{a,t}$ drawn from a Bernoulli [2] distribution with expected value $\mu_{a,t} \in [0, 1]$ is obtained. We denote by $\mu_t := (\mu_{1,t}, \mu_{2,t}, \ldots, \mu_{K,t})$ the expected reward tuple at time $t \in [T]$. Write $\mu := (\mu_1, \mu_2, \ldots, \mu_T)$ to be the expected reward profile of the bandit instance.

A bandit instance μ is defined to be *slowly-varying* with *drift limit* $\delta > 0$ (denoted as $\mu \in S_\delta$) if the expected reward profile satisfies

$$\forall a \in \mathcal{A}, t \in [T-1], |\mu_{a,t} - \mu_{a,t+1}| \leq \delta. \tag{1}$$

In other words, for an arm $a \in \mathcal{A}$ at a time step t, the expected reward $\mu_{a,t}$ can drift in value to some $\mu_{a,t+1}$ (at the next time-step) by at most δ.

Definition 1 (Regret)**.** *For a bandit instance μ, the (expected) regret incurred by an algorithm (or policy)* ALG *is* $R(\text{ALG}) = \sum_{t=1}^{T} \mu_t^* - \mathbb{E}\left[\mu_{\text{ALG}(t),t}\right]$, *where* $\mu_t^* = \max_{a \in \mathcal{A}} \mu_{a,t}$ *is the mean reward of the optimal arm at time t, and* $\text{ALG}(t) \in \mathcal{A}$ *is the arm played by* ALG *at time* $t \in [T]$.

Note that this is the *dynamic regret*, where the performance benchmark at each time-step is the expected reward of the optimal arm at that time-step (μ_t^*). This is a stronger (i.e., harsher) notion of regret compared to the classical notion of *static regret*, where the benchmark at all time-steps is the expected reward of the best single arm across the entire horizon in hindsight. The goal is to

[2] In general, our theoretical analysis and results hold for any sub-gaussian reward distributions with a suitably bounded variance.

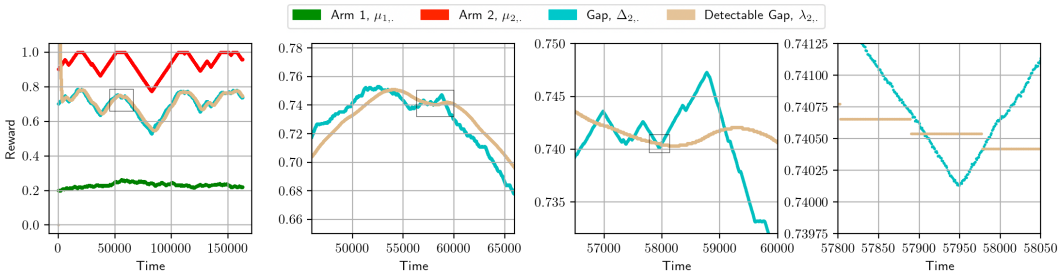


Figure 1: The figure depicts two arms' true reward means, and the sub-optimal arm's gap profile, and the detectable gap profile of one problem instance illustrated at 4 zoom levels. The box in each image is the region zoomed in for the next image (to its right). First, the detectable gaps (λ) reasonably positively correlates with the gaps (Δ), and initially (in general, whenever there is not a sufficient window size to detect a gap), for a short span, it takes the form $\sqrt{c_0 K \log T/t}$. Second, λ roughly trails Δ as it depends on a recent window (into the past) of Δ at every time-step. Third, λ is 'smoother' compared to Δ, as by design, its values are averaged over a window of samples. Fourth, although λ is described as a continuous optimization problem over $[0, 1]$, it actually transforms into a discrete optimization problem over different integral window sizes w. Thus, λ has a 'piece-wise constant' appearance.

learn to play arms to achieve low (dynamic) regret for any problem instance $\mu \in S_\delta$. We assume that the algorithm has access to the drift limit δ (or a suitable upper bound on it). This is reasonable as in practice there is often domain specific information available in advance about the drift of the quantity in consideration. We now introduce a novel characterization of a non-stationary bandit instance, and highlight its significance to the algorithm design and analysis to follow later.

Detectable Gap Profile. In the *stationary* stochastic multi-armed bandit problem, the suboptimality *gap* of some arm a is $\Delta_a := \mu^* - \mu_a$, where μ_a is the expected reward of arm a and μ^* is the expected reward of the optimal arm. These $\Delta_.$ quantities essentially characterise the attainable regret rate[3] of the problem instance. With *non-stationary* reward distributions, the notion of a (time-invariant) gap must be generalised to a *gap profile* For arm a, $\Delta_a := (\Delta_{a,t})_t$ where $\Delta_{a,t} := \mu_t^* - \mu_{a,t}$ is the difference at time t between the expected rewards of the optimal arm and arm a.

In a continuously non-stationary setting, however, this instantaneous arm gap Δ_t does not sufficiently capture the state of the problem instance (in terms of statistical distinguishability of arms) at the particular time t. It does not contain information about the nature of non-stationarity in the temporal neighborhood of that time-step t. We overcome this by introducing the notion of a *detectable gap profile* λ, which intuitively helps to characterise how hard it is to reliably estimate which arm is optimal and by how much at any time t. This is a derived quantity expressed in terms of the gap profile taken over a local window of time leading up to t.

Definition 2 (Detectable gap)**.** *For an arm $a \in \mathcal{A}$, we define its detectable gap $\lambda_a := (\lambda_{a,t})_t$ where,*

$$\lambda_{a,t} = \begin{cases} \max\limits_{b \in \mathcal{A} \setminus \{a\}} \left\{ \lambda \in (0,1] : \dfrac{1}{w(\lambda)} \sum\limits_{t'=t-w(\lambda)+1}^{t} \mu_{b,t'} - \mu_{a,t'} \geq \lambda \right\} & \text{, if such } \lambda \text{ exists.} \\ \sqrt{\dfrac{c_0 \log T}{t}} & \text{, otherwise.} \end{cases} \tag{2}$$

where, the summation is over a contiguous time-period of size $w(\lambda) := \lceil c_0 K \log T/\lambda^2 \rceil$ that terminates at t, and $c_0 = 144$ is a constant [4].

[3] assuming bounded/sub-gaussian rewards

[4] In fact, in the remainder of the paper, all notations of the form c_i shall be suitable universal constants.

Intuitively, $\lambda_{a,t} = \alpha$ for some time-step t implies that a (sub-optimality) gap of α for arm a can be detected with high probability by observing samples from all arms in the past $\approx 1/\alpha^2$ time-steps. In other words, if an observer directly received the true means of arms after playing them and considers them as observations from bounded random variables, then the choice of contiguous window $w(\alpha)$ essentially is upto how far in the past the observer would have to look to separate the arm from some other arm with high probability using Hoeffding's concentration inequality.

Fig. 1 depicts the detectable gap profile (λ) and instantaneous gap profile (Δ) of the sub-optimal arm in a two-armed bandit instance. We shall see that the detectable gap profile better reflects the nature of a non-stationary problem instance. Specifically, for our algorithm, we shall obtain a regret upper bound (to be shown in Theorem 1) as a function of the detectable gap profile, λ, which is an instance-dependent quantity.

3 Algorithm Description

Successive Elimination (SE) is a well-known algorithmic recipe for solving stationary stochastic bandits, built upon the Explore-Then-Commit (ETC) paradigm (see Slivkins (2019) for a text-book treatment of SE and ETC) In its classical form, SE adaptively pulls all arms in a round-robin fashion until it distinguishes an optimal arm from a sub-optimal arm with a high probability. Then, it drops the inferior arm indefinitely and continues this procedure with the remaining arms. Our algorithm, SnR (**S**nooze a**n**d **R**espawn) gracefully adapts this Successive Elimination paradigm to the non-stationary bandits setup. It pulls arms in a round-robin fashion until it can assertively identify/distinguish a sub-optimal arm in a 'local average' sense, and then plays only the remaining arms for a period of time until the earlier distinguished sub-optimal arm can possibly become the optimal arm due to the non-stationarity.

Specifically, SnR (pseudocode in Algorithm 1) runs in a series of episodes for every arm. Every episode of an arm $a \in \mathcal{A}$ begins with an *active phase*, during which SnR pulls arm a in every round as a part of its round-robin play. At the end of every time-step (or every round [5] of round-robin play), it performs a statistical test (Line 6) to detect a *clear gap* in reward means of arm a compared to some other arm b. If the test succeeds, the algorithm concludes that arm a is sub-optimal (in the current temporal neighbourhood) and the active phase of arm a ends, and it possibly *snoozes* arm a for a certain period of inactivity (Lines 10-11), termed as the *passive phase*. During this passive phase of arm a, SnR plays only the other arms that are in the active phases in their current episodes. At the end of this passive phase, arm a *respawns* to become active, and then its next episode begins.

We introduce some notations to describe the statistical test and for further analysis.

Let e_a be the total number of episodes (indexed as $1, 2, \ldots, e_a$) of arm $a \in \mathcal{A}$ in a run of SnR. For every episode $i \in [e_a]$ of every arm a, let $t_{a,i}$ denote the time after which the active phase of episode i of arm a begins. Note that $t_{a,i+1}$ corresponds to the time at which episode i of arm a ends. We also always have $t_{a,1} = 0$ and $t_{a,e_a+1} = T$, the end time of the final episode, for all arms a. Let $g_{a,i}$ denote the time at which the statistical test on line 6 of the algorithm passes, and write $\tau_{a,i} := g_{a,i} - t_{a,i}$ to denote the duration of the active phase of episode i of arm a. Let $w_{a,i}$ denote the number of rounds elapsed in the active phase, or equivalently, the number of times arm a has been played in the active phase $[t_{a,i} + 1, g_{a,i}]$.

For an arm a at time t in its ith episode, let $\omega_1^a, \omega_2^a, \ldots, \omega_w^a \leq t$ be the time steps in which arm a was played in its last w pulls, i.e., the the pull from each of the last w rounds of arm-play until time t. Write $\widehat{\mu}_{a,t}(w) := \frac{1}{w}\sum_{x=1}^{w} \widehat{\mu}_{a,\omega_x^a}$ to denote the empirical reward mean (or simply empirical mean) of arm a at time t measured/calculated from it's last w pulls. Here, $\widehat{\mu}_{a,\omega_x^a}$ is the actual reward the algorithm obtains on pulling arm a at time ω_x^a. Note that the true reward means $\mu_{a,\omega_\cdot^a}$ for different time-steps corresponding to these w pulls need not be the same.

[5] the terms 'time-step' and 'round' are sometimes used interchangeably in the bandit literature, but, here, a round corresponds to a sequence of time steps in a round-robin play of SnR in which all active arms are pulled once.

At a particular time, for a passive (snoozed) arm $a \in S$, denote by $\rho(a)$ the arm which arm a was compared against in the statistical test that resulted in the most recent snoozing of a. For an arm a, at a particular time, let $\bar{a}$ denote the arm that transitively eliminated (or caused snoozing of) arm a. Precisely, recursively define [6] $\bar{a} := \overline{\rho(a)}$ if a is passive, and $\bar{a} := a$ when a is active.

With these definitions, we define the statistical test as follows.

Definition 3 ($\widehat{\lambda}$-inferiority)**.** *Let $\widehat{\lambda} > 0$. At time t, an arm a in the active phase of episode i, is said to be $\widehat{\lambda}$-inferior to an arm b (written as $\bar{b} >_{\widehat{\lambda}} a$) if, for a window of $w := \lceil c_0 \log T / \widehat{\lambda}^2 \rceil$ rounds that falls entirely within episode i of arm a, we have*

$$LCB_{\bar{b},t}(w) > UCB_{a,t}(w) + 2r(w) - K\delta,$$

and the inequality holds for no other $\widehat{\lambda}' > \widehat{\lambda}$.

In the above definition, we write $LCB_{a,t}(w) := \widehat{\mu}_{a,t}(w) - r(w)$ (sim. $UCB_{a,t}(w) := \widehat{\mu}_{a,t}(w) + r(w)$) to denote the lower (sim. upper) confidence bound of an arm a's recent reward mean, and $r(w) := \sqrt{2 \log T / w}$ is the accuracy radius with w samples. Note that arm a is compared against arm $\bar{b}$ which can denote different arms in the w rounds in the window of consideration. Essentially, in time periods in which arm b is passive, arm $\bar{b}$ acts as a (superior) proxy for arm b. The constraint that all w rounds fall within the same (current) episode of arm a ensures that the rounds from which arm a's samples are observed are consecutive.

Using this statistical test (Line 6), SNR evaluates, at a time t, if some arm is $\widehat{\lambda}$-inferior to another (as in definition 3). The size of the window, w, in which the empirical means are calculated is dynamically chosen based on the empirical detectable gap, $\widehat{\lambda}$, itself that is being tested for. If the test passes, then SNR decides that arm a is sub-optimal and computes a sub-optimality buffer period, of duration buf $= 2/\delta . \sqrt{\log T / w_{a,i}}$ (Line 9), that begins at $t_{a,i}$, the start of the current episode. By time t, if the buffer period has not fully elapsed, we snooze the sub-optimal arm for a passive phase that runs until the end of the buffer period.

Optionally, we refer the reader to Appendix E for graphical illustrations (Figs. 6 and 7) of example algorithmic trajectories (with active/passive phases and statistical tests).

4 Theoretical Guarantees

Our first main result is a regret bound our algorithm in terms of the detectable gap profile of a slowly-varying bandit instance.

Theorem 1. *[Instance-dependent regret bound] If* SNR *is run with a drift limit parameter δ on a problem instance $\mu \in S_\delta$, then its expected regret , $R(\text{SNR})$, is upper bounded by*

$$c_6 \sum_{a \in \mathcal{A}} \sum_{j=1}^{m} \frac{1}{\lambda_{a,min}(j)} . \log T + c_8,$$

where $m = T/\tau$ is the number of blocks, each of length not more than $\tau = \min\{T, c_9 \delta^{-2/3} \log^{1/3} T\}$, and for every block $j \in [m]$ spanning a time period $b_j := [(j-1)\tau + 1, j\tau] \cap [T]$, define $\lambda_{a,min}(j) := \min_{t \in b_j} \lambda_{a,t}$. Here, c_is are suitable constants.

We also have the following upper bound on the worst-case regret (over instances in S_δ) for SNR.

Theorem 2. *[Minimax Upper bound] If* SNR *is run with a drift limit parameter δ on any problem instance $\mu \in S_\delta$, then it incurs an expected regret of $O(T\delta^{1/3} K^{1/3} \log^{1/3} T)$, where K is the number of arms.*

[6] the quantities $\rho(a), \bar{a}$ depend on the time t in consideration. But we refrain from explicity marking t in the notation as it shall be obvious from the context of usage.

Algorithm 1 SNR: Plays the non-stationary slowly-varying bandit problem instance

Input: Time horizon T, a set of arms $\mathcal{A} = \{1, 2, \ldots, K\}$ with sample access, the drift limit δ.
Output: Play an arm for every time-step.

```
Initialize active arms set A = 𝒜, snoozed arms set S = ∅.
Initialize episode index i(a) ← 1 for all arms a ∈ 𝒜.
for t = 1, 2, ..., T do
    x ← Least recently pulled arm in A.                    ▷ Play active arms in round-robin fashion
    Pull arm x and observe reward μ̂_{x,t}.
    if ∃ arms a, b ∈ A, ∃λ̂ ∈ [0, 1] such that b̄ >_λ̂ a then          ▷ As in Definition 3
        i ← i(a), arm a's current episode.
        Statistical test success time, g_{a,i} = t, Active phase duration, τ_{a,i} = g_{a,i} − t_{a,i}.
        Compute sub-optimality buffer/snooze period, buf = (2/δ) √(log T / w_{a,i}).
        if buf > τ_{a,i} then
            Snooze arm a, update A ← A \ {a}, and S ← S ∪ {(a, t_{a,i} + buf)}
        else
            Increment episode i(a) ← i + 1, then t_{a,i} ← t.     ▷ Episode ends without passive period
        end if
    end if
    if ∃(a, s) ∈ S : t ≥ s then
        Respawn arm a, update S ← S \ {(a, s)}, and A ← A ∪ {a}.
        Increment episode i(a) ← i + 1, then t_{a,i} ← t.     ▷ Episode ends after passive period elapses
    end if
end for
```

We finally complement the worst-case regret upper bound for SNR with a matching (upto logarithmic factors) universal minimax regret lower bound for any algorithm:

Theorem 3. *[Minimax Lower Bound] For any algorithm* ALG *and a drift limit $\delta > 0$, there exists a problem instance $\mu \in S_\delta$ such that,* ALG *incurs a expected regret of $\Omega(T\delta^{1/3}K^{1/3})$, where K is the number of arms.*

4.1 Discussion

The minimax regret lower bound we obtain establishes (constructively) that if the drift limit $\delta = \Omega(1)$, then it is impossible to achieve a sub-linear (in time) regret guarantee for any algorithm. The interesting problem space is thus when $\delta = o(1)$ as a function of total time T.

A basic sanity check is to evaluate our results for the stationary bandits setting, that is when $\delta = 0$, and the gap Δ is unchanged over time. In that case, in Theorem 1, the size of a block is $\tau = T$. From the detectable gap definition, either with $\lambda_t = \Delta$ from the first assignment, or $\lambda_t > \Delta$ from the second, we have $\lambda_t \geq \Delta$ at all times t. This gives us a regret bound of $O\left(\frac{1}{\Delta}\log T\right)$. One can also observe that SNR behaves similar to the classical Successive Elimination algorithm in this regime. After it distinguishes the optimal arm from the sub-optimal, that is, the statistical test passes, it computes a sub-optimality buffer $\lim_{\delta\to 0} \text{buf} = {}^{2}/_{\delta} \cdot \sqrt{\log T / w_{a,i}} \to \infty$, an infinite passive phase, i.e., it snoozes the sub-optimal arm indefinitely.

Moving on, for non-stationary instances with very small drift limits—specifically, for $\delta \leq O\left(T^{-3/2} \cdot \log^{1/2} T\right)$— applying Theorem 1 still yields a block size of $\tau = T$. With a similar analysis, this results in a regret bound of $O\left(\frac{1}{\Delta}\log T\right)$. It is noteworthy that, despite the mild non-stationarity of the instance, SNR achieves a logarithmic regret, which is typical in a stationary bandit instance.

This result in Theorem 2 is comparable with that of Besbes et al. (2014) who work on a total variation-budgeted setting. In our setting, a drift limit of δ per time-step translates to a cumulative drift limit of $T\delta$ over the entire time horizon. Precisely, for an arm a, $\sum_{t=1}^{T-1} |\mu_{a,t} - \mu_{a,t+1}| \leq T\delta$. This cumulative drift limit quantity is termed the total variation budget V_T in theirs, and we, here,

have $V_T = T\delta$. Substituting this in their regret upper bound $O(T^{2/3}V_T^{1/3}K^{1/3}\log^{1/3}T)$, we get $O(T\delta^{1/3}K^{1/3}\log^{1/3}T)$, the same upper bound as ours. Thus, we note that our minimax upper bound matches that of Besbes et al. (2014) which accommodates a more general setting.

On a more interesting note, on substituting $V_T = T\delta$ in their regret lower bound $\Omega(T^{2/3}V_T^{1/3}K^{1/3})$, we get $\Omega(T\delta^{1/3}K^{1/3})$, the same lower bound expression that we have established. That is, our minimax lower bound matches that of Besbes et al. (2014). This crucially establishes that the seemingly easier problem of slowly-varying bandits is at least as hard (in a minimax regret sense) as the more general problem of total variation-budgeted bandits. It can be interpreted that the ability/power of a total variation-budgeted problem instance to conserve budgets for long periods of time and manifest drastic non-stationarity in short bursts does not add to the difficulty (in terms of suffering higher regret) in playing the instance.

5 Proof Sketches

The formal proofs of all results are deferred to the Appendices B to D in the interest of space.

Proof sketch for Theorem 1 (Instance-dependent regret bound) We first restrict our attention to a 'good' event where all empirical means fall close to the reward means, which occurs with high probability. We establish useful properties about the nature of the active and passive phases, the sizes of episodes, the connection between the regret and the detectable gap, λ.

We show in Lemma 2 that in every episode, if an arm is snoozed, then it remains sub-optimal for the entirety of the passive phase. Towards this, we show that, for an arm a in its episode i, when the statistical test passes, we have, at some point of time (say t') in the active phase, that true reward means of the arms are well separated, i.e., $\mu_{b,t'} - \mu_{a,t'} \geq f$ for some gap f and some arm b. Then, we argue that the snooze period is carefully chosen based on this guaranteed true gap f, the drift limit δ, and the active phase duration $\tau_{a,i}$ such that arm a remains sub-optimal in it. As only the other arms are played in the passive phase, our algorithm incurs regret for arm a only during the active phase, which we try to bound next.

Next, we show in Lemma 4 a trajectory-dependent upper bound on arm a's regret $R_{a,i}$, in episode i, based on $\tau_{a,i}$, the duration of the active phase. We use the fact that the statistical test did not pass at time $g_{a,i}-1 = t_{a,i}+\tau_{a,i}-1$, specifically the fact that arm a was not identified to be $\widehat{\lambda}$-inferior to any other arm. We deduce that, their empirical means, and by the good event assumption, also their true means over a period of time were not well separable after playing the active arms in the round-robin fashion for $\tau_{a,i}-1$ time steps. Put simply, for these $\tau_{a,i}-1$ time-steps, the gaps between arm a and other better arms were small. This helps us get the following bound the regret of arm a in episode i as a function of the active phase duration $\tau_{a,i}$ and the rounds of samples observed $w_{a,i}$, or more precisely over carefully constructed blocks of time (as in Definition 5) that partition the active phase $\tau_{a,i}$, where $w(k)$s are the rounds/sample count within each block.

$$R_{a,i} \lesssim 1 + \sum_{k=1}^{j} c_2\sqrt{w(k).\log T}. \tag{3}$$

We draw the reader's attention to the fact that this expression in Equation (3) is an algorithmic trajectory-dependent quantity as it is a function of the active phase duration of the arm a that depends on the stochasticity/randomness in the observed rewards. Next, we convert this expression to a a problem instance-dependent quantity independent of the trajectory. Specifically, we go on to make a connection between the trajectory-specific active phase duration $\tau_{\cdot,\cdot}$ and the instance-specific detectable gap profile, λ.

Lemma 7 shows that, at time t in the active phase of some episode of arm a, if sufficiently many rounds w have passed, i.e., we have $\lambda_{a,t} \gtrsim \sqrt{c_0 \log T/w}$, then our algorithm *will* detect a gap in arms sufficient to declare $\bar{b} >_{\widehat{\lambda}} a$ for some arm b, and the statistical test shall pass. However, given that the

statistical test did not pass in the period upto $g_{a,i}-1 = t_{a,i}+\tau_{a,i}-1$, or more specifically at a certain points $(\alpha_k)_{k\in[j]}$ of choice (as shall be described in Definition 5) we derive the instance-dependent episodic regret bound: $R_{a,i} \lesssim \sum_{k=1}^{j} {}^{\log T}/_{\lambda_{\alpha_k}}$ in Lemma 6.

The remainder of the proof involves collecting all the episodic regrets of all arms (Lemma 11), smartly accounting them into blocks of time based on minimum possible size of any episode (Claims 9 and 10) to arrive at the final bound.

Proof sketch of Theorem 2 (Minimax regret bound). For every episode i of any arm a, we first consider the average/per time-step regret $R^{ave}_{a,i} = \frac{R_{a,i}}{t_{a,i+1}-t_{a,i}}$ and upper bound it. We argue that the blocks (where a block k is of time period $\tau(k)$ with $w(k)$ rounds of samples) are constructed within each episode's active phase in such a design (in Definition 5) that we have $\tau(k) \approx \frac{1}{\delta}\sqrt{{}^{\log T}/_{w(k)}}$.

As a tool of convenience, we introduce the notion of *inflated episodes* which are obtained by rearranging (or re-accounting) timesteps between episodes of different arms, thereby altering their durations but not their regret expressions. We show that the block size constraints extend to the inflated episode as $\widetilde{\tau}(k) \geq \tau(k) \approx \frac{1}{\delta}\sqrt{{}^{\log T}/_{w(k)}}$. By nature of construction of inflated episodes, we show that each round, after inflation, accounts one time-step for each of the K arms, thereby establishing that ${}^{\widetilde{\tau}(k)}/_{K} = w(k)$. From Equation (3), we have the regret in an episode i of arm a dependent on the breakdown of active phase duration $\tau_{a,i}$ into some j blocks as $R_{a,i} \lesssim 1 + \sum_{k=1}^{j} c_2\sqrt{w(k).\log T}$, where $w(k)$ and $\widetilde{\tau}(k)$ are the number of rounds and time steps respectivey of block k of an inflated episode.

We finally argue that the average per-time-step regret of each block in an inflated episode is suitably bounded as $O(\delta^{1/3}K^{-2/3}\log^{1/3}T)$. We multiply this by the length of the horizon, T, and the number of arms K to show the desired bound.

Proof sketch for Theorem 3 (Worst-case regret lower bound). We derive this result with information-theoretic arguments commonly employed in bandit literature (Garivier & Kaufmann, 2016). First, in Lemma 12, we adapt the standard 'change of measure inequality' for sequential sampling (see Garivier et al. (2019) for example) to a version that can accommodate non-stationary reward distributions. Second, we break the time horizon into suitably small blocks of size m, establish lower bounds within each of them, and finally aggregate them to arrive at the lower bound expression.

Towards proving the lower bound within a block, we consider a base instance ν, where all arms are identical and stationary with $Ber\,({}^{1}/_{2})$ reward distributions. We design a confusing instance, ν', where all but one arm are stationary with $Ber\,({}^{1}/_{2})$ rewards, whereas, the other arm (say, arm 1, with a distribution of $\nu'_{1,t}$ at time t within the block) exhibits non-stationarity as follows:

$$\nu'_{i,t} = \begin{cases} Ber\left(\frac{1}{2}\right) & \text{if } i = 2,3,\ldots,K \\ Ber\left(\frac{1}{2}+\frac{t-1}{m}.\varepsilon\right) & \text{if } i = 1, t \leq \left\lceil \frac{m}{2} \right\rceil \\ Ber\left(\frac{1}{2}+\frac{m-t}{m}.\varepsilon\right) & \text{if } i = 1, t > \left\lceil \frac{m}{2} \right\rceil. \end{cases} \tag{4}$$

Essentially, within a block, in the confusing instance, arm 1 starts identical to the other arms with $Ber\,({}^{1}/_{2})$ rewards, and for the first half of the block, drifts (by a suitably chosen value ε) upwards and away from a mean of ${}^{1}/_{2}$, and reaches a maximum reward mean gap at half-way point. Then, for the remainder of the block, it drifts towards (downwards) a mean of ${}^{1}/_{2}$ and reaches back the $Ber\,({}^{1}/_{2})$ reward distribution at the end of the block.

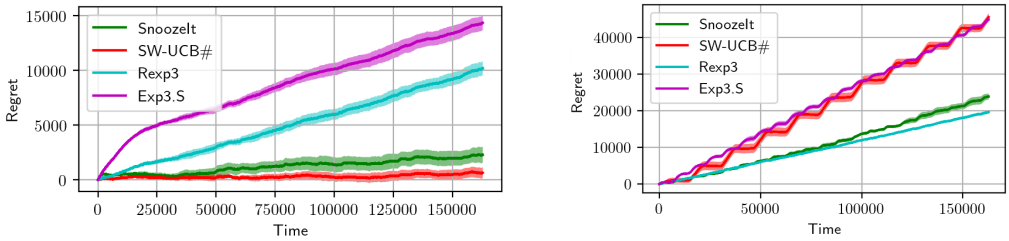
With this setup, we argue that when presented with a randomly chosen bandit problem instance between ν and ν', any algorithm is condemned to incur in expectation the stated minimum regret.

6 Experiments

We present some numerical experiments to better illustrate the characteristics of our algorithm. In this section, first, we compare the performance of our algorithm SnR against other algorithms in the literature for the slowly-varying (and also the closely related total variation-budgeted) bandits problem in Section 6.1. Second, we show the algorithmic trajectory of SnR for two problem instances in Appendix E.2.

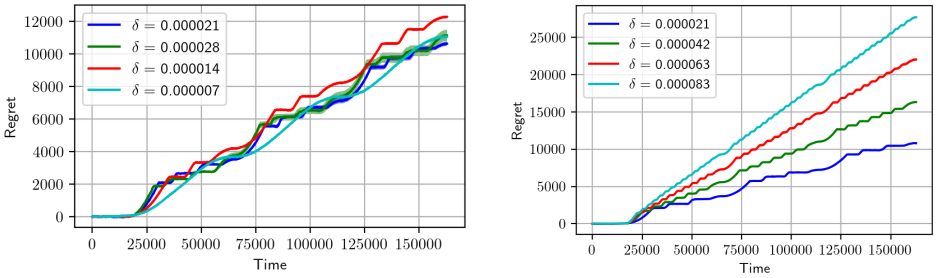
6.1 Comparison with Literature

We numerically evaluate the performance of SnR on synthetic two-armed bandit problem instances. We also implement algorithms from the literature and conduct a comparative study with different problem instances. In particular, we consider the following algorithms; (1) Rexp3 (Besbes et al., 2014), that divides the time into blocks upfront and repeatedly runs the Exp3 algorithm from scratch in each block, (2) Sw-Ucb# (Wei & Srivatsva, 2018), a soft exploration algorithm based on the arms' reward means' upper confidence bounds that are computed over a sliding window that enlarges with time. (3) Exp3.S (Besbes et al., 2019), a 'smoother' variant of Exp3 algorithm. We note that Rexp3 and Exp3.S are originally proposed for the more general total variation-budgeted setting. Also, all these algorithms have knowledge of the drift limit δ (or the appropriate drift parameter) in advance.



(a) (b)

Figure 2: Comparison of average regret among the 4 algorithms (SnR, Exp3.S, Rexp3, and Sw-Ucb#).



(a) (b)

Figure 3: Comparison of average regret incured by SnR on 4 problem instances characterized by different drift limit δ values.

We consider a time horizon of $T = \lfloor e^{12} \rfloor \simeq 1,60,000$, and report results averaged from 10 runs, the translucent regions (of the same colour) around the curves mark 1 standard deviation. The rewards are drawn from a Gaussian distribution with the specified mean, and a variance of ¼. [7].

First, in Fig. 2a, we run all the algorithms on a problem instance with a low value of drift $\delta = 1/(10 \cdot c_9 \log T) \simeq 2.1 \times 10^{-5}$ where the arms are well-separated and the identity of the optimal arm remains unchanged throughout (problem instance depicted in Fig. 4a in Appendix E.1). We observe that REXP3 and EXP3.S perform poorer compared to SW-UCB# and SNR. While SNR and SW-UCB# are very adaptive in their behaviour to the observed empirical reward means, we see that REXP3 predetermines the sizes of blocks (tailored to the drift parameter), and EXP3.S, after tuning the weights, additionally boosts up the weights of the arms equally without taking into account the observed empirical gaps.

Second, in Fig. 2b, we run all algorithms on a problem instance with a relatively larger value of drift $\delta = 1/(c_9 \log T) \simeq 2.1 \times 10^{-3}$, where both arms have short alternating stretches of stationarity and drifts, with the identity of the optimal arm toggling with every drift (problem instance depicted in Fig. 4b in Appendix E.1). Interestingly, we observe that SW-UCB# performs poorer compared to SNR. We hypothesise that SNR makes better use of short frequent stretches of stationarity: With the arms oscillating as we have here, when the sliding window encompasses the optimal stretches (and the drifts) of both arms, SW-UCB# does not take into account the recency of the samples within the sliding window. SNR, on the other hand, operates with dynamically-sized windows based on the observed detectable gaps that always capture the most suitably sized number of recent samples.

Third, in Fig. 3a, we evaluate the performance of our algorithm on different instances. In particular, we consider 4 instances with similar structure that differ only by the drift limit δ imposed, and have common periods of stationarity and drift (problem instance depicted in Fig. 5a in Appendix E.1). We observe that the average regret does not correlate with an increase in drift limit δ. The well separated nature of the arms in instances with a higher δ improves performance as increased detected (empirical) gaps lengthen the snooze periods. This potentially offsets the higher regret caused due to more frequent episodes with larger δ values leading to shorter snooze periods.

Finally, in Fig. 3b, we consider another set of 4 structurally similar instances with different δ, where arms have equal total cumulative drift and almost common periods of stationarity and drift (problem instance depicted in Fig. 5b in Appendix E.1). With the same maximum gaps among instances, the regret increases with drift limit δ, which is in line with what our theory lays down.

7 Conclusion & Future Work

In this paper, we studied the slowly-varying non-stationary bandits and provided multiple theoretical results that help better understand the difficulty of that class of problems. We came up with the construct of detectable gap profile which enabled us to show an instance-dependent characterization of the regret. We believe that our characterization of detectable gap profile is a fundamental property of non-stationarity in bandits, and may hold the key to more refined performance analysis, beyond merely the slowly-varying setting considered here. On that note, one interesting direction to pursue would be to characterize instant-dependent regret bounds in the more general total variation-budgeted setting.

It would also be interesting to explore the suitability of other flavours of algorithms such as Thompson sampling and Upper Confidence Bound (UCB) based approaches, especially if such algorithms can be designed in a parameter-free fashion, given the state-of-the-art algorithms provide such minimax guarantees in certain non-stationary settings.

[7] Indeed, the choice of a Gaussian draw with ¼ variance fits in with our theoretical analysis, specifically, the Hoeffding's inequality usage in Claim 1.

References

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pp. 138–158. PMLR, 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27:199–207, 2014.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337, 2019.

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

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pp. 696–726. PMLR, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087. PMLR, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: Learning to optimize under nonstationarity. *Management Science*, 68(3):1696–1713, 2022.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pp. 521–529. PMLR, 2014.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027, 2016.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pp. 174–188. Springer, 2011.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Su Jia, Qian Xie, Nathan Kallus, and Peter I Frazier. Smooth non-stationary bandits. *arXiv preprint arXiv:2301.12366*, 2023.

Zohar S Karnin and Oren Anava. Multi-armed bandits: Competing with optimal sequences. *Advances in Neural Information Processing Systems*, 29:199–207, 2016.

N Bora Keskin and Assaf Zeevi. Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research*, 42(2):277–307, 2017.

Baekjin Kim and Ambuj Tewari. Randomized exploration for non-stationary stochastic linear bandits. In *Conference on Uncertainty in Artificial Intelligence*, pp. 71–80. PMLR, 2020.

Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pp. 6437–6447. PMLR, 2020.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pp. 1739–1776. PMLR, 2018.

Anne Gael Manegueu, Alexandra Carpentier, and Yi Yu. Generalized non-stationary bandits. *arXiv preprint arXiv:2102.00725*, 2021.

Subhojyoti Mukherjee and Odalric-Ambrym Maillard. Distribution-dependent and time-uniform bounds for piecewise iid bandits. *arXiv preprint arXiv:1905.13159*, 2019.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

Yoan Russac, Olivier Cappé, and Aurélien Garivier. Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113*, 2020.

Aadirupa Saha and Shubham Gupta. Optimal and efficient dynamic regret algorithms for non-stationary dueling bandits. In *International Conference on Machine Learning*, pp. 19027–19049. PMLR, 2022.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pp. 343–354, 2008.

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

David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

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

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. *Advances in neural information processing systems*, 29, 2016.

Lai Wei and Vaibhav Srivatsva. On abruptly-changing and slowly-varying multiarmed bandit problems. In *2018 Annual American Control Conference (ACC)*, pp. 6291–6296. IEEE, 2018.

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

Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 495–504, 2018.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 746–755. PMLR, 2020.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Miscellaneous

We present in Table 2 a compilation of frequently used notations in this paper.

Notation	Description
$\mathcal{A}$	Set of all K arms.
$\mu_{a,t}$	Expected reward of arm a at time t.
μ_t^*	Expected reward of the optimal arm at time t.
$\widehat{\mu}_{a,t}$	Empirical reward of arm a at time t.
$\Delta_{a,t}$	$\mu_t^* - \mu_{a,t}$. Sub-optimality gap of arm a at time t.
$\lambda_{a,t}$	'Detectable gap' of arm a at time t (as in Equation (2)).
e_a	Number of episodes of arm a in an algorithmic run.
$t_{a,i}$	Time step after which active phase of episode i of arm a begins.
$g_{a,i}$	Time step at which statistical test of episode i of arm a passes.
$\tau_{a,i}$	$g_{a,i} - t_{a,i}$. Duration of active phase of episode i of arm a.
$w_{a,i}$	Number of rounds in the active phase of episode i of arm a,
	(or equivalently) the number of times arm a is played in its episode i.
$\omega_1^a, \omega_2^a, \ldots, \omega_w^a$	List of time steps of most recent w pulls of arm a until mentioned time step.
$\mu_{a,t}(w)$	$\frac{1}{w}\sum_{x=1}^{w} \mu_{a,\omega_x^a}$. Expected reward mean of arm a at time t in its last w pulls.
$\widehat{\mu}_{a,t}(w)$	$\frac{1}{w}\sum_{x=1}^{w} \widehat{\mu}_{a,\omega_x^a}$. Empirical reward mean of arm a at time t in its last w pulls.
$\bar{a}$	(the arm that most recently snoozed)* arm a at mentioned time.
$R_{a,i}$	Regret of playing arm a in its episode i.
$R_a(j)$	Regret of playing arm a in its episode blocks accounted to block j.
R_a	Regret of playing arm a in entire time horizon.
$w(k)$	Number of rounds in block k of the mentioned arm and episode (as in Definition 5).
$\tau(k)$	Duration of block k of the mentioned arm and episode (as in Definition 5).
$\widetilde{\tau}_{a,i}$	Duration of 'inflated' active phase of episode i of arm a.
$\widetilde{\tau}(k)$	Duration of 'inflated' block k of the mentioned arm and episode.

Table 2: Notations and their verbose descriptions.

A.1 On Computability of λ-inferiorty

The statistical test (Line 6) of our algorithm, SNR, describes the identification of $\widehat{\lambda}$-inferiority of an arm as an optimization problem over the continuous domain $\widehat{\lambda} \in [0, 1]$. However, with the integrality constraint of the sample window size w corresponding to a $\widehat{\lambda} \in [0, 1]$, we have $w \in D \subseteq$

$\{1, 2, \ldots, t\}$ belongs to a finite countable domain of size not more than t. Thus, the identification of a $\widehat{\lambda}$-inferior arm becomes a discrete optimization problem. Also, at all times, for every arm b, it is feasible to keep track of the indentity of $\bar{b}$ and the empirical means $\widehat{\mu}_{\bar{b},\cdot}(w)$ for different rounds of any window w. Thus, we note that the statistical test is deterministic and tractable.

B Proof of Theorem 1

We proceed to derive an instance-dependent regret upper bound expression for SNR. Towards that, we first establish useful properties about the nature of the active and passive phases, the sizes of episodes, the connection between the regret and the detectable gap, λ.

Let $\mathcal{G}$ denote the *good* event where for all time-steps and all arms, for all valid window sizes, the empirical reward means of the arm (or its proxy in certain parts of the window) fall close to the true reward means. In other words,

$$\mathcal{G} := \{\forall a \in \mathcal{A}, t \in [T], \forall w : \widehat{\mu}_{\bar{a},t}(w) - r(w) < \mu_{\bar{a},t}(w) < \widehat{\mu}_{\bar{a},t}(w) + r(w)\}. \tag{5}$$

Claim 1. *The good event $\mathcal{G}$ occurs with high probability; specifically, $\mathbb{P}\{\mathcal{G}\} \geq 1 - \frac{2}{T}$.*

Proof. We work in the probability space in which the rewards from arm pulls $\widehat{\mu}_{a,t}$s are generated (as a tape) ahead of time, and for every time $t \in [T]$, arm $a \in \mathcal{A}$, we have $\widehat{\mu}_{a,t}$ is an independent sample from $Ber(\mu_{a,t})$.

Recalling the definitions in Section 3, for an arm a, the notion of $\bar{a}$ refers to the arm that transitively snoozed arm a in the rounds in which a is passive, and refers to arm a itself otherwise (when a is active). Applying this to the other definitions, we have that, for any arm a, for some time-step t, for a window of w rounds, the arm $\bar{a}$ was played by SNR at time steps $\omega_1^{\bar{a}}, \omega_2^{\bar{a}}, \ldots, \omega_w^{\bar{a}} \leq t$ in the last w rounds, and $\widehat{\mu}_{\bar{a},t}(w) := \frac{1}{w}\sum_{x=1}^{w} \widehat{\mu}_{\bar{a},\omega_x^{\bar{a}}}$ is the average of empirical reward means of arm $\bar{a}$ over those samples, while $\mu_{\bar{a},t}(w) := \frac{1}{w}\sum_{x=1}^{w} \mu_{\bar{a},\omega_x^{\bar{a}}}$ is the average of true reward means of arm $\bar{a}$ in those time steps $(\omega_t^{\bar{a}})_{t=1}^{w}$.

Note that the terms $\widehat{\mu}_{\bar{a},t}(w)$ (sim. $\mu_{\bar{a},t}(w)$) and $\widehat{\mu}_{a,t}(w)$ (sim. $\mu_{a,t}(w)$) denote the same quantity if arm a is active in all of the w rounds in consideration.

Now, the probability of the average of empirical reward means $\widehat{\mu}_{\bar{a},t}(w)$ deviating from the average of true reward means $\mu_{\bar{a},t}(w)$ by more than a confidence radius $r(w) := \sqrt{\frac{2\log T}{w}}$ is upper bounded using Chernoff-Hoeffding inequality as follows:

$$\mathbb{P}\{|\widehat{\mu}_{\bar{a},t}(w) - \mu_{\bar{a},t}(w)| \geq r(w)\} \leq 2e^{-2w.r(w)^2} = 2e^{-2w.\frac{2\log T}{w}} = \frac{2}{T^4}.$$

Taking a union bound over all arms $a \in \mathcal{A}$ (given $|\mathcal{A}| < T$), all time steps $t \in [T]$, all windows of rounds w (where $w \leq T$) gives

$$\mathbb{P}\{\exists a \in \mathcal{A}, t \in [T], w, \text{ s.t. } |\widehat{\mu}_{a,t}(w) - \mu_{a,t}(w)| \geq r(w)\} \leq \frac{2}{T}$$
$$\implies \mathbb{P}\{\forall a \in \mathcal{A}, t \in [T], \forall w : |\widehat{\mu}_{a,t}(w) - \mu_{a,t}(w)| \leq r(w)\} \geq 1 - \frac{2}{T}.$$

□

For the rest of the analysis in this section until Theorem 1 is restated, we shall assume that event $\mathcal{G}$ occurs.

Lemma 2 (Sub-optimality of snoozed arms)**.** *A snoozed arm is never optimal during the passive phase. Precisely, for an arm a in its episode i, for some arm b, we have $\bar{b} >_{\widehat{\lambda}} a$ at time $g_{a,i}$. Then, for all times $t \in [g_{a,i} + 1, t_{a,i+1}]$, we have $\mu_{a,t} \leq \mu_t^*$.*

Proof. Note that if there is no passive phase in the episode, i.e., when arm a is not snoozed, the result is vacuously true. Thus, what is to be shown is only the case where arm a is snoozed for a non-empty passive phase $[g_{a,i}+1, t_{a,i+1}]$. We prove this lemma in two steps. First, we show (in Claim 3) that when SNR detects $\bar{b} >_{\widehat{\lambda}} a$ at time $g_{a,i}$, then, for a period of time culminating at $g_{a,i}$, some arm α has a larger average true reward mean than that of arm a by a certain margin. Second, we argue that the duration of the passive phase for which arm a is snoozed is chosen based on this margin such that it remains sub-optimal for the entirety of the passive phase owing to the drift limit.

At the end of the active phase of arm a's episode i, at $g_{a,i}$, SNR detects for some $\widehat{\lambda} \in [0,1]$ and some arm b that $\bar{b} >_{\widehat{\lambda}} a$. We make the following claim about the sub-optimality gap of arm a from another arm at some point in time during this active phase.

Claim 3. *There exists a time-step $t' \in [t_{a,i} + {}^{K}/_{2}, g_{a,i}]$ such that for some arm α, we have*

$$\mu_{\alpha,t'} - \mu_{a,t'} \geq \frac{\widehat{\lambda}}{3} - K\delta \geq 4\sqrt{\frac{\log T}{w_{a,i}}} - K\delta.$$

Proof. As arm $\bar{b} >_{\widehat{\lambda}} a$ at time $t = g_{a,i}$, by Definition 3, we have, for a window of w rounds

$$\begin{aligned}
& LCB_{\bar{b},t}(w) > UCB_{a,t}(w) + 2r(w) - K\delta \\
\implies & \widehat{\mu}_{\bar{b},t}(w) - r(w) > \widehat{\mu}_{a,t}(w) + r(w) + 2r(w) - K\delta \\
\overset{(a)}{\implies} & \mu_{\bar{b},t}(w) > \widehat{\mu}_{\bar{b},t}(w) - r(w) > \widehat{\mu}_{a,t}(w) + r(w) + 2r(w) - K\delta > \mu_{a,t}(w) + 2r(w) - K\delta \\
\implies & \mu_{\bar{b},t}(w) - \mu_{a,t}(w) > 2r(w) - K\delta \\
\implies & \mu_{\bar{b},t}(w) - \mu_{a,t}(w) > 2\sqrt{\frac{2\log T}{w}} - K\delta = 2\sqrt{\frac{2\log T.\widehat{\lambda}^2}{72\log T}} - K\delta = 2\sqrt{\frac{\widehat{\lambda}^2}{36}} - K\delta = \frac{\widehat{\lambda}}{3} - K\delta.
\end{aligned} \tag{6}$$

The implication (a) is due to the occurence of event $\mathcal{G}$. Note that $\widehat{\lambda} \in [0,1]$ is constrained (again, as in Definition 3) by the number of samples w available in the current episode i as follows:

$$\left\lceil \frac{c_0 \log T}{\widehat{\lambda}^2} \right\rceil \leq w_{a,i} \implies \widehat{\lambda} \geq \sqrt{\frac{c_0 \log T}{w_{a,i}}}. \tag{7}$$

We use the short-hand $f := \frac{1}{3}\sqrt{\frac{c_0 \log T}{w_{a,i}}} = 4\sqrt{\frac{\log T}{w_{a,i}}}$ for ease of expression. Expanding the terms $\mu_{a,t}(w)$ and $\mu_{\bar{b},t}(w)$ by their definitions, we have

$$\mu_{\bar{b},t}(w) - \mu_{a,t}(w) = \frac{1}{w}\sum_{i=1}^{w} \mu_{\bar{b},\omega_i^{\bar{b}}} - \mu_{a,\omega_i^a} \overset{(a)}{>} \frac{\widehat{\lambda}}{3} - K\delta \overset{(b)}{\geq} f - K\delta. \tag{8}$$

The (a) is due to Equation (6) and the (b) is due to $f \leq \frac{\widehat{\lambda}}{3}$ from Equation (7).

The average of w terms being larger than $f - \delta$ implies that atleast one term is larger than $f - \delta$. Thus, for some $1 \leq i \leq w$, we have $\mu_{\bar{b},\omega_i^{\bar{b}}} - \mu_{a,\omega_i^a} > f - \delta$. The two time steps $\omega_i^{\bar{b}}$ and ω_i^a from the same round i can not be more than $K-1$ time steps apart by the nature of round-robin play of arms. With the drift limit implication as in Equation (1), we have $\mu_{\bar{b},\omega_i^a} - \mu_{a,\omega_i^a} > f - K\delta$. Thus, we get for some $t' \in [t_{a,i} + {}^{K}/_{2}, g_{a,i}]$ that $\mu_{\bar{b},t'} - \mu_{a,t'} \geq f - K\delta$. This completes the proof of the claim with the arm $\alpha = \bar{b}$ at time t'. □

We continue to use the short-hand $f := 4\sqrt{\frac{\log T}{w_{a,i}}}$ for ease of expression. From Claim 3, at some time $t' \in [t_{a,i} + {}^{K}/_{2}, g_{a,i}]$, we have for some arm b that $\mu_{b,t'} - \mu_{a,t'} > f - K\delta$. By adhering to the drift limit, we have $|\Delta_{a,t} - \Delta_{a,t+1}| \leq 2\delta$ for all time-steps $t \in [T-1]$. Thus, for the identity of the optimal arm to change, i.e., for the sub-optimality gap of at least $f - K\delta$ of arm a to get exhausted/overturned, it requires a minimum of ${}^{(f-K\delta)}/_{2\delta} = {}^{f}/_{2\delta} - {}^{K}/_{2}$ time-steps to pass. Thus, for all $t'' \in [t', t' + {}^{f}/_{2\delta} - {}^{K}/_{2}]$, and thus for all $t'' \in [g_{a,i} + 1, t_{a,i} + {}^{f}/_{2\delta}]$ (which is a shorter sub-period) we have $\mu_{a,t''} \leq \mu_{b,t''} \leq \mu^*_{t''}$. i.e., arm a is sub-optimal in the time period $[g_{a,i} + 1, t_{a,i} + {}^{f}/_{2\delta}]$.

Once SNR detects $\bar{b} >_{\hat{\lambda}} a$ at time t, it computes the duration of the sub-optimality buffer period as $\text{buf} = {}^{f}/_{2\delta}$ (in Line 9). If $t_{a,i} + \text{buf} \leq g_{a,i}$ or equivaently $\text{buf} \leq \tau_{a,i}$, the arm a is not snoozed for a passive phase. Otherwise, as in Line 11 of the agorithm, the sub-optimal arm a is snoozed until time-step $t_{a,i+1} = t_{a,i} + \text{buf} = t_{a,i} + {}^{f}/_{2\delta}$, i.e., the passive phase runs for the period $[g_{a,i+1}, t_{a,i} + {}^{f}/_{2\delta}]$. We have already shown that during this period arm a is sub-optimal. □

Thus, thanks to Lemma 2, during the passive phase of an episode, not playing the snoozed arm does not lead to any regret.

We now bound the expected regret incurred by the sub-optimal arm a in episode i, Towards that, we define the notion of regret in an episode:

Definition 4 (Episode Regret)**.** *In an algorithmic run of* SNR *with a drift limit parameter δ on a problem instance $\mu \in S_\delta$, the regret for arm a in its episode i defined as*

$$R_{a,i} = \sum_{t=t_{a,i}+1}^{t_{a,i+1}} \left(\mu^*_t - \mu_{\text{SNR}(t),t}\right) \times \mathbb{1}\left\{\text{SNR}(t) = a\right\},$$

where $\mu^*_t = \max_{a\in\mathcal{A}} \mu_{a,t}$ is the mean reward of the optimal arm at time t, $\text{SNR}(t) \in \mathcal{A}$ *is the arm pulled by* SNR *at time $t \in [T]$.*

Here, $\mathbb{1}\{X\}$ is the indicator random variable that takes the value 1 if event X happens, and 0 otherwise.

Next, we upper bound the regret $R_{a,i}$ as a function of $w_{a,i}$, the number of times arm a is pulled in its episode i. Towards that, we state a key definition about partitioning the active phase into blocks of time as follows:

Definition 5 (Blocks in an active phase of episode)**.** *For an episode i of arm a, divide the time period $[t_{a,i} + 1, g_{a,i} - 1]$ of size $\tau_{a,i} - 1$ (consisting of $w_{a,i} - 1$ rounds) into $j \geq 1$ consecutive blocks of time periods, where every block $k \in [j]$ spans the time period $[\alpha_{k-1} + 1, \alpha_k]$ of size $\tau(k)$ which contains a window of $w(k)$ rounds of samples.*

The size of every block $k \in [j-1]$ is $\tau(k)$ and is chosen such that $\tau(k) = \frac{1}{\delta}\sqrt{\frac{\log T}{w(k)}}$ and for the last block, where $k = j$, such that $\tau(k) \leq \frac{1}{\delta}\sqrt{\frac{\log T}{w(k)}}$. Note that the active phase's first time step $\alpha_0 = t_{a,i}$ and last time step $\alpha_j = g_{a,i} - 1$ here.

Lemma 4 (Episode regret trajectory-dependent upper bound)**.** *In an algorithmic run of* SNR *with a drift limit parameter δ on a problem instance $\mu \in S_\delta$, the regret for arm a in its episode i is upper bounded as*

$$R_{a,i} \leq 1 + \sum_{k=1}^{j} c_2 \sqrt{w(k).\log T}.$$

where $c_2 = 10$ is a constant, and terms j and $w(k)$s are as in Definition 5.

Proof. From Definition 4,

$$
\begin{aligned}
R_{a,i} &= \sum_{t=t_{a,i}+1}^{t_{a,i+1}} \left(\mu_t^* - \mu_{\text{SNR}(t),t}\right) \times \mathbb{1}\left\{\text{SNR}(t) = a\right\} \\
&\overset{\text{(a)}}{=} \sum_{t=t_{a,i}+1}^{g_{a,i}} \left(\mu_t^* - \mu_{\text{SNR}(t),t}\right) \times \mathbb{1}\left\{\text{SNR}(t) = a\right\} \\
&\overset{\text{(b)}}{\leq} 1 + \sum_{t=t_{a,i}+1}^{g_{a,i}-1} \left(\mu_t^* - \mu_{\text{SNR}(t),t}\right) \times \mathbb{1}\left\{\text{SNR}(t) = a\right\}. \qquad (9)
\end{aligned}
$$

The (a) is due to the snoozed arm not being played in the entire passive phase $[g_{a,i}+1, t_{a,i+1}]$. The (b) is obtained by trivially upper bounding the regret at time $g_{a,i}$ by 1.

Recall that SNR rotates the active arms (i.e., play in a round-robin fashion) in the time period $[t_{a,i}+1, g_{a,i}-1]$, and the statistical test did not pass in that period. Thus, we have that there was no $\widehat{\lambda} \in [0,1]$ (and a corresponding window of $w := \left\lceil \frac{c_1 \log T}{\widehat{\lambda}^2} \right\rceil$ rounds) for which $\bar{b} >_{\widehat{\lambda}} a$ at any time $t \in [t_{a,i}+1, g_{a,i}-1]$.

Applying Definition 3 at time this time t with some valid window of w rounds, we have for all arms $b \in \mathcal{A} \setminus \{a\}$ that

$$
\begin{aligned}
LCB_{\bar{b},t}(w) \leq& UCB_{a,t}(w) + 2r(w) - K\delta \\
\implies \widehat{\mu}_{\bar{b},t}(w) - r(w) \leq& \widehat{\mu}_{a,t}(w) + r(w) + 2r(w) - K\delta \\
\overset{\text{(a)}}{\implies} \mu_{\bar{b},t}(w) - r(w) - r(w) \leq& \widehat{\mu}_{b,t}(w) - r(w) \\
\leq \widehat{\mu}_{a,t}(w) + r(w) + 2r(w) - K\delta \leq& \mu_{a,t}(w) + r(w) + r(w) + 2r(w) - K\delta \\
\implies \mu_{\bar{b},t}(w) - \mu_{a,t}(w) \leq& 6r(w) - K\delta \\
\iff \mu_{\bar{b},t}(w) - \mu_{a,t}(w) \leq& 6\sqrt{\frac{2\log T}{w}} - K\delta \\
\overset{\text{(b)}}{\implies} \mu_{b,t}(w) - \mu_{a,t}(w) \leq& 6\sqrt{\frac{2\log T}{w}} - K\delta = c_1\sqrt{\frac{\log T}{w}} - K\delta, \qquad (10)
\end{aligned}
$$

where $c_1 = \sqrt{72}$ is a constant. The (a) is due to occurrence of the good event $\mathcal{G}$. The (b) is due to $\mu_{\bar{b},t}(w) \geq \mu_{b,t}(w)$ as per Lemma 2. Note that Equation (10) is a generic inequality to upper bound the gap of an arm a to another single arm b whilst the statistical test keeps failing.

Continuing the analysis from Equation (10) that upper bounds the reward mean gap of arm a from another single arm b, we next try to upper bound the reward mean gap of arm a from that of the optimal arm at each time-step. For this, we constrain the number of rounds, w, of samples available from the current episode for the statistical test, and the time period τ that spans the w rounds as follows:

Claim 5. *At time t, if the statistical test doesn't detect that arm a is λ-inferior with a window of w rounds of samples that spans a time period of size τ from the current episode, such that $\tau \leq \frac{1}{\delta}\sqrt{\frac{\log T}{w}}$, then the sub-optimality of arm a is upper bounded as follows: for all $t' \in [t-\tau+1, t]$, we have*

$$
\mu_{t'}^* - \mu_{a,t'} \leq c_2\sqrt{\frac{\log T}{w}} - K\delta.
$$

Proof. From Equation (10), for all arms $b \in \mathcal{A} \setminus \{a\}$, we have $\mu_{b,t}(w) - \mu_{a,t}(w) \leq c_1\sqrt{\frac{\log T}{w}} - K\delta$. Expanding the terms, we have $\frac{1}{w}\sum_{x=1}^{w} \mu_{b,\omega_x^b} - \mu_{a,\omega_x^a} \leq c_1\sqrt{\frac{\log T}{w}} - K\delta$. Thus, we have that for

some round $\ell \in [w]$,

$$\mu_{b,\omega_\ell^b} - \mu_{a,\omega_\ell^a} \leq c_1 \sqrt{\frac{\log T}{w}} - K\delta. \tag{11}$$

By the drift limit implication, for any time steps $t', t'' \in [t - \tau + 1, t]$, we have for all arms b that $\mu_{b,t'} - \mu_{b,t''} \leq \tau.\delta \leq \sqrt{\frac{\log T}{w}}$. Combining this with Equation (11), for all time steps $t' \in [t-\tau+1, t]$ and for all arms $b \in \mathcal{A} \setminus \{a\}$, we have that $\mu_{b,t'} - \mu_{a,t'} \leq \left(c_1\sqrt{\frac{\log T}{w}} - K\delta\right) + \left(\sqrt{\frac{\log T}{w}}\right) = c_2\sqrt{\frac{\log T}{w}} - K\delta$, where $c_2 = 10 \geq c_1 + 1$ is a constant. Consequentially, as it holds for all other arms b, $\max_{b\in\mathcal{A}\setminus\{a\}} \mu_{b,t'} - \mu_{a,t'} = \mu^*_{t'} - \mu_{a,t'} \leq c_2\sqrt{\frac{\log T}{w}} - K\delta$. This completes the proof of the claim. □

Using the results obtained in Claim 5 that constraints the window w of samples and time period τ that spans the window, we proceed with bounding the episodic regret $R_{a,i}$ by similarly breaking down the active phase $\tau_{a,i}$ and total rounds $w_{a,i}$ of samples observed into suitably sized blocks of time periods as described in Definition 5.

Next, bringing back Equation (9), we have

$$\begin{aligned}
R_{a,i} \leq & 1 + \sum_{t=t_{a,i}+1}^{g_{a,i}-1} (\mu^*_t - \mu_{\text{SNR}(t),t}) \times \mathbb{1}\{\text{SNR}(t) = a\} \\
\leq & 1 + \sum_{k=1}^{j} \sum_{t=\alpha_{k-1}+1}^{\alpha_k} (\mu^*_t - \mu_{\text{SNR}(t),t}) \times \mathbb{1}\{\text{SNR}(t) = a\} \\
\overset{\text{(a)}}{\leq} & 1 + \sum_{k=1}^{j} \left(c_2\sqrt{\frac{\log T}{w(k)}} - K\delta\right).w(k) \\
\leq & 1 + \sum_{k=1}^{j} c_2\sqrt{w(k)\log T}.
\end{aligned} \tag{12}$$

Here, (a) is due to Claim 5. This completes the proof of the Lemma 4. □

We see that Lemma 4 characterizes the episodic regret as a function of algorithm-run (trajectory) dependent quantities—such as the window of samples, $w_{a,i}$s, with a constrained break-up of the active phase of size $\tau_{a,i}$. Next, we strive to get an instance dependent characterization of the episodic regret in terms of the detectable gap λ by drawing connections to the samples observed, $w_{a,i}$ (or $w(k)$s to be precise).

Lemma 6 (Episode regret instance-dependent upper bound)**.** *In an algorithmic run of* SNR *with a drift limit parameter δ on a problem instance $\mu \in S_\delta$, the regret in episode $i \in [e]$ of an arm a is upper bounded as*

$$R_{a,i} \leq c_4 \sum_{k=1}^{j} \frac{1}{\lambda_{\alpha_k}} \log T + c_5,$$

where $c_4 = 120, c_5 = 1$ are constants, and terms j and α_ks are as in Definition 5.

Proof. Towards proving this lemma, we make an important observation about the connection between the detectable gaps λ and the success of the statistical test used in our algorithm.

Lemma 7. *[Sufficient condition for positive statistical test] At some time t in the active phase of some episode of arm a when w rounds have passed, if we have $\lambda_{a,t} > \sqrt{\frac{c_0 \log T}{w}}$, then the statistical test passes, i.e., algorithm* SNR *at time t declares $\bar{b} >_{\widehat{\lambda}} a$ for some arm b. Here, $c_0 = 144$ is a constant.*

Proof. If $\lambda_{a,t} = \sqrt{\frac{c_0 \log T}{t}}$ (from second assignment in Eqn. 2), we always have $\lambda_{a,t} \leq \sqrt{\frac{c_0 \log T}{w}}$ as $w \leq t$. But, given that we have its negation as the premise of the Lemma statement, we assume the description of detectable gap λ as quantified by the first assignment in Equation (2). While this description guarantees over an average of reward mean gaps computed over a window of contiguous time steps, to make a comparison with the algorithm's statistical test, we desire similar guarantees about the average of reward mean gaps computed over a sparser set of time-steps from the window, specifically, from one time step per round over the window. We make the following claim towards that:

Claim 8. *The averages of the true reward means of arm a and some arm b at time t computed over a window of $w \geq \frac{c_0 \log T}{\lambda_{a,t}^2}$ rounds spanning a time period of at most $\frac{c_3 K \log T}{\lambda_{a,t}^2}$ obeys $\mu_{b,t}(w) - \mu_{a,t}(w) \geq \lambda_{a,t} - K\delta$. Here, $c_3 = 72$ is a constant.*

Proof. Let b be the arm that was compared with arm a in determining the value of $\lambda_{a,t}$. From Equation (2), we have for a contiguous window of size $\tau' := \frac{c_0 K \log T}{\lambda_{a,t}^2}$ that spans $[s = t - \tau' + 1, t]$, we have

$$\begin{aligned} &\frac{1}{\tau'} \sum_{t'=s}^{t} \mu_{b,t'} - \mu_{a,t'} \geq \lambda_{a,t} \\ \overset{\text{(a)}}{\implies} &\frac{1}{w} \sum_{x=1}^{w} \mu_{b,\omega_x^b} - \mu_{a,\omega_x^a} \geq \lambda_{a,t} - K\delta. \end{aligned}$$

The (a) is due to the δ drift limit implication and the fact that the ω_x^b and ω_a^b time steps are at most K steps apart between themselves and between two consecutive rounds (i.e., between ω_x^b and ω_{x+1}^b), and the fact that all w time steps are from within the time period of $[s = t - \tau' + 1, t]$ for which $\lambda_{a,t}$ was originally computed. □

From Claim 8, we have

$$\begin{aligned} &\mu_{b,t}(w) - \mu_{a,t}(w) \geq \lambda_{a,t} - K\delta \\ \overset{\text{(a)}}{\implies} &\widehat{\mu}_{b,t}(w) - \widehat{\mu}_{a,t}(w) \geq \lambda_{a,t} - 2r(w) - K\delta \\ \overset{\text{(b)}}{\implies} &\widehat{\mu}_{\bar{b},t}(w) - \widehat{\mu}_{a,t}(w) \geq \lambda_{a,t} - 2r(w) - K\delta \\ \implies &LCB_{\bar{b},t}(w) - UCB_{a,t}(w) \geq \lambda_{a,t} - 4r(w) - K\delta \\ &= \sqrt{\frac{c_3 K \log T}{\tau'}} - 4\sqrt{\frac{2 \log T}{w}} - K\delta \\ &\overset{\text{(c)}}{\geq} \sqrt{\frac{c_3 \log T}{w}} - 4\sqrt{\frac{2 \log T}{w}} - K\delta \\ &= 2r(w) - K\delta, \end{aligned} \tag{13}$$

where $c_3 = 4 \geq \sqrt{c_0} - 4\sqrt{2}$ is a constant. Here, (a) is due to occurence of $\mathcal{G}$, the (b) is due to $\widehat{\mu}_{\bar{b},t}(w) \geq \widehat{\mu}_{b,t}(w)$ from Lemma 2, and (c) uses $w \geq {}^{\tau'}/_{K}$.

The inequality in Equation (13) is the condition for the algorithm to declare $\bar{b} >_{\widehat{\lambda}} a$ (as in Definition 3) at time t. Thus, the statistical test passes. This completes the proof of Lemma 7. □

Now, recall that the statistical test did not pass in the time period $[t_{a,i}+1, g_{a,i}-1]$. Specifically, the test did not pass in the time-steps $\alpha_1, \alpha_2, \ldots, \alpha_j = g_{a,i}-1$ associated with the active phase of arm a's episode i. By Lemma 7, for all $k \in [j]$, we have $\lambda_{\alpha_k} \leq \sqrt{\frac{c_0 \log T}{w(k)}}$ or equivalently, $\sqrt{w(k)} \leq \frac{\sqrt{c_0 \log T}}{\lambda_{\alpha_k}}$. Substituting this in Equation (12), we get

$$
\begin{aligned}
R_{a,i} &\leq 1 + \sum_{k=1}^{j} c_2 \sqrt{w(k).\log T} \\
&\leq 1 + \sum_{k=1}^{j} c_2 \frac{\sqrt{c_0 \log T}}{\lambda_{\alpha_k}} \sqrt{\log T} \\
&= \sum_{k=1}^{j} c_4 \frac{1}{\lambda_{\alpha_k}} \log T + 1 \\
&= c_4 \sum_{k=1}^{j} \frac{1}{\lambda_{\alpha_k}} \log T + c_5.
\end{aligned}
$$

This completes the proof of Lemma 6. □

Next, we upper bound the overall regret of an arm a over all episodes. Towards that, we argue that the time steps α_ks that are a part of the episodic regret upper bound in Lemma 6 are well spaced between themselves (as in Claim 9) and between different episodes (from Claim 10). We next present two claims in that regard.

Claim 9. *In the regret upper bound expressions in Lemma 6, for every episode i of arm a, the time steps $\alpha_1, \alpha_2, \ldots, \alpha_j$ that characterize the duration of the contiguous time blocks and the regret expression are spaced as follows: for all $k \in [j-1]$, we have $\alpha_k - \alpha_{k-1} \geq \delta^{-2/3} \log^{1/3} T$.*

Proof. By definition (as in Lemma 4), we have the time period $[t_{a,i+1}, g_{a,i}-1]$ divided into j consecutive blocks of time periods, where every block $k \in [j]$ spans the time period $[\alpha_{k-1}+1, \alpha_k]$ of size $\tau(k)$ which contains a window of $w(k)$ rounds of samples. The sizes of every block $k \in [j-1]$ are chosen such that $\tau(k) = \frac{1}{\delta}\sqrt{\frac{\log T}{w(k)}}$. Trivially, $\tau(k) \geq w(k)$. Thus, we have $\tau(k) \geq \delta^{-2/3} \log^{1/3} T$, completing the proof of Claim 9. □

Claim 10 (Minimum duration of an episode)**.** *In an algorithmic run of* SNR*, for any arm a, the duration of any episode $i \in [e_a - 1]$ (except the last one) is lower bounded as follows: $t_{i+1} - t_i \geq c_9 \delta^{-2/3} \log^{1/3} T$. Here, $c_9 = 2^{2/3}$ is a constant.*

Proof. At the end of the active phase (of duration $\tau_{a,i}$) at time-step $g_{a,i}$, the sub-optimality buffer computed is buf $= \frac{2}{\delta}\sqrt{\frac{\log T}{w_{a,i}}}$. In Lines 10-11, the algorithm decides what the snooze duration of arm a should be, thus determining $t_{a,i+1}$, the time when the episode i of arm a shall end, based on the following two cases.

Case $\tau_{a,i} \geq$ buf : Equivalently, we have $\tau_{a,i} \geq \frac{2}{\delta}\sqrt{\frac{\log T}{w_{a,i}}}$. The sub-optimal arm does not get snoozed, so, episode i terminates at $t_{a,i+1} = g_{a,i} = t_{a,i} + \tau_{a,i}$.

Case $\tau_{a,i} <$ buf : Equivalently, we have $\tau_{a,i} < \frac{2}{\delta}\sqrt{\frac{\log T}{w_{a,i}}}$. The sub-optimal arm gets snoozed until time $t_{a,i} +$ buf $= t_{a,i} + \frac{2}{\delta}\sqrt{\frac{\log T}{w_{a,i}}}$. The end of the passive phase marks the end of episode i, thus, $t_{a,i+1} = t_{a,i} + \frac{2}{\delta}\sqrt{\frac{\log T}{w_{a,i}}}$.

From these two cases, we can see that $t_{a,i+1} = t_{a,i} + \max\left\{\tau_{a,i}, \frac{2}{\delta}\sqrt{\frac{\log T}{w_{a,i}}}\right\}$. As $\tau_{a,i} \geq w_{a,i}$, to lower bound $t_{a,i+1} - t_{a,i}$, we minimize the quantity $\max\left\{w_{a,i}, \frac{2}{\delta}\sqrt{\frac{\log T}{w_{a,i}}}\right\}$. As the two quantities grow oppositely with $w_{a,i}$, the minimum occurs when

$$w_{a,i} = \frac{2}{\delta}\sqrt{\frac{\log T}{w_{a,i}}} \quad \Leftrightarrow \quad w_{a,i} = 2^{2/3}\delta^{-2/3}\log^{1/3} T.$$

Thus, we have $t_{a,i+1} - t_{a,i} \geq c_9\delta^{-2/3}\log^{1/3} T$, completing the proof of Claim 10. □

Lemma 11 (Single arm instance-dependent regret upper bound)**.** *In an algorithmic run of* SNR *with a drift limit parameter δ on a problem instance $\mu \in S_\delta$, the regret of an arm a is upper bounded as*

$$R_a = \sum_{i\in[e_a]} R_{a,i} \leq c_6 \sum_{j=1}^{m} \frac{1}{\lambda_{a,min}(j)}.\log T + c_7,$$

where $m = {}^{T}/_{\tau}$ is the number of blocks, each of length not more than $\tau = \min\left\{T, c_3\delta^{-2/3}K^{1/3}\log^{1/3} T\right\}$, and for every block $j \in [m]$ spanning a time period $b_j := [(j-1)\tau + 1, j\tau] \cap [T]$, define $\lambda_{a,min}(j) := \min_{t\in b_j} \lambda_{a,t}$. Here, c_is are suitable constants.

Proof. To show this upper bound, we partition the time-horizon into *blocks* of suitably small size, and bound the regret in episodes (from Lemma 6) by accounting each term in them to some block.

Recall that in an algorithmic run of SNR, e_a is the number of episodes that arm a runs for. These episodes, indexed $1, 2, \ldots, e_a$, start after times $t_{a,1} = 0, t_{a,2}, \ldots, t_{a,e_a}$ (and a hypothetical $t_{e_a+1} = T$ marks the end of the last episode e) respectively. Note that the collection of episode time periods $([t_{a,i} + 1, t_{a,i+1}])_{i\in[e_a]}$ for each arm a partitions the entire time horizon $[T]$. Thus, the total regret incurred by the algorithm (in a particular run [8]) is $R(\text{SNR}) = \sum_{a\in\mathcal{A}}\sum_{i\in[e_a]} R_{a,i}$.

For analysing the total regret, we partition the time horizon into $m = {}^{T}/_{\tau}$ blocks (indexed as $1, 2, \ldots, m$), each of size $\tau = c_9\delta^{-2/3}\log^{1/3} T$, where $c_9 = 2^{2/3}$ is a constant. Each block $j \in [m]$ spans the time period $b_j := [(j-1)\tau + 1, j\tau] \cap [T]$.

As stated earlier, the chosen size of a block τ is the minimum duration of an episode stated in Claim 10. Thus, a block period b_j overlaps with a maximum of two episodes.

Note that the chosen block size τ is also the minimum size of a block (except the last one) as per Claim 9 in the episodic regret upper bound expression in Lemma 6. Thus, a block period b_j overlaps with at most two episodic regret blocks from a particular episode.

Thus, for any block $j \in [m]$, there are at most 2 episodes, say $i, i+1$, and at most 4 corresponding episodic regret blocks end times, say α_k, α_{k+1} for episode i and $\alpha_{k'}, \alpha_{k'+1}$ for episode $i+1$, intersect with the block period b_j. We account the corresponding four block regrets to block j. We denote this regret to block $j \in [m]$ as

$$R_a(j) \leq c_4.\left(\frac{1}{\lambda_{a,\alpha_k}} + \frac{1}{\lambda_{a,\alpha_{k+1}}} + \frac{1}{\lambda_{a,\alpha'_k}} + \frac{1}{\lambda_{a,\alpha_{k'+1}}}\right).\log T + 4c_5 \tag{14}$$

As the α_ks in Equation (14) are algorithm trajectory dependent quantities, and every block $j \in [m]$ accounts for at most a constant number of regret terms, we further upper bound $R_a(j)$ as follows:

[8] while the quantities e_as, $\tau_{a,i}$s, $w_{a,i}$ etc. are random variables, we use them here as quantities obtained after a single algorithmic trajectory/run.

$$
\begin{aligned}
R_a(j) \leq & 4c_4 . \max_{t \in b_j} \frac{1}{\lambda_{a,t}} . \log T + 4c_5 \\
& c_6 . \frac{1}{\min_{t \in b_j} \lambda_{a,t}} . \log T + c_7 \\
& c_6 . \frac{1}{\lambda_{a,min}(j)} . \log T + c_7,
\end{aligned} \tag{15}
$$

where we write $\lambda_{a,min}(j) := \min_{t \in b_j} \lambda_{a,t}$, and $c_6 = 4c_4 = 480$, $c_7 = 4c_5 = 4$ are constants.

Note that every regret term of every episode from Lemma 6 is accounted to some block $j \in [m]$. By this accounting criteria, we modify the expression for regret of arm a from a summation over episodes to a summation over blocks as follows:

$$
R_a = \sum_{i \in e_a} R_{a,i} \leq \sum_{j \in [m]} R_a(j) \overset{\text{(a)}}{\leq} \sum_{j \in [m]} c_6 . \frac{1}{\lambda_{a,min}(j)} . \log T + c_7,
$$

where (a) is from Equation (15).

This completes the proof of the Lemma 11. □

Next, we come to the final part of the proof. We show the instance-dependent regret bound of SnR over the entire time horizon as an accumulation of regret bounds of all arms:

$$
R(\text{SnR}) = \sum_{a \in \mathcal{A}} R_a \overset{\text{(a)}}{\leq} \sum_{a \in \mathcal{A}} \sum_{j=1}^{m} c_6 \frac{1}{\lambda_{a,min}(j)} + c_7, \tag{16}
$$

where (a) is from Lemma 11.

For the remainder of the proof, we drop the implicit assumption that $\mathcal{G}$ occurs, and upper bound the conditional expected regret of our algorithm SnR when high probable event $\mathcal{G}$ occurs using Equation (16), and then generously upper bound the conditional regret when $\mathcal{G}'$ occurs by T (a regret of 1 for every time-step). Along with Claim 1, we have

$$
\begin{aligned}
& \mathbb{E}\left[R(\text{SnR})\right] \\
= & \mathbb{E}\left[R(\text{SnR}|\mathcal{G})\right] . \mathbb{P}\left\{\mathcal{G}\right\} + \mathbb{E}\left[R(\text{SnR}|\mathcal{G}')\right] . \mathbb{P}\left\{\mathcal{G}'\right\} \\
\leq & \left(\sum_{a \in \mathcal{A}} \sum_{j=1}^{m} c_6 . \frac{1}{\lambda_{a,min}(j)} + c_7 \right) + T . \frac{2}{T} \\
= & c_6 \sum_{a \in \mathcal{A}} \sum_{j=1}^{m} \frac{1}{\lambda_{a,min}(j)} . \log T + c_8,
\end{aligned} \tag{17}
$$

where $c_6 = 480$ and $c_8 = c_7 + 2 = 6$ are constants. This completes the proof and leads to the Theorem.

Theorem 1. *[Instance-dependent regret bound] If* SnR *is run with a drift limit parameter δ on a problem instance $\mu \in S_\delta$, then its expected regret , $R(\text{SnR})$, is upper bounded by*

$$
c_6 \sum_{a \in \mathcal{A}} \sum_{j=1}^{m} \frac{1}{\lambda_{a,min}(j)} . \log T + c_8,
$$

where $m = T/\tau$ is the number of blocks, each of length not more than $\tau = \min\{T, c_9 \delta^{-2/3} \log^{1/3} T\}$, and for every block $j \in [m]$ spanning a time period $b_j := [(j-1)\tau + 1, j\tau] \cap [T]$, define $\lambda_{a,min}(j) := \min_{t \in b_j} \lambda_{a,t}$. Here, c_is are suitable constants.

C Proof of Theorem 2

In this section, we show a minimax (instance-independent) upper bound for the regret incurred by SNR that depends on the time horizon T, the drift limit δ, and the number of arms K but, is independent of the actual reward mean profile μ. Towards this, for every episode, we upper bound the *average regret* or per-time-step regret.

Then, we shall show a minimax regret that is bounded by T times the maximum average regret incurred in any episode.

Theorem 2. *[Minimax Upper bound] If* SNR *is run with a drift limit parameter δ on any problem instance $\mu \in S_\delta$, then it incurs an expected regret of $O(T\delta^{1/3}K^{1/3}\log^{1/3} T)$, where K is the number of arms.*

Proof of Theorem. As shown in Equation (17), the conditional regret when $\mathcal{G}'$ occurs can be bounded as $\mathbb{E}\left[R(\text{SNR}|\mathcal{G}')\right].\mathbb{P}\left\{\mathcal{G}'\right\} \leq 2$. Thus, it is sufficient to prove the required bound under the assumption of occurrence of $\mathcal{G}$.

We begin our proof from Equation (12) that was used as an intermediary part of the proof of the earlier Theorem 1. The regret of an arm a in its episode i as per Equation (12) is

$$\begin{aligned} R_{a,i} \leq & 1 + \sum_{k=1}^{j} c_2 \sqrt{w(k).\log T} \\ \leq & \sum_{k=1}^{j} c_{10} \sqrt{w(k).\log T}, \end{aligned} \tag{18}$$

where j is the number of blocks partitioning the active phase and $w(k)$ is the number of rounds (or the number of times arm a is pulled) in block $k \in [j]$, as per Definition 5.

The total regret for our algorithm over the entire time horizon over all arms is

$$\begin{aligned} R(\text{SNR}) &= \sum_{a\in\mathcal{A}} \sum_{i\in[e_a]} R_{a,i} \\ &\overset{\text{(a)}}{\leq} \sum_{a\in\mathcal{A}} \sum_{i\in[e_a]} \sum_{k=1}^{j} c_{10} \sqrt{w(k).\log T} \\ &\leq KT \max_{a\in\mathcal{A}, i\in[e_a]} \frac{\sum_{k=1}^{j} c_{10}\sqrt{w(k).\log T}}{t_{a,i+1} - t_{a,i}}, \end{aligned} \tag{19}$$

where (a) uses Equation (18).

The final inequality is due to the fact that each time-step is a part of only one episode of every arm. Note that $\sum_{k=1}^{j} c_{10}\sqrt{w(k).\log T}/t_{a,i+1}-t_{a,i}$ is the averaged regret per-time-step of episode i of arm a. The episode of period $[t_{a,i}+1, t_{a,i+1}]$ can comprise of both active and passive phases, and we have $\tau_{a,i}/K \leq w_{a,i}$ as not all K arms are necessarily active (and thus played) in every round.

Inflated episodes. For convenience in handling these quantities, we introduce an analytical construct that we call *inflated episodes*. For an episode i of an arm a, we obtain an inflated episode (with inflated active and passive phase) by interleaving into it time steps from episodes of other arms. Precisely, we do the following. Consider some round in the algorithmic run/trajectory of size $x \leq K$, i.e., x distinct arms are pulled once each during that round. We have x arms in their active phases and $K - x$ arms in their passive phases for that round. There are x time steps in each of the $K - x$ passive arms in the current round, thus a total of $x(K - x)$ passive time steps. We inflate the current round of the x active arms by assigning $K - x$ time steps each, thus each of the $x(K - x)$ passive time steps are added into the round of some active arm. This inflated round is a part of the

inflated active phase (and inflated episode) of the active arm. Now, the round of each of the x active arms have K timesteps each, and the round of each of the $K - x$ passive arms have no time steps.

By repeating this exercise for all rounds of an algorithmic trajectory, we obtain the complete set of inflated episodes. Observe that every time step removed from the passive phase of some arm gets into the inflated active phase of some other arm. Thus, we have

$$\sum_{a\in\mathcal{A}}\sum_{i\in[e_a]} t_{a,i+1} - t_{a,i} = \sum_{a\in\mathcal{A}}\sum_{i\in[e_a]} \tau_{a,i} + (t_{a,i+1} - g_{a,i}) = \sum_{a\in\mathcal{A}}\sum_{i\in[e_a]} \widetilde{\tau}_{a,i},$$

where $\widetilde{\tau}_{a,i}$ is the length of the inflated active phase of episode i of arm a, and such inflated episodes don't have a passive phase.

Now, we make use this property and continue from Equation (19) by replacing the averaged per-time-step regret of an episode with the averaged per-time-step regret of an inflated episode as follows:

$$\begin{aligned} R(\textsc{SnR}) &= KT. \max_{a\in\mathcal{A}, i\in[e_a]} \frac{\sum_{k=1}^{j} c_{10}\sqrt{w(k).\log T}}{\widetilde{\tau}_{a,i}} \\ &\leq KT. \max_{a\in\mathcal{A}, i\in[e_a], k\in[j]} c_{10}\frac{\sqrt{w(k).\log T}}{\widetilde{\tau}(k)} \end{aligned} \tag{20}$$

where $\widetilde{\tau}(k)$ is the length of the inflated block $k \in [j]$.

What remains to be shown is that $\frac{\sqrt{w(k).\log T}}{\widetilde{\tau}(k)}$ term in Equation (20) is $O(\delta^{1/3}K^{-2/3}\log^{1/3} T)$.

By the nature of our construction of inflated episodes where each round has K time steps, we have for all arms a and episodes $i \in [e_a]$ that $\tau_{a,i} \leq \widetilde{\tau}_{a,i} = K.w_{a,i}$. As the construction (i.e., moving time steps from passive arms to active arms) happens at a round level, we have for every block $k \in [j]$ that $\tau(k) \leq \widetilde{\tau}(k) = K.w(k)$.

Thus, we have

$$\begin{aligned} \widetilde{\tau}(k) \geq \tau(k) &= \frac{1}{\delta}.\sqrt{\frac{\log T}{w(k)}} \\ &\geq \frac{1}{\delta}.\sqrt{\frac{\log T}{w(k)}} \\ \implies \widetilde{\tau}(k) &\geq \delta^{-2/3}K^{-1/6}\log^{-1/6} T. \end{aligned}$$

Substituting this in averaged inflated block regret expresssion $\frac{\sqrt{w(k).\log T}}{\widetilde{\tau}(k)}$ to minimize, we have

$$\begin{aligned} \frac{\sqrt{w(k).\log T}}{\widetilde{\tau}(k)} &= \sqrt{\frac{\log T}{K.\widetilde{\tau}(k)}} \\ &\leq \frac{\log^{1/2} T.\delta^{-1/3}.K^{-1/6}\log^{-1/6} T}{K^{1/2}} \\ &= \delta^{1/3}K^{-2/3}\log^{1/3} T. \end{aligned}$$

This completes the proof of the Theorem.

□

D Proof of Theorem 3

Theorem 3. *[Minimax Lower Bound] For any algorithm* ALG *and a drift limit $\delta > 0$, there exists a problem instance $\mu \in S_\delta$ such that,* ALG *incurs a expected regret of $\Omega(T\delta^{1/3}K^{1/3})$, where K is the number of arms.*

Proof of Theorem. Towards proving this theorem, first, we shall state and prove some useful information theoretic lemmas. Then, we will divide the time horizon into smaller blocks and lower bound the expected regret of each block using those lemmas. Finally, we shall aggregate the regrets of the individual blocks by adhering to problem specific limitations, specifically, the drift limit δ, to arrive at the final overall lower bound.

The Change of Measure Inequality presented in Lemma 12 generalizes that of Garivier & Kaufmann (2016) by accommodating non-stationary reward distributions of arms.

Lemma 12 (Non-stationary Change of Measure Inequality)**.** *Let ν and ν' be two non-stationary bandit instances (sets of reward distributions for each time-step) with k arms over time horizon $[T]$. For any bandit algorithm* ALG*, for any random variable Z with values in $[0,1]$ that is fully determinable from the trajectory (history) of an algorithmic run, H_T, i.e., Z is $\sigma(H_T)$- measurable, we have*

$$\sum_{i=1}^{k}\sum_{t=1}^{T} KL\left(\nu_{i,t}, \nu'_{i,t}\right) \mathbb{E}_\nu\left[\mathbb{1}\left\{\text{ALG}(t) = i\right\}\right] \geq KL\left(Ber\left(\mathbb{E}_\nu\left[Z\right]\right), Ber\left(\mathbb{E}_{\nu'}\left[Z\right]\right)\right) \tag{21}$$

where, $\mathbb{E}_\nu[X]$ is the expected value of random variable X under bandit instance ν, ALG(t) *is the arm played by* ALG *at time-step t, $KL(a,b)$ is the Kullback-Leibler divergence between distributions a and b, $Ber(a)$ is the Bernoulli distribution with expectation a.*

Proof. We prove this inequality by establishing two intermediate results:

$$\sum_{i=1}^{k}\sum_{t=1}^{T} KL\left(\nu_{i,t}, \nu'_{i,t}\right) \mathbb{E}_\nu\left[\mathbb{1}\left\{\text{ALG}(t) = i\right\}\right] = KL\left(\mathbb{P}_\nu^{H_{T+1}}, \mathbb{P}_{\nu'}^{H_{T+1}}\right), \quad \text{and} \tag{22}$$

$$KL\left(\mathbb{P}_\nu^{H_{T+1}}, \mathbb{P}_{\nu'}^{H_{T+1}}\right) \geq KL\left(Ber\left(\mathbb{E}_\nu\left[Z\right]\right), Ber\left(\mathbb{E}_{\nu'}\left[Z\right]\right)\right). \tag{23}$$

Here, $\mathbb{P}_\nu^{H_{T+1}}$ (resp. $\mathbb{P}_{\nu'}^{H_{T+1}}$) is the probability distribution under instance ν (resp. ν') of the algorithmic trajectory $H_{T+1} = (U_1, I_1, Y_1, \ldots, U_T, I_T, Y_T, U_{T+1})$. And, U_t, I_t, Y_t are random variables that correspond to internal randomness, arm pulled, and reward obtained respectively at time t.

We start with the right-hand-side (RHS) of step 1 (Equation (22)) and show it's equality to the left-hand-side (LHS). By definition,

$$KL\left(\mathbb{P}_\nu^{H_T}, \mathbb{P}_{\nu'}^{H_T}\right) = \sum_{h_{T+1}} \mathbb{P}_\nu\left\{H_{T+1} = h_{T+1}\right\} \log \frac{\mathbb{P}_\nu\left\{H_{T+1} = h_{T+1}\right\}}{\mathbb{P}_{\nu'}\left\{H_{T+1} = h_{T+1}\right\}},$$

where $h_{T+1} := (u_1, i_1, y_1, \ldots, u_T, i_T, y_T, u_{T+1})$ is a realisation of a trajectory of a bandit algorithm. We continue by writing the sought-after divergence as

$$\sum_{h_{T+1}} \mathbb{P}_\nu\left\{H_{T+1} = h_{T+1}\right\} \log\left(\frac{f_u(u_1).\mathbb{P}_\nu\left\{I_1 = i_1 | U_1 = u_1\right\}.\mathbb{P}_\nu\left\{Y_1 = y_1 | U_1 = u_1, I_1 = i_1\right\}.f_u(u_2)\ldots.}{f_u(u_1).\mathbb{P}_{\nu'}\left\{I_1 = i_1 | U_1 = u_1\right\}.\mathbb{P}_{\nu'}\left\{Y_1 = y_1 | U_1 = u_1, I_1 = i_1\right\}.f_u(u_2)\ldots.}\right).$$

The internal randomness function $f_u(\cdot)$ of the algorithm ALG is instance-agnostic. Also, the probability distribution of I_t under instances ν and ν' are identical when conditioned upon the the trajectory $(u_1, i_1, y_1, \ldots, u_{t-1}, i_{t-1}, y_{t-1}, u_t)$. We continue by writing the sought-after divergence as

$$\sum_{h_{T+1}} \mathbb{P}_\nu \left\{H_{T+1} = h_{T+1}\right\} \log \prod_{t=1}^{T} \frac{\nu_{i_t,t}(y_t)}{\nu'_{i_t,t}(y_t)} = \sum_{h_{T+1}} \mathbb{P}_\nu \left\{H_{T+1} = h_{T+1}\right\} \sum_{t=1}^{T} \log \frac{\nu_{i_t,t}(y_t)}{\nu'_{i_t,t}(y_t)}.$$

We represent the above expression as an expectation over all possible trajectories h_t under instance ν. At time step t, the deterministic arm played i_t is replaced by the random variable I_t. We continue the derivation as follows.

$$\begin{aligned}
KL\left(\mathbb{P}_\nu^{H_T}, \mathbb{P}_{\nu'}^{H_T}\right) &= \mathbb{E}_\nu \left[\sum_{t=1}^{T} \log \frac{\nu_{I_t,t}(y_t)}{\nu'_{I_t,t}(y_t)}\right] \\
&\overset{\text{(a)}}{=} \mathbb{E}_\nu \left[\sum_{t=1}^{T} \log \frac{\nu_{I_t,t}(y_t)}{\nu'_{I_t,t}(y_t)} \sum_{i=1}^{k} \mathbb{1}\left\{I_t = i\right\}\right] \\
&= \sum_{i=1}^{k} \sum_{t=1}^{T} \mathbb{E}_\nu \left[\mathbb{1}\left\{I_t = i\right\} \log \frac{\nu_{i,t}(y_t)}{\nu'_{i,t}(y_t)}\right] \\
&\overset{\text{(b)}}{=} \sum_{i=1}^{k} \sum_{t=1}^{T} \mathbb{E}_\nu \left[\mathbb{1}\left\{I_t = i\right\} \mathbb{E}_\nu \left[\log \frac{\nu_{i,t}(y_t)}{\nu'_{i,t}(y_t)}\right] | \mathbb{1}\left\{I_t = i\right\}\right] \\
&= \sum_{i=1}^{k} \sum_{t=1}^{T} \mathbb{E}_\nu \left[\mathbb{1}\left\{I_t = i\right\} KL(\nu_{i,t}, \nu'_{i,t})\right] \\
&= \sum_{i=1}^{k} \sum_{t=1}^{T} KL(\nu_{i,t}, \nu'_{i,t}) \mathbb{E}_\nu \left[\mathbb{1}\left\{I_t = i\right\}\right].
\end{aligned}$$

The (a) is due to the fact that at every time t, exactly one arm is played. In (b), the inner expectation is over all realisations of y_t. This completes the proof of Equation (22).

For the proof for Equation (23), to avoid repetition, we refer the reader to Garivier et al. (2019). This completes the proof of the Lemma 12. □

Next, we shall divide the time horizon into smaller blocks, and use this lemma to prove a lower bound on each of them individually. We shall finally aggregate them to get the final lower bound. Divide the time horizon $[T] := 1, 2, \ldots, T$ into blocks of size m, to be determined later. We get T/m[9] blocks in total. Throughout this section, we trade the notion of mean rewards μ (or $\mu_{i,t}$) in favour of the more generic notion of reward distributions ν (or $\nu_{i,t}$). Also, the time horizon of the block is $[m]$ instead of the global time horizon $[T]$.

Construction of 'confusing' problem instances. Consider two reward distributions ν, ν'. Let ν be a stationary instance with identical arms with $Ber\left(1/2\right)$ rewards, i.e., for $i \in \mathcal{A}$, all time steps $t \in [m]$, we have $\nu_{i,t} \sim Ber\left(1/2\right)$. One of the K arms is played at most in m/K number of time steps, i.e., $\exists i \in \mathcal{A} : \mathbb{E}_\nu\left[N_i\right] \leq m/K$, where N_i is the number of times arm i is played in the block (by the end of time step m). Without loss of generality, assume that it is arm $i = 1$ that satisfies the above condition, i.e., $\mathbb{E}_\nu\left[N_1\right] \leq m/K$.

We construct bandit problem instance ν' in such a way that the lesser played arm in instance ν (arm 1) is optimal in ν'.

[9] For technical clarity, we assume integrality of all quantities suitably.

In ν', all arms other than arm 1 have a stationary (for all times $t \in [m]$) $Ber\left(1/2\right)$ reward distribution. Whereas, arm 1 has a $Ber\left(1/2\right)$ reward distribution at the beginning and ending time-step of the block, but, has reward distributions with larger means (than $1/2$) in the intervening time-steps. Precisely,

$$\nu'_{i,t} = \begin{cases} Ber\left(\frac{1}{2}\right) & \text{if } i = 2, 3, \ldots, K \\ Ber\left(\frac{1}{2} + \frac{t-1}{m}.\varepsilon\right) & \text{if } i = 1, t \leq \left\lceil \frac{m}{2} \right\rceil \\ Ber\left(\frac{1}{2} + \frac{m-t}{m}.\varepsilon\right) & \text{if } i = 1, t > \left\lceil \frac{m}{2} \right\rceil. \end{cases} \tag{24}$$

We apply the non-stationary change of measure inequality stated in Lemma 12 to the instances ν and ν' with a choice of $Z = N_1/m$, the play fraction of the arm that is underplayed in ν, but is optimal in ν'.

We first upper bound the LHS before plugging it into the inequality:

$$\begin{aligned}
&\sum_{i=1}^{K}\sum_{t=1}^{m} KL\left(\nu_{i,t}, \nu'_{i,t}\right) \mathbb{E}_{\nu}\left[\mathbb{1}\left\{\textsc{Alg}(t) = i\right\}\right] \\
&\overset{(a)}{=} \sum_{t=1}^{m} KL\left(\nu_{1,t}, \nu'_{1,t}\right) \mathbb{E}_{\nu}\left[\mathbb{1}\left\{\textsc{Alg}(t) = 1\right\}\right] \\
&= \sum_{t=1}^{\left\lceil \frac{m}{2} \right\rceil} KL\left(Ber\left(\frac{1}{2}\right), Ber\left(\frac{1}{2} + \frac{t-1}{m}.\varepsilon\right)\right) \mathbb{E}_{\nu}\left[\mathbb{1}\left\{\textsc{Alg}(t) = 1\right\}\right] + \\
&\qquad \sum_{t=\left\lceil \frac{m}{2} \right\rceil + 1}^{m} KL\left(Ber\left(\frac{1}{2}\right), Ber\left(\frac{1}{2} + \frac{m-t}{m}.\varepsilon\right)\right) \mathbb{E}_{\nu}\left[\mathbb{1}\left\{\textsc{Alg}(t) = 1\right\}\right] \\
&\leq \sum_{t=1}^{\left\lceil \frac{m}{2} \right\rceil} KL\left(Ber\left(\frac{1}{2}\right), Ber\left(\frac{1+\varepsilon}{2}\right)\right) \mathbb{E}_{\nu}\left[\mathbb{1}\left\{\textsc{Alg}(t) = 1\right\}\right] + \\
&\qquad \sum_{t=\left\lceil \frac{m}{2} \right\rceil + 1}^{m} KL\left(Ber\left(\frac{1}{2}\right), Ber\left(\frac{1+\varepsilon}{2}\right)\right) \mathbb{E}_{\nu}\left[\mathbb{1}\left\{\textsc{Alg}(t) = 1\right\}\right] \qquad (25) \\
&= KL\left(Ber\left(\frac{1}{2}\right), Ber\left(\frac{1+\varepsilon}{2}\right)\right) \sum_{t=1}^{m} \mathbb{E}_{\nu}\left[\mathbb{1}\left\{\textsc{Alg}(t) = 1\right\}\right] \leq \varepsilon^2.\mathbb{E}_{\nu}\left[N_1\right]. \qquad (26)
\end{aligned}$$

The (a) is due to the fact that all arms other than arm 1 are identical in both the instances, i.e., for all time-steps $t \in [m]$ and arms $i \in \mathcal{A} \setminus \{1\}$, we have $KL\left(\nu_{i,t}, \nu'_{i,t}\right) = 0$.

The Equation (25) upper bounds the information gathered about the distinguishability of the arm 1 (from stationary arm 1 with $Ber\left(1/2\right)$ rewards in instance ν) as if it were statinonary with $Ber\left(\frac{1+\varepsilon}{2}\right)$ rewards at all times $t \in [m]$ in ν'.

We now plug the Equation (26) into Equation (21) with a choice of $Z = {}^{N_1}/_{m}$ to get

$$\begin{aligned}
\varepsilon^2 \mathbb{E}_\nu\left[N_1\right] &\geq KL\left(Ber\left(\frac{\mathbb{E}_\nu\left[N_1\right]}{m}\right), Ber\left(\frac{\mathbb{E}_{\nu'}\left[N_1\right]}{m}\right)\right) \\
&\overset{(a)}{\geq} 2\left(\frac{\mathbb{E}_\nu\left[N_1\right]}{m} - \frac{\mathbb{E}_{\nu'}\left[N_1\right]}{m}\right)^2 \qquad \text{(By Pinsker's Inequality)} \\
\implies \sqrt{\frac{1}{2}.\varepsilon^2 \mathbb{E}_\nu\left[N_1\right]} &\geq \frac{\mathbb{E}_{\nu'}\left[N_1\right]}{m} - \frac{\mathbb{E}_\nu\left[N_1\right]}{m} \\
\implies \frac{\mathbb{E}_{\nu'}\left[N_1\right]}{m} &\leq \frac{\mathbb{E}_\nu\left[N_1\right]}{m} + \sqrt{\frac{1}{2}.\varepsilon^2 \mathbb{E}_\nu\left[N_1\right]} \\
&\overset{(b)}{\leq} \frac{1}{K} + \sqrt{\frac{1}{2}.\varepsilon^2.\frac{m}{K}}.
\end{aligned}$$

Here, (a) is by Pinsker's inequality, and (b) is due to $\mathbb{E}_\nu\left[N_1\right] \leq {}^{m}/_{K}$. By fixing $\varepsilon = \sqrt{{}^{K}/_{8m}}$, we get $\frac{\mathbb{E}_{\nu'}[N_1]}{m} \leq {}^{1}/_{K} + {}^{1}/_{4} \leq {}^{3}/_{4}$, since $K \geq 2$. Thus, the sub-optimal arms in instance ν' are played for more than a constant fraction of times in expectation, i.e., $\sum_{i \in \mathcal{A}\setminus\{1\}} \mathbb{E}_{\nu'}\left[N_i\right] \geq {}^{m}/_{4}$. In instance ν', let these sub-optimal arms be played by ALG at time steps $t_1, t_2, \ldots, t_x$ for some $x \geq {}^{m}/_{4}$. Then, the expected regret in the block (denoted by $R^b(\text{ALG})$) is

$$\mathbb{E}_{\nu'}\left[R^b(\text{ALG})\right] = \sum_{i=1}^{x} \mu^*_{t_i} - \mu_{\text{ALG}(t_i),t_i} = \sum_{i=1}^{x} \mu_{1,t_i} - \mu_{2,t_i},$$

where the second equality is because arm 1 is optimal throughout the block and the reward mean of any sub-optimal arm is identical to that of arm 2.

By design (in Equation (24)), the set of reward mean gaps, $\{|\mu_{1,t} - \mu_{2,t}|\}_{t\in[m]}$, throughout the block $[m]$ is $\{0, 0, \frac{\varepsilon}{m}, \frac{\varepsilon}{m}, \frac{2\varepsilon}{m}, \ldots, \frac{\varepsilon}{2}\}$. We lower bound $\sum_{i=1}^{x} \mu_{1,t_i} - \mu_{2,t_i}$ with the least possible sum of $x \geq {}^{m}/_{4}$ values from the set of gaps as follows:

$$\begin{aligned}
\mathbb{E}_{\nu'}\left[R^b(\text{ALG})\right] &= \sum_{i=1}^{x} \mu_{1,t_i} - \mu_{2,t_i} \\
&\geq 2.\sum_{i=1}^{m/8} \frac{(i-1).\varepsilon}{m} = \frac{2\varepsilon}{m} \sum_{i=0}^{m/8-1} i \\
&= \frac{2\varepsilon}{m}.\frac{m^2 - 8m}{64} = \frac{\varepsilon}{32}.(m-8) \overset{(a)}{=} \sqrt{\frac{K}{8m}}.\frac{m-8}{32} \\
&= \frac{\sqrt{k}}{64\sqrt{2}}.\left(m^{1/2} - 8/m^{1/2}\right) = \Omega\left(\sqrt{Km}\right).
\end{aligned} \tag{27}$$

The (a) is due to $\varepsilon = \sqrt{{}^{K}/_{8m}}$.

Also, at each time step in $t \in [m]$, the maximum reward mean gap between any two arms is $\mu_{1,t} - \mu_{2,t} \leq \varepsilon$. So, the regret in the block is trivially upper bounded as follows:

$$\mathbb{E}_{\nu'}\left[R^b(\text{ALG})\right] \leq m.\varepsilon = m.\sqrt{{}^{K}/_{8m}} = O\left(\sqrt{Km}\right). \tag{28}$$

The following lemma is a consequence of Equations (27) and (28).

Lemma 13. *For a block of time period m, there exists a bandit instance ν' (as in expression 24) such that, for any algorithm* ALG*, its expected regret $\mathbb{E}_{\nu'}\left[R^b(\text{ALG})\right]$ is $\Theta(\sqrt{Km})$.*

Next, we aggregate the regret across T/m blocks to get the overall regret. Note that the instances ν and ν' have a drift limit implication as follows:

$$\delta \geq \frac{\varepsilon}{m} \quad \iff \quad \delta \geq \frac{K^{1/2}}{\sqrt{8}.m^{3/2}} \quad \iff \quad m^{3/2} \geq \frac{1}{\sqrt{8}}.\frac{K^{1/2}}{\delta} \quad \iff \quad \sqrt{m} \geq \frac{1}{\sqrt{2}}.\frac{K^{1/6}}{\delta^{1/3}}. \tag{29}$$

Now, the total regret is lower bounded by the number of blocks multiplied by the lower bound of regret within each block. Thus,

$$\begin{aligned}\mathbb{E}\left[R(\textsc{Alg})\right] &= \frac{T}{m} \times \Theta(\sqrt{Km}) = \frac{T.K^{1/2}}{\Theta(\sqrt{m})} \\ &\overset{(a)}{\geq} T.K^{1/2}.\sqrt{2}\delta^{1/3}K^{-1/6} = \Omega\left(TK^{1/3}\delta^{1/3}\right),\end{aligned}$$

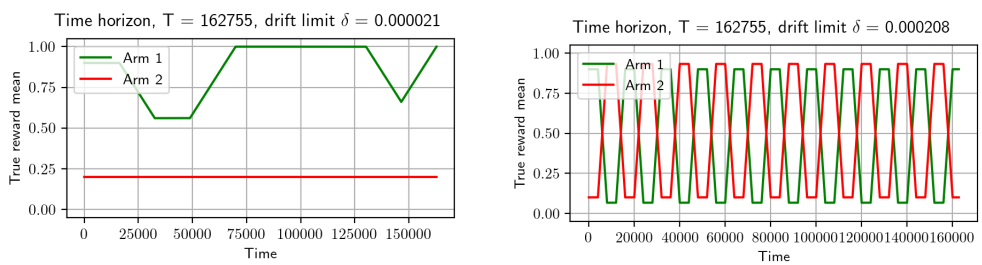
where (a) uses Equation (29). This completes the proof of the Theorem 3. □

E Additonal simulations

In this section, we plot supporting graphs that were not included in the main paper.

E.1 Problem Instances

We plot the problem instances whose regret plots were analysed in the main paper Section 6.

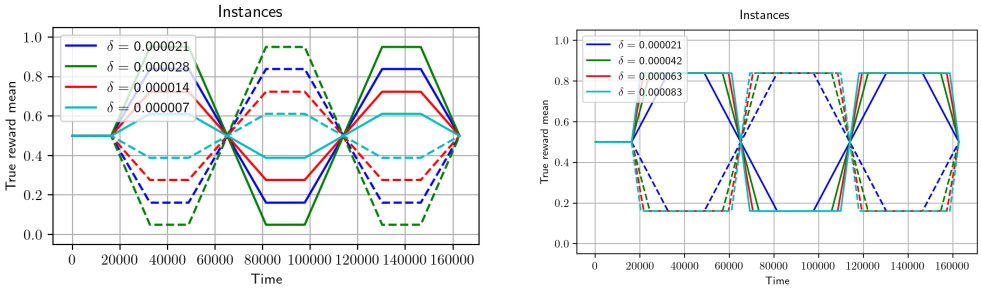


(a) (b)

Figure 4: Illustration of problem instances, i.e., the true reward means of the arms. In (a), the arms are well separated with no change in optimal arm's identity throughout. Here, the drift limit $\delta = 1/10.c_9.\log T \simeq 0.000021$. In (b), the arms experience short stretches of stationarity and drift alternately. The optimal arm's identity toggles with every drift. Here, we have a relatively large drift limit $\delta = 1/c_9.\log T \simeq 0.00021$. The choice of δs are arbitrary to well describe the differences in instances.

E.2 Algorithmic Trajectories

In this subsection, we show two algorithmic trajectories (for two different instances) that can help better understand the functioning of our algorithm SNR in Figs. 6 and 7.



(a) (b)

Figure 5: Two sets of 4 structurally similar problems with varying drift limits δ. In (a), the instances have common periods of stationarity and drift. The amount of drift varies with the corresponding $\delta = a, 2a, 3a, 4a$, for $a = {}^{1}\!/_{c_{10} \log T} \simeq 0.000007$ values. In (b), the instances have equal total cumulative drift. To achieve that drift, the duration of drift varies with the corresponding $\delta = a, 2a, 3a, 4a$, for $a = {}^{1}\!/_{c_{11} \log T} \simeq 0.000021$ values. The choice of δs are arbitrary to well describe the differences in instances.

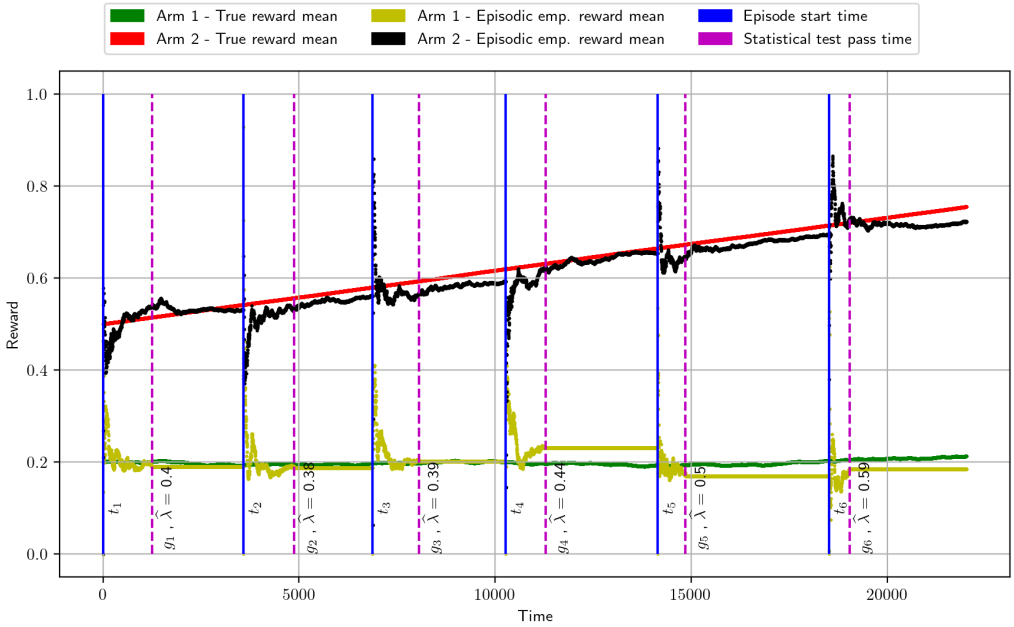


Figure 6: Illustration of the trajectory of SNR for a problem instance whose gap, $\Delta_{2,t}$, and thus the detectable gap, $\lambda_{2,t}$ (not shown in picture) increases with time t. One can observe that the statistical test passes sooner (a shorter active phase $[t_{2,i} + 1, g_{2,i}]$ of episode i) for larger observed gaps, $\widehat{\lambda}$s. Also, the snooze period is more (a longer passive phase $[g_{2,i} + 1, t_{2,i+1}]$) for larger observed gaps. The empirical means plotted are measured from the beginning of the episode till the current time-step. Note that the empirical means of the sub-optimal arm remains unchanged during the passive phase due to it not being played.

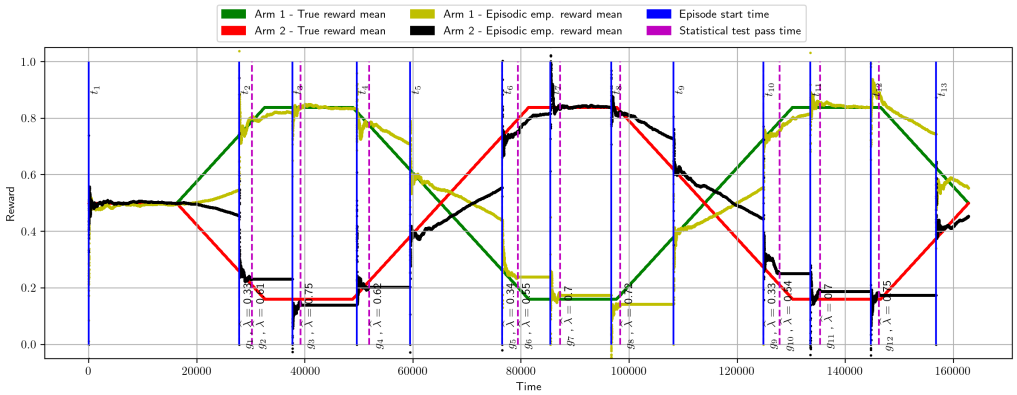


Figure 7: Illustration of the trajectory of SNR for a problem instance with oscillating arms. Note that episodes $1, 5$, and 9 do not have a passive phase. Essentially, compared with the length of the active phases, the gaps detected $\widehat{\lambda}$s were not sufficiently large to warrant snoozing the sub-optimal arm. However, stretches of time where the arms are well separated enjoy substantial passive phases.