An Information-Theoretic Analysis of Nonstationary Bandit Learning

Seungki Min [1] **Daniel Russo** [2]

Abstract

In nonstationary bandit learning problems, the decision-maker must continually gather information and adapt their action selection as the latent state of the environment evolves. In each time period, some latent optimal action maximizes expected reward under the environment state. We view the optimal action sequence as a stochastic process, and take an information-theoretic approach to analyze attainable performance. We bound per-period regret in terms of the *entropy rate of the optimal action process*. The bound applies to a wide array of problems studied in the literature and reflects the problem's information structure through its information-ratio.

1. Introduction

We study the problem of learning in interactive decision-making. Across a sequence of time periods, a decision-maker selects actions, observes outcomes, and associates these with rewards. They hope to earn high rewards, but this may require investing in gathering information.

Most of the literature studies stationary environments — where the likelihood of outcomes under an action is fixed across time.[1] Efficient algorithms limit costs required to converge on optimal behavior. We study the design and analysis of algorithms in nonstationary environments, where converging on optimal behavior is impossible.

In our model, the latent state of the environment in each time period is encoded in a parameter vector. These parameters are unobservable, but evolve according to a known stochastic process. The decision-maker hopes to earn high rewards by adapting their action selection as the environment evolves. This requires continual learning from interaction and striking a judicious balance between exploration and exploitation. Uncertainty about the environment's state cannot be fully resolved before the state changes and this necessarily manifests in suboptimal decisions. Strong performance is impossible under adversarial forms of nonstationarity but is possible in more benign environments. Why are A/B testing, or recommender systems, widespread and effective even though nonstationarity is a ubiquitous concern? Quantifying the impact different forms of nonstationarity have on decision-quality is, unfortunately, quite subtle.

Our contributions. We provide a novel information-theoretic analysis that bounds the inherent degradation of decision-quality in changing environments. Note that the latent state evolution of the environment induces a latent optimal action process — where the optimal action at any time step is that one that maximizes expected reward conditioned on the current environment parameter. We bound per-period regret in terms of the *entropy-rate of the optimal action process*. The entropy rate of a stochastic process is a fundamental concept in the theory of communications. We use the entropy rate to measure the extent to which the evolving state of the environment manifests in surprising and erratic evolution of the optimal action process. Subsection 1.1 gives an example of nonstationarity, inspired by A/B testing, in which the entropy rate of the parameter process is large but the entropy rate of the action process is small. We believe this distinction is essential.

We enrich this result in two ways. First, we provide a matching lower bound. This exhibits a sequence of problems with varying entropy rate under which no algorithm could meaningfully outperform our upper bounds. Second, we provide stylized upper bounds on the entropy rate, expressed in terms of the number of changes in the optimal action or the 'effective time horizon', a new notion introduced in this paper. Combined with our general result, they give rise to regret bounds that are rather interpretable and consistent with the previous results in the literature.

In addition to the problem's entropy rate, our general bounds depend on the algorithm's information ratio. First introduced by Russo & Van Roy (2016), the information ratio measures the per-period price an algorithm pays to acquire

*Equal contribution [1]Department of Industrial and Systems Engineering, KAIST [2]Division of Decision Risk and Operations, Columbia Business School. Correspondence to: Seungki Min <skmin@kaist.ac.kr>, Daniel Russo <djr2174@gsb.columbia.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]An alternative style of result lets the environment change, but tries only to compete with the best fixed action in hindsight.

new information. It has been shown to properly capture the complexity of learning in a range of widely studied problems, and recent works link it to generic limits on when efficient learning is possible (Lattimore, 2022; Foster et al., 2022).

Because the information-ratio framework covers many of the most important sequential learning problems, our framework applies to *nonstationary variants* of many of the most important sequential learning problems. A secondary contribution of our work is extending information-ratio analysis to cover contextual bandits, resolving an open question highlighted by Neu et al. (2022). See Section 4.5.

This work emphasizes understanding of the limits of attainable performance. Thankfully, most results apply to Thompson sampling (TS), one of the most widely used learning algorithms in practice. In some problems, TS is far from optimal, and better bounds are attained with Information-Directed Sampling (Russo & Van Roy, 2018).

1.1. An illustrative Bayesian model of nonstationarity

Consider a multi-armed bandit environment where two types of nonstationarity coexist – a common variation that affects the performance of all arms, and idiosyncratic variations that affect the performance of individual arms separately. More explicitly, let us assume that the mean reward of arm a at time t is given by

$$\mu_{t,a} = \theta_t^{\mathrm{cm}} + \theta_{t,a}^{\mathrm{id}},$$

where $(\theta_t^{\mathrm{cm}})_{t\in\mathbb{N}}$ and $(\theta_{t,a}^{\mathrm{id}})_{t\in\mathbb{N}}$'s are latent stochastic processes describing common and idiosyncratic disturbances. While deferring the detailed description to Appendix B, we introduce two hyperparameters τ^{cm} and τ^{id} in our generative model to control the time scale of these two types of variations.[2]

Inspired by real-world A/B tests (Wu et al., 2022), we imagine a two-armed bandit instance involving a common variation that is much more erratic than idiosyncratic variations. Common variations reflect exogenous shocks to user behavior which impacts the reward under all treatment arms. Figure 1 visualizes such an example, a sample path generated with the choice of $\tau^{\mathrm{cm}} = 10$ and $\tau^{\mathrm{id}} = 50$. Observe that the optimal action A_t^* has changed only five times throughout 1,000 periods. Although that the environment itself is highly nonstationary and unpredictable due to the common variation term, the optimal action sequence $(A_t^*)_{t\in\mathbb{N}}$ is relatively stable and predictable since it depends only on the idiosyncratic variations.

[2]We assume that $(\theta_t^{\mathrm{cm}})_{t\in\mathbb{N}}$ is a zero-mean Gaussian process satisfying $\mathrm{Cov}[\theta_s^{\mathrm{cm}}, \theta_t^{\mathrm{cm}}] = \exp\left(-\frac{1}{2}\left(\frac{t-s}{\tau^{\mathrm{cm}}}\right)^2\right)$ so that τ^{cm} determines the volatility of the process. Similarly, the volatility of $(\theta_{t,a}^{\mathrm{id}})_{t\in\mathbb{N}}$ is determined by τ^{id}.

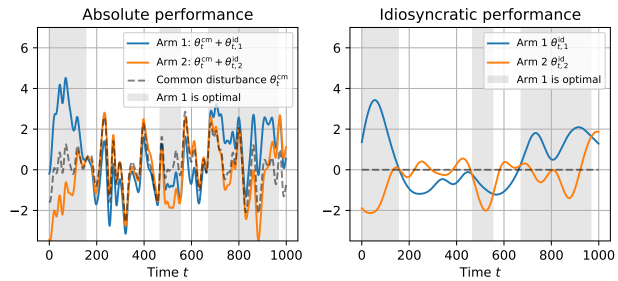

Figure 1. A two-arm bandit environment with two types of nonstationarity – a common variation $(\theta_t^{\mathrm{cm}})_{t\in\mathbb{N}}$ generated with a time-scaling factor $\tau^{\mathrm{cm}} = 10$, and idiosyncratic variations $(\theta_{t,a}^{\mathrm{id}})_{t\in\mathbb{N},a\in\mathcal{A}}$ generated with a time-scaling factor $\tau^{\mathrm{id}} = 50$. While absolute performance of two arms are extremely volatile (left), their idiosyncratic performances are relatively stable (right).

Now we ask — *How difficult is this learning task? Which type of nonstationarity determines the difficulty?* A quick numerical investigation shows that the problem's difficulty is mainly determined by the frequency of optimal action switches, rather than volatility of common variation.

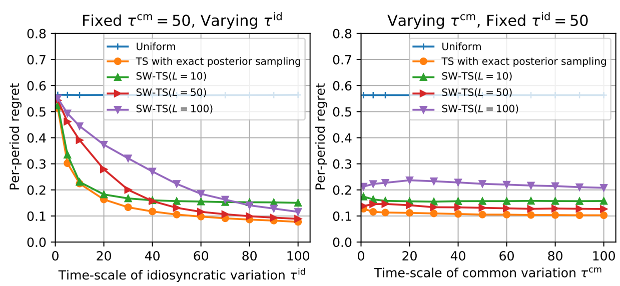

Figure 2. Performance of algorithms in two-armed bandit environments, with difference choices of time-scaling factors τ^{cm} (common variation) and τ^{id} (idiosyncratic variations). Each data point reports per-period regret averaged over 1,000 time periods and 1,000 runs of simulation.

See Figure 2, where we report the effect of τ^{cm} and τ^{id} on the performance of several bandit algorithms (namely, Thompson sampling with exact posterior sampling,[3] and Sliding-Window TS that only uses recent $L \in \{10, 50, 100\}$ observations; see Appendix B for the details). Remarkably, their performances appear to be sensitive only to τ^{id} but not to τ^{cm}, highlighting that nonstationarity driven by common variation is benign to the learner.

We remark that our information-theoretic analyses predict this result. Theorem 4.8 shows that the complexity of a nonstationary environment can be sufficiently characterized

[3]In order to perform exact posterior sampling, it exploits the specified nonstationary structure as well as the values of τ^{cm} and τ^{id}.

by the entropy rate of the optimal action sequence, which should depend only on τ^{id} but not on τ^{cm} in this example. Theorem 4.4 further expresses the entropy rate in terms of effective horizon, which corresponds to τ^{id} in this example.

1.2. Comments on the use of prior knowledge

A substantive discussion of Bayesian, frequentist, and adversarial perspectives on decision-making uncertainty is beyond the scope of this short paper. We make two quick observations. First, where does a prior like the one in Figure 1 come from? One answer is that company may run many thousands of A/B tests, and an informed prior may let them transfer experience across tests (Azevedo et al., 2019). In particular, experience with past tests may let them calibrate τ^{id}, or form hierarchical prior where τ^{id} is also random. Second, Thompson sampling with a stationary prior is perhaps the most widely used bandit algorithm. One might view the model in Section 1.1 as a more conservative way of applying TS that guards against a certain magnitude of nonstationarity.

1.3. Literature review

Most existing theoretical studies on nonstationary bandit experiments adopt adversarial or frequentist viewpoints in the modeling of nonstationarity, typically falling into two categories – "switching environments" and "drifting environments".

Switching environments consider a situation where underlying reward distributions change at unknown times (often referred to as changepoints or breakpoints). Denoting the total number of changes over T periods by N, it was shown that the cumulative regret $\tilde{O}(\sqrt{NT})$ is achievable: e.g., Exp3.S (Auer et al., 2002; Auer, 2002), Discounted-UCB (Kocsis & Szepesvári, 2006), Sliding-Window UCB (Garivier & Moulines, 2008), and more complicated algorithms that actively detect the changepoints (Auer et al., 2019; Chen et al., 2019). More recent studies improve upon this result by showing that $\tilde{O}(\sqrt{ST})$ is achievable where S only counts the number of best arm switches (Abbasi-Yadkori et al., 2022; Suk & Kpotufe, 2022). Our results reveal that Thompson sampling also achieves the regret bound $\tilde{O}(\sqrt{ST})$ (see Theorem 4.8 with Theorem 4.3) in a wide range of problems beyond k-armed bandits (see Section 4.4).

Another stream of work considers *drifting environments*. Denoting the total variation in the underlying reward distribution by V (often referred to as variation budget, e.g., $V := \sum_{t=2}^{T} \|\theta_t - \theta_{t-1}\|_\infty$), it was shown that the cumulative regret $\tilde{O}(V^{1/3}T^{2/3})$ is achievable (Besbes et al., 2014; 2015; Cheung et al., 2019). Building a tight connection between these results and ours is an important direction for future work. We comment on this in the conclusion.

We adopt *Bayesian viewpoints* to describe nonstationary environments: changes in the underlying reward distributions (more generally, changes in outcome distributions) are driven by a stochastic process. Such a viewpoint dates back to the earliest work of Whittle (1988) which introduces the term 'restless bandits' and has motivated subsequent work (Slivkins & Upfal, 2008; Chakrabarti et al., 2008; Jung & Tewari, 2019). On the other hand, since Thompson sampling (TS) has gained its popularity as a Bayesian bandit algorithm, its variants have been proposed for nonstationary settings accordingly: e.g., Dynamic TS (Gupta et al., 2011), Discounted TS (Raj & Kalyani, 2017), Sliding-Window TS (Trovo et al., 2020), TS with Bayesian changepoint detection (Mellor & Shapiro, 2013; Ghatak, 2020), and Predictive Sampling (Liu et al., 2023). Although the Bayesian framework can flexibly model various types of nonstationarity, this literature rarely presents performance guarantees that apply to a broad class of models.

Our analysis adopts an *information-theoretic approach* introduced by Russo & Van Roy (2016), which has been motivating design and analysis of TS-like algorithms for complicated online optimization problems (Russo & Van Roy, 2018; Liu et al., 2018; Dong et al., 2019; Hao et al., 2021; Lattimore & Gyorgy, 2021; Russo & Van Roy, 2022; Neu et al., 2022; Liu et al., 2023). Our work can be seen as a natural extension of Russo & Van Roy (2016) to nonstationary bandit problems, systematically inheriting the wealth of the previous results established for stationary bandit problems (see Section 4.4). Notably, a recent work of Liu et al. (2023) also adopts the information-theoretic approach for nonstationary settings. We leave a discussion in the conclusion.

2. Problem Setup

A decision-maker interacts with a changing environment across rounds $t \in \mathbb{N} := \{1, 2, 3, \ldots\}$. In period t, the decision-maker selects some action A_t from a finite set $\mathcal{A}$, observes an outcome O_t, and associates this with reward $R_t = R(O_t, A_t)$ that depends on the outcome and action through a known utility function $R(\cdot)$.

There is a function g, an i.i.d sequence of disturbances $W = (W_t)_{t\in\mathbb{N}}$, and a sequence of latent environment states $\theta = (\theta_t)_{t\in\mathbb{N}}$ taking values in Θ, such that outcomes are determined as

$$O_t = g(A_t, \theta_t, W_t). \tag{1}$$

Write potential outcomes as $O_{t,a} = g(a, \theta_t, W_t)$ and potential rewards as $R_{t,a} = R(O_{t,a}, a)$. Equation (1) is equivalent to specifying a known probability distribution over outcomes for each choice of action and environment state.

The decision-maker wants to earn high rewards even as the environment evolves, but cannot directly observe the environment state or influence its evolution. Specifically, the decision-maker's actions are determined by some choice of policy $\pi = (\pi_1, \pi_2, \ldots)$. At time t, an action $A_t = \pi_t(\mathcal{F}_{t-1}, \tilde{W}_t)$ is a function of the observation history $\mathcal{F}_{t-1} = (A_1, O_1, \ldots, A_{t-1}, O_{t-1})$ and an internal random seed $\tilde{W}_t$ that allows for randomness in action selection. Reflecting that the seed is exogenously determined, assume $\tilde{W} = (\tilde{W}_t)_{t\in\mathbb{N}}$ is jointly independent of the outcome disturbance process W and state process $\theta = (\theta_t)_{t\in\mathbb{N}}$. That actions do not influence the environment's evolution can be written formally through the conditional independence relation $(\theta_s)_{s\geq t+1} \perp \mathcal{F}_t \mid (\theta_\ell)_{\ell\leq t}$.

The decision-maker wants to select a policy π that accumulates high rewards as this interaction continues. They know all probability distributions and functions listed above, but are uncertain about how environment states will evolve across time. To perform 'well', they need to continually gather information about the latent environment states and carefully balance exploration and exploitation.

Rather than measure the reward a policy generates, it is helpful to measure its regret. We define the *T-period per-period regret* of a policy π to be

$$\bar{\Delta}_T(\pi) := \frac{\mathbb{E}_\pi\left[\sum_{t=1}^T \left(R_{t,A_t^*} - R_{t,A_t}\right)\right]}{T},$$

where the latent optimal action A_t^* is a function of the latent state θ_t satisfying $A_t^* \in \arg\max_{a\in\mathcal{A}} \mathbb{E}[R_{t,a} \mid \theta_t]$. We further define the *regret rate* of policy π as its limit value,

$$\bar{\Delta}_\infty(\pi) := \limsup_{T\to\infty} \bar{\Delta}_T(\pi).$$

It measures the (long-run) per-period degradation in performance due to uncertainty about the environment state.

Remark 2.1. The use of a limit supremum and Cesàro averages is likely unnecessary under some technical restrictions. For instance, under Thompson sampling applied to Examples 2.4–2.7, if the latent state process $(\theta_t)_{t\in\mathbb{N}}$ is ergodic, we conjecture that $\bar{\Delta}_\infty(\pi) = \lim_{t\to\infty} \mathbb{E}_\pi\left[R_{t,A_t^*} - R_{t,A_t}\right]$.

Our analysis proceeds under the following assumption, which is standard in the literature.

Assumption 2.2. There exists σ such that, conditioned on $\mathcal{F}_{t-1}$, $R_{t,a}$ is sub-Gaussian with variance proxy σ^2.

2.1. 'Stationary processes' in 'nonstationary bandits'

The way the term 'nonstationarity' is used in the bandit learning literature could cause confusion as it conflicts with the meaning of 'stationarity' in the theory of stochastic process, which we use elsewhere in this paper.

Definition 2.3. A stochastic process $X = (X_t)_{t\in\mathbb{N}}$ is (strictly) stationary if for each integer t, the random vector $(X_{1+m}, \ldots, X_{t+m})$ has the same distribution for each choice of m.

'Nonstationarity', as used in the bandit learning literature, means that *realizations* of the latent state θ_t may differ at different time steps. The decision-maker can gather information about the current state of the environment, but it may later change. Nonstationarity of the stochastic process $(\theta_t)_{t\in\mathbb{N}}$, in the language of probability theory, arises when apriori there are predictable differences between environment states at different timesteps – e.g., if time period t is nighttime then rewards tend to be lower than daytime. It is often clearer to model predictable differences like that through contexts, as in Example 2.7.

2.2. Examples

Many interactive decision-making problems can be naturally written as special cases of our general protocol, where actions generate outcomes that are associated with rewards.

Our first example describes a bandit problem with independent arms, where outcomes generate information only about the selected action.

Example 2.4 (k-armed bandit)**.** Consider a website who can display one among k ads at a time and gains one dollar per click. For each ad $a \in [k] := \{1, \ldots, k\}$, the potential outcome/reward $O_{t,a} = R_{t,a} \sim \text{Bernoulli}(\theta_{t,a})$ is a random variable representing whether the ad a is clicked by the t^{th} visitor if displayed, where $\theta_{t,a} \in [0, 1]$ represents its click-through-rate. The platform only observes the reward of the displayed ad, so $O_t = R_{t,A_t}$.

Full information online optimization problems fall at the other extreme. There the potential observation $O_{t,a}$ does not depend on the chosen action a, so purposeful information gathering is unnecessary. The next example was introduced by Cover (1991) and motivates such scenarios.

Example 2.5 (Log-optimal online portfolios)**.** Consider a small trader who has no market impact. In period t they have wealth W_t which they divide among k possible investments. The action A_t is chosen from a feasible set of probability vectors, with $A_{t,i}$ denoting the proportion of wealth invested in stock i. The observation is $O_t \in \mathbb{R}_+^k$ where $O_{t,i}$ is the end-of-day value of \$1 invested in stock i at the start of the day and the distribution of O_t is parameterized by θ_t. Because the observation consists of publicly available data, and the trader has no market impact, O_t does not depend on the investor's decision. Define the reward function $R_t = \log\left(O_t^\top A_t\right)$. Since wealth evolves according to the equation $W_{t+1} = \left(O_t^\top A_t\right) W_t$,

$$\sum_{t=1}^{T-1} R_t = \log(W_T/W_1).$$

Many problems lie in between these extremes. We give two examples. The first is a matching problem. Many pairs of individuals are matched together and, in addition to the cumulative reward, the decision-maker observes feedback on the quality of outcome from each individual match. This kind of observation structure is sometimes called "semi-bandit" feedback (Audibert et al., 2014).

Example 2.6 (Matching)**.** Consider an online dating platform with two disjoint sets of individuals $\mathcal{M}$ and $\mathcal{W}$. On each day t, the platform suggests a matching of size k, $A_t \subset \{(m,w) : m \in \mathcal{M}, w \in \mathcal{W}\}$ with $|A_t| \leq k$. For each pair (m,w), their match quality is given by $\theta_{t,(m,w)}$. The platform observes the quality of individual matches, $O_t = \left(\theta_{t,(m,w)} : (m,w) \in A_t\right)$, and earns their average, $R_t = \frac{1}{k}\sum_{(m,w)\in A_t} \theta_{t,(m,w)}$.

Our final example is a contextual bandit problem. Here an action is itself more like a policy — it is a rule for assigning treatments on the basis of an observed context. Observations are richer than in the k-armed bandit. The decision-maker sees not only the reward a policy generated but also the context in which it was applied.

Example 2.7 (Contextual bandit)**.** Suppose that the website described in Example 2.4 can now access additional information about each visitor, denoted by $X_t \in \mathcal{X}$. The website observes the contextual information X_t, chooses an ad to display, and then observes whether the user clicks. To represent this task using our general protocol, we let the decision space $\mathcal{A}$ be the set of mappings from the context space $\mathcal{X}$ to the set of ads $\{1,\ldots,k\}$, the decision $A_t \in \mathcal{A}$ be a personalized advertising rule, and the observation $O_t = (X_t, R_t)$ contains the observed visitor information and the reward from applying the ad $A_t(X_t)$. Rewards are drawn according to $R_t \mid X_t, A_t, \theta_t \sim \text{Bernoulli}(\phi_{\theta_t}(X_t, A_t(X_t)))$, where $\phi_\theta : \mathcal{X} \times [k] \to [0,1]$ is a parametric click-through-rate model. Assume $X_{t+1} \perp (A_t, \theta) \mid X_t, \mathcal{F}_{t-1}$. This assumption means that advertising decisions cannot influence the future contexts and that parameters of the click-through rate model $\theta = (\theta_t)_{t\in\mathbb{N}}$ cannot be inferred passively by observing contexts.

3. Information Theoretic Preliminaries

The entropy of a discrete random variable X, defined by $H(X) = -\sum_x \mathbb{P}(X=x)\log(\mathbb{P}(X=x))$, measures the uncertainty in its realization. The entropy rate of a stochastic process $(X_1, X_2, \ldots)$ is the rate at which entropy of the partial realization $(X_1,\ldots,X_t)$ accumulates as t grows.

Definition 3.1. The *T-period entropy rate* of a stochastic process $X = (X_t)_{t\in\mathbb{N}}$, taking values in a discrete set, is

$$\begin{aligned}\bar{H}_T(X) &:= \frac{H([X_1, \ldots X_T])}{T} \\ &= \frac{1}{T}\sum_{t=1}^{T} H(X_t|X_{t-1},\ldots,X_1).\end{aligned}$$

The *entropy rate* is defined as its limit value:

$$\bar{H}_\infty(X) := \limsup_{T\to\infty} \bar{H}_T(X).$$

If X is a stationary stochastic process, then

$$\bar{H}_\infty(X) = \lim_{t\to\infty} H(X_t|X_{t-1},\ldots,X_1). \tag{2}$$

The form (2) is especially elegant. The entropy rate of a stationary stochastic process is the residual uncertainty in the draw of X_t which cannot be removed by knowing the draw of $X_{t-1},\ldots,X_1$. Processes that evolve quickly and erratically have high entropy rate. Those that tend to change infrequently (i.e., $X_t = X_{t-1}$ for most t) or change predictably will have low entropy rate.

4. Information-Theoretic Analysis of Dynamic Regret

We apply the information theoretic analysis of Russo & Van Roy (2016) and establish upper bounds on the per-period regret, expressed in terms of (1) the algorithm's information ratio, and (2) the entropy rate of the optimal action process.

4.1. Preview: special cases of our result

We begin by giving a special case of our result. It bounds the regret of Thompson sampling in terms of the reward variance proxy σ^2, the number of actions $|\mathcal{A}|$, and the entropy rate of the optimal action process $\bar{H}_T(A^*_{1:T})$. Thompson sampling is denoted by π^{TS} and is defined by the probability matching property:

$$\mathbb{P}(A_t = a \mid \mathcal{F}_{t-1}) = \mathbb{P}(A^*_t = a \mid \mathcal{F}_{t-1}), \tag{3}$$

which holds for all $t \in \mathbb{N}$, $a \in \mathcal{A}$. Actions are chosen by sampling from the posterior distribution of the optimal action.

Corollary 4.1. *Under any problem in the scope of our problem formulation,*

$$\bar{\Delta}_T(\pi^{\text{TS}}) \leq \sigma\sqrt{2\cdot|\mathcal{A}|\cdot\bar{H}_T(A^*)},$$

$$\textit{and}\quad \bar{\Delta}_\infty(\pi^{\text{TS}}) \leq \sigma\sqrt{2\cdot|\mathcal{A}|\cdot\bar{H}_\infty(A^*)}.$$

This result naturally covers a wide range of bandit learning tasks while highlighting that the entropy rate of optimal action process captures the level of degradation due to nonstationarity. Note that it includes as a special case the well-known regret upper bound established for a stationary k-armed bandit: when $A^*_1 = \ldots = A^*_T$, we have $H([A^*_1,\ldots,A^*_T]) = H(A^*_1) \leq \log k$, and thus $\bar{\Delta}_T(\pi^{\text{TS}}) \leq \tilde{O}(\sigma\sqrt{k/T})$.

According to (2), the entropy rate is small when the conditional entropy $H(A_t^* \mid A_1^*, \ldots, A_{t-1}^*)$ is small. That is, the entropy rate is small if most uncertainty in the optimal action A_t^* is removed through knowledge of the past optimal actions. Of course, Thompson sampling does not observe the environment states or the corresponding optimal actions, so its dependence on this quantity is somewhat remarkable.

The dependence of regret on the number of actions, $|\mathcal{A}|$, is unavoidable in a problem like the k-armed bandit of Example 2.4. But in other cases, it is undesirable. Our general results depend on the problem's information structure in a more refined manner. To preview this, we give another corollary of our main result, which holds for problems with full-information feedback (see Example 2.5 for motivation). In this case, the dependence on the number of actions completely disappears and the bound depends on the variance proxy and the entropy rate. The bound applies to TS and the policy π^{Greedy}, which chooses $A_t \in \arg\max_{a\in\mathcal{A}} \mathbb{E}[R_{t,a} \mid \mathcal{F}_{t-1}]$ in each period t.

Corollary 4.2. *For full information problems, where $O_{t,a} = O_{t,a'}$ for each $a, a' \in \mathcal{A}$, we have*

$$\bar{\Delta}_T(\pi^{\text{Greedy}}) \leq \bar{\Delta}_T(\pi^{\text{TS}}) \leq \sigma\sqrt{2 \cdot \bar{H}_T(A^*)},$$

$$\text{and} \quad \bar{\Delta}_\infty(\pi^{\text{Greedy}}) \leq \bar{\Delta}_\infty(\pi^{\text{TS}}) \leq \sigma\sqrt{2 \cdot \bar{H}_\infty(A^*)}.$$

4.2. Bounds on the entropy rate

Our results highlight that the difficulty arising due to the nonstationarity of the environment is sufficiently characterized by the entropy rate of the optimal action process, denoted by $\bar{H}_T(A^*)$ or $\bar{H}_\infty(A^*)$. We provide some stylized upper bounds on these quantities to aid in their interpretation and to characterize the resulting regret bounds in a comparison with the existing results in the literature.

Bound with the number of switches. The next theorem states that the T-period entropy rate $\bar{H}_T(A^*)$ can be bounded by $\tilde{O}(S_T/T)$ if the optimal action switches at most S_T times up to time T almost surely.

Theorem 4.3. *Suppose there exists $S_T \in \mathbb{N}$ that almost surely bounds the number of switches in the optimal action sequence occurring up to time T:*

$$\sum_{t=1}^{T} \mathbb{I}\{A_t^* \neq A_{t-1}^*\} \leq S_T \quad \textit{almost surely},$$

where we assume that $\mathbb{I}\{A_1^* \neq A_0^*\} = 1$. Then,

$$\bar{H}_T(A^*) \leq \frac{S_T}{T} \cdot \left(1 + \log\left(1 + \frac{T}{S_T}\right) + \log|\mathcal{A}|\right).$$

Combining this result with Corollary 4.1 gives the bound

$$\bar{\Delta}_T(\pi^{\text{TS}}) \leq \tilde{O}\left(\sigma\sqrt{|\mathcal{A}| \cdot \frac{S_T}{T}}\right),$$

which precisely recovers the recent results[4] established for switching bandits in the frequentist's setting (Suk & Kpotufe, 2022; Abbasi-Yadkori et al., 2022). Although other features of the environment may change erratically, a low regret is achievable if the optimal action switches infrequently.

Bound with the effective time horizon. We further refine the above result for the cases where the optimal action process $(A_t^*)_{t\in\mathbb{N}}$ is stationary. With $\tau_{\text{eff}} := 1/\mathbb{P}(A_t^* \neq A_{t-1}^*)$, the optimal action switches only once every τ_{eff} time periods in average.

We interpret τ_{eff} as the problem's "effective time horizon", which captures the average length of time before the identity of the optimal action changes. The effective time horizon τ_{eff} is long when the optimal action changes infrequently, so that, intuitively, a decision-maker could continue to exploit the optimal action for a long time if it were identified, achieving a low regret. The next theorem shows that the entropy rate $\bar{H}_T(A^*)$ is bounded by the inverse of τ_{eff}, regardless of T:

Theorem 4.4. *When the process $(A_t^*)_{t\in\mathbb{N}}$ is stationary,*

$$\bar{H}_T(A^*) \leq \frac{H(A_1^*)}{T} + \frac{1 + \log(\tau_{\text{eff}}) + H(A_t^* | A_t^* \neq A_{t-1}^*)}{\tau_{\text{eff}}}, \tag{4}$$

for every $T \in \mathbb{N}$, where

$$\tau_{\text{eff}} := \frac{1}{\mathbb{P}(A_t^* \neq A_{t-1}^*)}. \tag{5}$$

Combining this result with Corollary 4.1, and the fact that $\frac{H(A_1^*)}{T} \to 0$ as $T \to \infty$, we obtain

$$\bar{\Delta}_\infty(\pi^{\text{TS}}) \leq \tilde{O}\left(\sigma\sqrt{\frac{|\mathcal{A}|}{\tau_{\text{eff}}}}\right),$$

which closely mirrors familiar $O(\sqrt{k/T})$ regret bounds on the average per-period regret in bandit problems with k arms, T periods, and i.i.d rewards (Bubeck & Cesa-Bianchi, 2012), except that the effective time horizon replaces the problem's raw time horizon.

Theorem 4.4 can be seen as a refined version of Theorem 4.3, specialized to the problems with stationary switching processes. Below we give an example where the upper bound

[4] In the earlier literature, the per-period regret bounds appear to have a form of $\tilde{O}(\sqrt{|\mathcal{A}| \cdot N_T/T})$ where N_T represents the number of any distributional changes in the environment. Typically, $S_T \ll N_T$.

in Theorem 4.4 is nearly exact while Theorem 4.3 yields a vacuous result.

Example 4.5 (Piecewise stationary environment)**.** Suppose $(A_t^*)_{t\in\mathbb{N}}$ follows a switching process. With probability $1-\delta$ there is no change in the optimal action, whereas with probability δ there is a change-event and A_t is drawn uniformly from among the other $k-1 \equiv |\mathcal{A}|-1$ arms. Precisely, $(A_t^*)_{t\in\mathbb{N}}$ follows a Markov process with transition dynamics:

$$\mathbb{P}(A_{t+1}^* = a \mid A_t^* = a') = \begin{cases} 1-\delta & \text{if } a = a' \\ \delta/(k-1) & \text{if } a \neq a' \end{cases}$$

for $a, a' \in \mathcal{A}$. Then

$$\begin{aligned} \bar{H}_\infty(A^*) &= (1-\delta)\log\left(\frac{1}{1-\delta}\right) + \delta\log\left(\frac{k-1}{\delta}\right) \\ &= (1-\delta)\log\left(1+\frac{\delta}{1-\delta}\right) + \delta\log\left(\frac{k-1}{\delta}\right) \\ &\approx \delta + \delta\log((k-1)/\delta), \end{aligned}$$

where we used the approximation $\log(1+x) \approx x$. Plugging in $\tau_{\text{eff}} = 1/\delta$ and $H(A_t^* \mid A_t^* \neq A_{t-1}^*) = \log(k-1)$ yields,

$$\bar{H}_\infty(A^*) \approx \frac{1+\log(\tau_{\text{eff}}) + H(A_t^* \mid A_t^* \neq A_{t-1}^*)}{\tau_{\text{eff}}},$$

which matches the upper bound (4).

In terms of the maximal number of switches, we have $S_T = T$ since the optimal action can switch in every single period, although it is very unlikely. Invoking Theorem 4.3 with $S_T = T$ yields $\bar{H}_T(A_{1:T}^*) \lesssim \log k$ which is significantly looser than the bound (4).

Bound with the entropy rate of latent state process. Although it can be illuminating to consider the number of switches or the effective horizon, the entropy rate is a deeper quantity that better captures a problem's intrinsic difficulty. A simple but useful fact is that the entropy rate of the optimal action process cannot exceed that of the environment's state process:

Remark 4.6. Since the optimal action A_t is completely determined by the latent state θ_t, by data processing inequality,

$$\bar{H}_T(A^*) \leq H_T(\theta), \quad \bar{H}_\infty(A^*) \leq \bar{H}_\infty(\theta).$$

These bounds can be useful when the environment's nonstationarity has some temporal structure. The next example illustrates such a situation, in which the entropy rate of the latent state process can be directly quantified.

Example 4.7 (System with seasonality)**.** Consider a system that exhibits a strong intraday seasonality. Specifically, suppose that the system's hourly state (e.g., arrival rate) at time t can be modeled as

$$\theta_t = \xi_{\text{day}(t)} \cdot \mu_{\text{time-of-the-day}(t)} + \epsilon_t,$$

where $(\xi_d)_{d\in\mathbb{N}}$ is a sequence of i.i.d random variables describing the daily random fluctuation, $(\mu_h)_{h\in\{0,\ldots,23\}}$ is a known deterministic sequence describing the intraday pattern, and $(\epsilon_t)_{t\in\mathbb{N}}$ is a sequence of i.i.d random variables describing the hourly random fluctuation. Then we have

$$\bar{H}_\infty(A^*) \leq \bar{H}_\infty(\theta) = \frac{1}{24}H(\xi) + H(\epsilon),$$

regardless of the state-action relationship. Imagine that the variation within the intraday pattern μ is large so that the optimal action changes almost every hour (i.e., $S_T \approx T$ and $\tau_{\text{eff}} \approx 1$). In this case, the bound like above can be easier to compute and more meaningful than the bounds in Theorems 4.3 and 4.4.

4.3. Main result

The corollaries presented earlier are special cases of a general result that we present now. Define the (maximal) information ratio of an algorithm π by,

$$\Gamma(\pi) := \sup_{t\in\mathbb{N}} \underbrace{\frac{\left(\mathbb{E}\left[R_{t,A_t^*} - R_{t,A_t}\right]\right)^2}{I\left(A_t^*; (A_t, O_{t,A_t}) \mid \mathcal{F}_{t-1}\right)}}_{=:\Gamma_t(\pi)},$$

The per-period information ratio $\Gamma_t(\pi)$ was defined by Russo & Van Roy (2016) and presented in this form by Russo & Van Roy (2022). It is the ratio between the square of expected regret and the conditional mutual information between the optimal action and the algorithm's observation. It measures the cost, in terms of the square of expected regret, that the algorithm pays to acquire each bit of information about the optimum.

The next theorem shows that any algorithm's per-period regret is bounded by the square root of the product of its information ratio and the entropy rate of the optimal action sequence. The result has profound consequences, but follows easily by applying elegant properties of information measures.

Theorem 4.8. *Under any algorithm π,*

$$\bar{\Delta}_T(\pi) \leq \sqrt{\Gamma(\pi)\cdot\bar{H}_T(A^*)},$$

$$\textit{and} \quad \bar{\Delta}_\infty(\pi) \leq \sqrt{\Gamma(\pi)\cdot\bar{H}_\infty(A^*)}.$$

Proof. Use the shorthand notation $\Delta_t := \mathbb{E}[R_{t,A_t^*} - R_{t,A_t}]$ for regret, $G_t := I(A_t^*; (A_t, O_{t,A_t})|\mathcal{F}_{t-1})$ for information gain, and $\Gamma_t = \Delta_t^2/G_t$ for the information ratio at period t. Then,

$$\sum_{t=1}^T \Delta_t = \sum_{t=1}^T \sqrt{\Gamma_t}\sqrt{G_t} \leq \sqrt{\sum_{t=1}^T \Gamma_t}\sqrt{\sum_{t=1}^T G_t},$$

by Cauchy-Schwarz, and

$$\sqrt{\sum_{t=1}^{T}\Gamma_t}\sqrt{\sum_{t=1}^{T}G_t} \leq \sqrt{\Gamma(\pi)\cdot T\cdot\sum_{t=1}^{T}G_t}.$$

by definition of $\Gamma(\pi)$. We can further bound the information gain. This uses the chain rule, the data processing inequality, and the fact that entropy bounds mutual information:

$$\begin{aligned}\sum_{t=1}^{T}G_t &= \sum_{t=1}^{T} I(A_t^*;(A_t,O_{t,A_t})|\mathcal{F}_{t-1}) \\ &\leq \sum_{t=1}^{T} I([A_1^*,\ldots,A_T^*];(A_t,O_{t,A_t})|\mathcal{F}_{t-1}) \\ &= I([A_1^*,\ldots,A_T^*];\mathcal{F}_T) \\ &\leq H([A_1^*,\ldots,A_T^*]).\end{aligned}$$

Combining these results, we obtain

$$\begin{aligned}\bar{\Delta}_T(\pi) &\leq \frac{\sqrt{\Gamma(\pi)\cdot T\cdot H([A_1^*,\ldots,H_T^*])}}{T} \\ &= \sqrt{\Gamma(\pi)\cdot \bar{H}_T(A^*)}.\end{aligned}$$

Taking limit yields the bound on regret rate $\bar{\Delta}_\infty(\pi)$. $\square$

Remark 4.9. A careful reading of the proof reveals that it is possible to replace the entropy rate $\bar{H}_T(A^*)$ with the mutual information rate $T^{-1}I([A_1^*,\ldots,A_T^*];\mathcal{F}_T)$.

Remark 4.10. Following Lattimore & Gyorgy (2021), one can generalize the definition of information ratio,

$$\Gamma_\lambda(\pi) := \sup_{t\in\mathbb{N}} \frac{\left(\mathbb{E}\left[R_{t,A_t^*}-R_{t,A_t}\right]\right)^\lambda}{I\left(A_t^*;(A_t,O_{t,A_t})\mid\mathcal{F}_{t-1}\right)},$$

which immediately yields an inequality, $\bar{\Delta}_T(\pi) \leq \left(\Gamma_\lambda(\pi)\bar{H}_T(A^*)\right)^{1/\lambda}$ for any $\lambda \geq 1$.

4.4. Some known bounds on the information ratio

We list some known results about the information ratio. These were originally established for stationary bandit problems but immediately extend to nonstationary settings considered in this paper. Most results apply to Thompson sampling, and essentially all bounds apply to Information-directed sampling, which is designed to minimize the information ratio (Russo & Van Roy, 2018). The first four results were shown by Russo & Van Roy (2016) under Assumption 2.2.

Classical bandits. $\Gamma(\pi^{\mathrm{TS}}) \leq 2\sigma^2|\mathcal{A}|$, for bandit tasks with finite action set (e.g., Example 2.4).

Full information. $\Gamma(\pi^{\mathrm{TS}}) \leq 2\sigma^2$, for problems with full-information feedback (e.g., Example 2.5).

Linear bandits. $\Gamma(\pi^{\mathrm{TS}}) \leq 2\sigma^2 d$, for linear bandits of dimension d (i.e., $\mathcal{A}\subseteq\mathbb{R}^d$, $\Theta\subseteq\mathbb{R}^d$, and $\mathbb{R}[R_{t,a}|\theta_t] = a^\top\theta_t$).

Combinatorial bandits. $\Gamma(\pi^{\mathrm{TS}}) \leq 2\sigma^2\frac{d}{k^2}$, for combinatorial optimization tasks of selecting k items out of d items with semi-bandit feedback (e.g., Example 2.6).

Contextual bandits. See the below for a new result.

Logistic bandits. Dong et al. (2019) consider problems where mean-rewards follow a generalized linear model with logistic link function, and bound the information ratio by the dimension of the parameter vector and a new notion they call the ‘fragility dimension.’

Graph based feedback. With graph based feedback, the decision-maker observes not only the reward of selected arm but also the reward of its neighbors in feedback graph. One can bound the information ratio by the feedback graph’s clique cover number (Liu et al., 2018) or its independence number (Hao et al., 2022).

Sparse linear models. Hao et al. (2021) consider sparse linear bandits and show conditions under which the information ratio of Information-Directed Sampling in Remark 4.10 is bounded by the number of nonzero elements in the parameter vector.

Convex cost functions. Bubeck & Eldan (2016) and Lattimore (2020) study bandit learning problems where the reward function is known to be concave and bound the information ratio by a polynomial function of the dimension of the action space.

4.5. A new bound on the information ratio of contextual bandits

Contextual bandit problems are a special case of our formulation that satisfy the following abstract assumption. Reread Example 2.7 to get intuition.

Assumption 4.11. There is a set $\mathcal{X}$ and integer k such that $\mathcal{A}$ is the set of functions mapping $\mathcal{X}$ to $[k]$. The observation at time t is the tuple $O_t = (X_t, R_t) \in \mathcal{X}\times\mathbb{R}$. Define $i_t := A_t(X_t) \in [k]$. Assume that for each t, $X_{t+1} \perp (A_t, R_t) \mid X_t, \mathcal{F}_{t-1}$. and $R_t \perp A_t \mid (X_t, i_t, \theta_t)$.

Under this assumption, we provide an information ratio bound that depends on the number of arms k. It is a massive improvement over Corollary 4.1, which depends on the number of *decision-rules*.

Lemma 4.12. *Under Assumption 4.11,* $\Gamma(\pi^{\mathrm{TS}}) \leq 2\cdot\sigma^2\cdot k$.

Theorem 4.8 therefore bounds regret in terms of the entropy rate of the optimal decision rule process $(A_t^*)_{t\in\mathbb{N}}$, the number of arms k, and the reward variance proxy σ^2.

Neu et al. (2022) recently highlighted that information-ratio analysis seems not to deal adequately with context, and proposed a substantial modification which considers information gain about model parameters rather than optimal decision-rules. Lemma 4.12 appears to resolve this open question without changing the information ratio itself. Our bounds scale with the entropy of the optimal decision-rule, instead of the entropy of the true model parameter, as in Neu et al. (2022). By the data processing inequality, the former is always smaller. Our proof bounds the per-period information ratio, so it can be used to provide finite time regret bounds for stationary contextual bandit problems. Hao et al. (2022) provide an interesting study of variants of Information-directed sampling in contextual bandits with complex information structure. It is not immediately clear how that work relates to Lemma 4.12 and the information ratio of Thompson sampling.

The next corollary combines the information ratio bound above with the earlier bound of Theorem 4.4. The bound depends on the number of arms, the dimension of the parameter space, and the effective time horizon. No further structural assumptions (e.g., linearity) are needed. An unfortunate feature of the result is that it applies only to parameter vectors that are quantized at scale ϵ. The logarithmic dependence on ϵ is omitted in the $\tilde{O}(\cdot)$ notation, but displayed in the proof. When outcome distributions are smooth in θ_t, we believe this could be removed with careful analysis.

Corollary 4.13. *Under Assumption 4.11, if $\theta_t \in \{-1, -1+\epsilon, \ldots, 1-\epsilon, 1\}^p$ is a discretized p-dimensional vector, and the optimal policy process $(A_t^*)_{t\in\mathbb{N}}$ is stationary, then*

$$\bar{\Delta}_\infty(\pi^{\mathrm{TS}}) \leq \tilde{O}\left(\sigma\sqrt{\frac{p\cdot k}{\tau_{\text{eff}}}}\right).$$

5. Lower Bound

We provide an impossibility result through the next theorem, showing that no algorithm can perform significantly better than the upper bounds provided in the previous section. Our proof is built by modifying well known lower bound examples for stationary bandits.

Theorem 5.1. *Let $k > 1$ and $\tau \geq k$. There exists a nonstationary bandit problem instance $|\mathcal{A}| = k$ and $\tau_{\text{eff}} = \tau$, such that*

$$\inf_\pi \bar{\Delta}_\infty(\pi) \geq C \cdot \sigma\sqrt{\frac{|\mathcal{A}|}{\tau_{\text{eff}}}},$$

where C is a universal constant.

Remark 5.2. For the problem instance constructed in the proof, the entropy rate of optimal action process is $\bar{H}(A^*) \approx \log(|\mathcal{A}|)/\tau_{\text{eff}}$. This implies that the upper bound established in Corollary 4.1 is tight up to logarithmic factors, and so is the one established in Theorem 4.8.

6. Conclusion and Open Questions

We have provided a new information-theoretic analysis of interactive learning in changing environments. The results offer an intriguing measure of the difficulty of learning: the entropy rate of the optimal action process. A strength of the approach is that it applies to nonstationary variants of many of the most important learning problems. Instead of designing algorithms to make the proofs work, most results apply to Thompson sampling (TS), one of the most widely used bandit algorithms.

TS can explore too aggressively in nonstationary learning problems with short effective horizon. Namely, we conjecture that TS fails to achieve the regret bound $\tilde{O}(V^{1/3}T^{2/3})$ in the worst-case "drifting environments" (Besbes et al., 2014), and accordingly, our regret bounds could be too loose to characterize the best achievable performance. To resolve this, one should consider variants of TS that *satisfice*. We believe that one can still attain near-optimal performance by tracking an ('satisficing') action sequence which changes less frequently, and a synthesis of our analysis with the information-ratio analysis of satisficing in Russo & Van Roy (2022) could tighten our bounds.

Predictive Sampling (PS), recently proposed by Liu et al. (2023), would be a notable example aligned with this direction: PS improves upon TS by "deprioritizing acquiring information that quickly loses usefulness". The authors use an information-ratio analysis to bound an algorithm's 'foresight regret' in terms of 'predictive information.' Although it is nontrivial to precisely compare their bounds with ours or others in the literature, one can see that (1) their notion of regret is always no larger than the conventional regret we analyze, so our results can bound their foresight regret but not vice versa, and (2) when specialized to stationary k-armed bandits, our bound has a tighter dependency on the number of arms.[5] We believe that our results are complementary and it is an interesting research direction to synthesize their innovations with our analysis.

Acknowledgements

This research was supported by National R&D Program through the NRF funded by Ministry of Science and ICT (2022R1C1C1013402).

[5]When specialized to stationary k-armed bandits, their result roughly depends on the entropy of the model parameter (p.14) whereas ours depends on the entropy of the optimal action. The former scales as $O(k)$ whereas the later scales as $O(\log(k))$.

More generally, they comment that cumulative predictive information represents "the total new uncertainty that has been injected into the environment thus far" (p.12), which roughly corresponds to the cumulative entropy of environment's state process. The example in Section 1.1 suggests that information or entropy regarding the optimal action process may be more suitable.

References

Abbasi-Yadkori, Y., Gyorgy, A., and Lazic, N. A new look at dynamic regret for non-stationary stochastic bandits. *arXiv preprint arXiv:2201.06532*, 2022.

Audibert, J.-Y., Bubeck, S., and Lugosi, G. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

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

Azevedo, E. M., Deng, A., Montiel Olea, J. L., and Weyl, E. G. Empirical Bayes estimation of treatment effects with many A/B tests: An overview. In *AEA Papers and Proceedings*, volume 109, pp. 43–47, 2019.

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

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Bubeck, S. and Eldan, R. Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, pp. 583–589. PMLR, 2016.

Chakrabarti, D., Kumar, R., Radlinski, F., and Upfal, E. Mortal multi-armed bandits. In *Advances in neural information processing systems 21*, 2008.

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

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Learning to optimize under non-stationarity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087. PMLR, 2019.

Cover, T. Universal portfolios. *Mathematical Finance*, 1(1): 1–29, 1991.

Dong, S., Ma, T., and Van Roy, B. On the performance of Thompson sampling on logistic bandits. In *Conference on Learning Theory*, pp. 1158–1160, 2019.

Foster, D. J., Rakhlin, A., Sekhari, A., and Sridharan, K. On the complexity of adversarial decision making. In *Advances in Neural Information Processing Systems*, 2022.

Garivier, A. and Moulines, E. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

Ghatak, G. A change-detection-based Thompson sampling framework for non-stationary bandits. *IEEE Transactions on Computers*, 70(10):1670–1676, 2020.

Gupta, N., Granmo, O.-C., and Agrawala, A. Thompson sampling for dynamic multi-armed bandits. In *10th International Conference on Machine Learning and Applications*, volume 1, pp. 484–489. IEEE, 2011.

Hao, B., Lattimore, T., and Deng, W. Information directed sampling for sparse linear bandits. In *Advances in Neural Information Processing Systems 34*, pp. 16738–16750, 2021.

Hao, B., Lattimore, T., and Qin, C. Contextual information-directed sampling. In *International Conference on Machine Learning*, pp. 8446–8464. PMLR, 2022.

Jung, Y. H. and Tewari, A. Regret bounds for Thompson sampling in episodic restless bandit problems. In *Advances in Neural Information Processing Systems 32*, 2019.

Kocsis, L. and Szepesvári, C. Discounted UCB. In *2nd PASCAL Challenges Workshop*, 2006.

Lattimore, T. Improved regret for zeroth-order adversarial bandit convex optimisation. *Mathematical Statistics and Learning*, 2(3):311–334, 2020.

Lattimore, T. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. In *Conference on Learning Theory*, pp. 1547–1575. PMLR, 2022.

Lattimore, T. and Gyorgy, A. Mirror descent and the information ratio. In *Conference on Learning Theory*, pp. 2965–2992. PMLR, 2021.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Liu, F., Buccapatnam, S., and Shroff, N. Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Liu, Y., Van Roy, B., and Xu, K. Nonstationary bandit learning via predictive sampling. *arXiv preprint arXiv:2205.01970v5*, 2023.

Mellor, J. and Shapiro, J. Thompson sampling in switching environments with Bayesian online change point detection. In *Artificial Intelligence and Statistics*, pp. 442–450. PMLR, 2013.

Neu, G., Olkhovskaya, J., Papini, M., and Schwartz, L. Lifting the information ratio: An information-theoretic analysis of Thompson sampling for contextual bandits. *arXiv preprint arXiv:2205.13924*, 2022.

Raj, V. and Kalyani, S. Taming non-stationary bandits: A Bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

Russo, D. and Van Roy, B. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(1):1221–1243, 2016.

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

Russo, D. and Van Roy, B. Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*, 47 (4):2815–2839, 2022.

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

Suk, J. and Kpotufe, S. Tracking most significant arm switches in bandits. In *Conference on Learning Theory*, pp. 2160–2182. PMLR, 2022.

Trovo, F., Paladino, S., Restelli, M., and Gatti, N. Sliding-window Thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

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

Wu, Y., Zheng, Z., Zhang, G., Zhang, Z., and Wang, C. Non-stationary a/b tests. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2079–2089, 2022.

A. Proofs

Proof of Theorem 4.3. We count the number of possible optimal action sequence configurations, when there can be at most S_T switches up to time T:

$$\left|\left\{(a_1^*, \ldots, a_T^*) \in \mathcal{A}^T \,\middle|\, \sum_{t=1}^{T} \mathbb{I}\{a_t^* \neq a_{t-1}^*\} \le S_T\right\}\right| \le \binom{(T-1)+(S_T-1)}{S_T-1} \times |\mathcal{A}|^{S_T} \le \binom{T+S_T-1}{S_T} \times |\mathcal{A}|^{S_T},$$

where the first inequality is obtained by bounding the possible number of switching time configurations[6], and the second inequality uses the fact that $\binom{n}{k} \le \binom{n+1}{k+1}$. Note that for any $k \le n \in \mathbb{N}$,

$$\binom{n}{k} = \frac{n \times (n-1) \times \ldots \times (n-k+1)}{k!} \le \frac{n^k}{k!} \le \frac{n^k}{\sqrt{2\pi k}(k/e)^k} \le \frac{n^k}{(k/e)^k} = \left(\frac{en}{k}\right)^k,$$

where the second inequality uses Stirling. Therefore,

$$H(A_{1:T}^*) \le \log\left(\left(\frac{e(T+S_T-1)}{S_T}\right)^{S_T} \times |\mathcal{A}|^{S_T}\right) = S_T \times \left(1 + \log\left(1 + \frac{T-1}{S_T}\right) + \log|\mathcal{A}|\right).$$

By observing $\log\left(1+\frac{T-1}{S_T}\right) \le \log\left(1+\frac{T}{S_T}\right)$, we obtain the desired result.

Proof of Theorem 4.4. Let $Z_t := \mathbb{I}\{A_t^* \neq A_{t-1}^*\}$, an indicator of a "switch". Then, $\tau_{\text{eff}}^{-1} = \mathbb{P}(Z_t = 1)$ and

$$\bar{H}_T(A^*) = \frac{1}{T}\left[H(A_1^*) + H(A_2^*|A_1^*) + \ldots + H(A_T^*|A_{1:(T-1)^*})\right].$$

For $t \ge 1$, we bound each term in this sum as

$$\begin{aligned}
H(A_t^*|A_{1:t-1}^*) &= H(A_t^*|A_{1:t-1}^*) + \underbrace{H(Z_t|A_{1:t-1}^*, A_t^*)}_{=0} \\
&= H((Z_t, A_t^*)|A_{1:t-1}^*) \\
&= H(Z_t|A_{1:t-1}^*) + H(A_t^*|Z_t, A_{1:t-1}^*) \\
&\le H(Z_t) + H(A_t^*|Z_t, A_{1:t-1}^*) \\
&= H(Z_t) + \mathbb{P}(Z_t = 1)H(A_t^*|Z_t = 1, A_{1:t-1}^*) \\
&\quad + \underbrace{\mathbb{P}(Z_t = 0)H(A_t^*|Z_t = 0, A_{1:t-1}^*)}_{=0} \\
&\le H(Z_t) + \mathbb{P}(Z_t = 1)H(A_t^*|Z_t = 1).
\end{aligned}$$

With $\delta := \tau_{\text{eff}}^{-1}$,

$$\begin{aligned}
&H(Z_t) + \mathbb{P}[Z_t = 1]H(A_t^*|Z_t = 1) \\
&= \delta\log(1/\delta) + (1-\delta)\log(1/(1-\delta)) + \delta H(A_t^*|Z_t = 1) \\
&= \delta\log(1/\delta) + (1-\delta)\log(1+\delta/(1-\delta)) + \delta H(A_t^*|Z_t = 1) \\
&\le \delta\log(1/\delta) + \delta + \delta H(A_t^*|Z_t = 1) \\
&= \frac{1}{\tau_{\text{eff}}}\left[\log(\tau_{\text{eff}}) + 1 + H(A_t^*|Z_t = 1)\right].
\end{aligned}$$

[6] One can imagine a two-dimensional grid of size $T \times S_T$, represented with coordinates $((t,s))_{t\in\{1,\ldots,T\}, s\in\{1,\ldots,S_T\}}$. A feasible switching time configuration corresponds to a path from $(1,1)$ to (T, S_T) that consists of $T-1$ rightward moves and $S_T - 1$ upward moves (whenever a path makes an upward move, from (t,s) to $(t, s+1)$, we can mark that a switch occurs at time t). The number of such paths is given by $\binom{(T-1)+(S_T-1)}{S_T-1}$.

Proof of Remark 4.10. Let $\Delta_t := \mathbb{E}[R_{t,A_t^*} - R_{t,A_t}]$, $G_t := I(A_t^*; (A_t, O_{t,A_t})|\mathcal{F}_{t-1})$, and $\Gamma_{\lambda,t} := \Delta_t^{\lambda}/G_t$. Then, $\Gamma_\lambda(\pi) = \sup_{t\in\mathbb{N}} \Gamma_{\lambda,t}$, and we have

$$\begin{aligned}
\bar{\Delta}_T(\pi) &= T^{-1}\sum_{t=1}^{T}\Delta_t \\
&= T^{-1}\sum_{t=1}^{T}\Gamma_{\lambda,t}^{1/\lambda}G_t^{1/\lambda} \\
&\leq \Gamma_\lambda(\pi)^{1/\lambda}\cdot\left(T^{-1}\sum_{t=1}^{T}G_t^{1/\lambda}\right) \\
&\overset{(a)}{\leq} \Gamma_\lambda(\pi)^{1/\lambda}\cdot\left[T^{-1}\left(\sum_{t=1}^{T}G_t\right)^{1/\lambda}\cdot\left(\sum_{t=1}^{T}1\right)^{1-1/\lambda}\right] \\
&= \Gamma_\lambda(\pi)^{1/\lambda}\cdot\left(T^{-1}\sum_{t=1}^{T}G_t\right)^{1/\lambda} \\
&\overset{(b)}{\leq} \Gamma_\lambda(\pi)^{1/\lambda}\cdot\bar{H}_T(A^*)^{1/\lambda},
\end{aligned}$$

where step (a) uses Hölder's inequality, and step (b) uses $T^{-1}\sum_{t=1}^{T}G_t \leq \bar{H}_T(A^*)$.

Proof of Lemma 4.12. Recall the definition, $\Gamma(\pi) := \sup_{t\in\mathbb{N}}\Gamma_t(\pi)$ where

$$\Gamma_t(\pi) = \frac{\left(\mathbb{E}\left[R_{t,A_t^*} - R_{t,A_t}\right]\right)^2}{I\left(A_t^*; (A_t, O_{t,A_t}) \mid \mathcal{F}_{t-1}\right)}.$$

Our goal is to bound the numerator of $\Gamma_t(\pi^{\mathrm{TS}})$ in terms of the denominator.

Let $\mathbb{E}_t[\cdot] := \mathbb{E}[\,\cdot \mid X_t, \mathcal{F}_{t-1}]$ denote the conditional expectation operator which conditions on observations prior to time t AND the context at time t. Similarly, define the probability operation $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid X_t, \mathcal{F}_{t-1})$ accordingly. Define $I_t(\cdot;\cdot)$ to be the function that evaluates mutual information when the base measure is $\mathbb{P}_t$.

The law of iterated expectations states that for any real valued random variable Z, $\mathbb{E}[\mathbb{E}_t[Z]] = \mathbb{E}[Z]$. The definition of conditional mutual information states that for any random variables Z_1, Z_2,

$$\mathbb{E}\left[I_t(Z_1; Z_2)\right] = I(Z_1; Z_2 \mid X_t, \mathcal{F}_{t-1}). \tag{6}$$

Under Assumption 4.11, there exists a function $\mu : \Theta \times [k] \times \mathcal{X} \to \mathbb{R}$ such that

$$\mu(\theta', i, x) = \mathbb{E}\left[R_{t,A_t} \mid \theta_t = \theta', i_t = i, A_t\right]. \tag{7}$$

This specifies expected rewards as a function of the parameter and chosen arm, regardless of the specific decision-rule used.

The definition of Thompson sampling is the probability matching property on decision-rules, $\mathbb{P}\left(A_t^* = a \mid \mathcal{F}_{t-1}\right) = \mathbb{P}\left(A_t^* = a \mid \mathcal{F}_{t-1}\right)$, for each $A \in \mathcal{A}$. It implies the following probability matching property on arms: with $i_t^* := A_t^*(X_t) \in [k]$.

$$\mathbb{P}_t\left(i_t^* = i\right) = \mathbb{P}\left(A_t^*(X_t) = i \mid \mathcal{F}_{t-1}, X_t\right) = \mathbb{P}\left(A_t(X_t) = i \mid \mathcal{F}_{t-1}, X_t\right) = \mathbb{P}_t\left(i_t = i\right),$$

which holds for each $i \in [k]$.

With this setup, repeating the analysis in Proposition 3, or Corollary 1, of Russo & Van Roy (2016) implies, immediately, that

$$\left(\mathbb{E}_t\left[\mu(\theta_t, X_t, i_t^*) - \mu(\theta_t, X_t, i_t)\right]\right)^2 \leq 2\cdot\sigma^2\cdot k\cdot I_t\left(i_t^*; (i_t, R_t)\right). \tag{8}$$

(Conditioned on context, one can repeat the same proof to relate regret to information gain about the optimal arm.) Now, we complete the proof:

$$
\begin{aligned}
\left(\mathbb{E}\left[R_{t,A_t^*} - R_{t,A_t}\right]\right)^2 &\overset{(a)}{=} \left(\mathbb{E}\left[\mu(\theta_t, X_t, i_t^*) - \mu(\theta_t, X_t, i_t)\right]\right)^2 \\
&\overset{(b)}{\leq} \mathbb{E}\left[\left(\mathbb{E}_t\left[\mu(\theta_t, X_t, i_t^*) - \mu(\theta_t, X_t, i_t)\right]\right)^2\right] \\
&\overset{(c)}{\leq} 2 \cdot \sigma^2 \cdot k \cdot \mathbb{E}\left[I_t\left(i_t^*; (i_t, R_t)\right)\right] \\
&\overset{(d)}{=} 2 \cdot \sigma^2 \cdot k \cdot I\left(i_t^*; (i_t, R_t) \mid X_t, \mathcal{F}_{t-1}\right) \\
&\overset{(e)}{\leq} 2 \cdot \sigma^2 \cdot k \cdot I\left(A_t^*; (i_t, R_t) \mid X_t, \mathcal{F}_{t-1}\right) \\
&\overset{(f)}{\leq} 2 \cdot \sigma^2 \cdot k \cdot I\left(A_t^*; (A_t, R_t) \mid X_t, \mathcal{F}_{t-1}\right) \\
&\overset{(g)}{=} 2 \cdot \sigma^2 \cdot k \cdot \left[I\left(A_t^*; (A_t, X_t, R_t) \mid \mathcal{F}_{t-1}\right) - I\left(A_t^*; X_t \mid \mathcal{F}_{t-1}\right)\right] \\
&\overset{(h)}{\leq} 2 \cdot \sigma^2 \cdot k \cdot I\left(A_t^*; (A_t, X_t, R_t) \mid \mathcal{F}_{t-1}\right) \\
&\overset{(i)}{=} 2 \cdot \sigma^2 \cdot k \cdot I\left(A_t^*; (A_t, O_t) \mid \mathcal{F}_{t-1}\right),
\end{aligned}
$$

where step (a) uses (7), step (b) is Jensen's inequality, step (c) applies (8), step (d) is (6), steps (e) and (f) apply the data processing inequality, step (g) uses the chain-rule of mutual information, step (h) uses that mutual information is non-negative, and step (i) simply recalls that $O_t = (X_t, R_t)$.

Proof of Corollary 4.13. If $\theta_t \in \{-1, -1+\epsilon, \ldots, 1-\epsilon, 1\}^p$ is a discretized p dimensional vector, and the optimal policy process $(A_t^*)_{t\in\mathbb{N}}$ is stationary, then, by Theorem 4.4,

$$
\begin{aligned}
\bar{H}_T(A^*) &\leq \frac{1 + \log(\tau_{\text{eff}}) + H(A_t^* | A_t^* \neq A_{t-1}^*)}{\tau_{\text{eff}}} \\
&\leq \frac{1 + \log(\tau_{\text{eff}}) + H(\theta_t)}{\tau_{\text{eff}}} \\
&\leq \frac{1 + \log(\tau_{\text{eff}}) + p\log(2/\epsilon)}{\tau_{\text{eff}}}.
\end{aligned}
$$

Combining this with Theorem 4.8 and the information ratio bound in Lemma 4.12 gives

$$
\bar{\Delta}_T(\pi^{\text{TS}}) \leq \tilde{O}\left(\sigma\sqrt{\frac{p \cdot d}{\tau_{\text{eff}}}}\right).
$$

Proof of Theorem 5.1. We start with a proof sketch. Our proof is built upon a well-known result established for stationary bandits: there exists a stationary (Bayesian) bandit instance such that any algorithm's (Bayesian) cumulative regret is lower bounded by $\Omega(\sqrt{nk})$ where n is the length of time horizon.

More specifically, we set $n = \Theta(\tau_{\text{eff}}) \in \mathbb{N}$ and construct a nonstationary environment by concatenating independent stationary Gaussian bandit instance blocks of length n, i.e., the mean rewards changes periodically every n time steps. In each block (of length n), the best arm has mean reward $\epsilon > 0$ and the other arms has zero mean reward, where the best arm is drawn from k arms uniformly and independently per block. When $\epsilon = \Theta(\sqrt{k/n})$, no algorithm can identify this best arm within n samples, and hence the cumulative regret should increase by $\Omega(n\epsilon)$ per block. Consequently, the per-period regret $\bar{\Delta}_\infty(\pi)$ should be $\Omega(\epsilon) = \Omega(\sqrt{k/n}) = \Omega(\sqrt{k/\tau_{\text{eff}}})$. In our detailed proof, we additionally employ some randomization trick in determination of changepoints in order to ensure that the optimal action sequence $(A_t^*)_{t\in\mathbb{N}}$ is stationary and $\mathbb{P}(A_t^* \neq A_{t-1}^*) = \tau_{\text{eff}}^{-1}$ exactly. Now, we give the formal proof.

Proof. We will consider Gaussian bandit instances throughout the proof. Without loss of generality, we assume $\sigma = 1$ and the noise variances are always one.

We begin by stating a well-known result for the stationary bandits, adopted from Lattimore & Szepesvári (2020, Exercise 15.2): With $\epsilon = (1-1/k)\sqrt{k/n}$, for each $i \in \{1,\dots,k\}$, let mean reward vector $\mu^{(i)} \in \mathbb{R}^k$ satisfy $\mu_a^{(i)} = \epsilon \mathbb{I}\{i=a\}$. It is shown that, when $k>1$ and $n \geq k$, under any algorithm π

$$\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}^{\pi}_{\mu^{(i)}}\left[\sum_{t=1}^{n}(R_{t,i}-R_{t,A_t})\right] \geq \frac{1}{8}\sqrt{nk}, \tag{9}$$

where $\mathbb{E}^{\pi}_{\mu^{(i)}}\left[\sum_{t=1}^{n}(R_{t,i}-R_{t,A_t})\right]$ is the (frequentist's) cumulative regret of π in a k-armed Gaussian bandit instance specified by the time horizon length n and mean reward vector $\mu^{(i)}$ (i.e., the reward distribution of arm a is $\mathcal{N}(\mu_a^{(i)}, 1^2)$). Considering a uniform distribution over $\{\mu^{(1)},\cdots,\mu^{(k)}\}$ as a prior, we can construct a Bayesian K-armed bandit instance of length n such that $\mathbb{E}[\sum_{t=1}^{n}(R_{t,A^*}-R_{t,A_t})] \geq \sqrt{nk}/8$ under any algorithm.

Given $\tau_{\text{eff}} \geq 2$, set $\tilde{\tau} = \frac{k-1}{k}\tau_{\text{eff}}$, $n = \lfloor \tilde{\tau} \rfloor$, and $p = \tilde{\tau} - \lfloor \tilde{\tau} \rfloor$. Let N be the random variable such that equals n with probability p and equals $n+1$ with probability $1-p$, so that $\mathbb{E}[N] = \tilde{\tau}$. We construct a *stationary* renewal process $(T_1, T_2, \dots)$ whose inter-renewal time distribution is given by the distribution of N. That is, $T_{j+1} - T_j \overset{\text{d}}{=} N$ for all $j \in \mathbb{N}$, and T_1 is drawn from the equilibrium distribution of its excess life time, i.e.,

$$\mathbb{P}(T_1 = x) = \begin{cases} 1/\tilde{\tau} & \text{if } x \leq n, \\ p/\tilde{\tau} & \text{if } x = n+1, \\ 0 & \text{if } x > n+1, \end{cases} \qquad \forall x \in \mathbb{N}.$$

Since the process $(T_1, T_2, \dots)$ is a stationary renewal process,

$$\mathbb{P}\left(\text{renewal occurs at } t\right) = \mathbb{P}\left(\exists j, T_j = t\right) = \frac{1}{\mathbb{E}[N]} = \frac{1}{\tilde{\tau}}, \quad \forall t \in \mathbb{N}.$$

We now consider a nonstationary Gaussian bandit instance where the mean reward vector is (re-)drawn from $\{\mu^{(1)},\cdots,\mu^{(k)}\}$ uniformly and independently at times $T_1, T_2, \dots$. As desired, the effective horizon of this bandit instance matches the target τ_{eff}:

$$\mathbb{P}\left(A_t^* \neq A_{t-1}^*\right) = \mathbb{P}\left(\exists j, T_j = t\right) \times \mathbb{P}\left(A_t^* \neq A_{t-1}^* | \exists j, T_j = t\right) = \frac{1}{\tilde{\tau}} \times \left(1 - \frac{1}{k}\right) = \frac{1}{\tau_{\text{eff}}}.$$

Since $T_{j+1} - T_j \geq n$,

$$\mathbb{E}\left[\sum_{t=T_j}^{T_{j+1}-1}(R_{t,A_t^*}-R_{t,A_t})\right] \geq \mathbb{E}\left[\sum_{t=T_j}^{T_j+n-1}(R_{t,A_t^*}-R_{t,A_t})\right] = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\sum_{t=T_j}^{T_j+n-1}(R_{t,A_t^*}-R_{t,A_t})\middle| A_{T_j}^* = i\right] \geq \frac{1}{8}\sqrt{nk},$$

where the last inequality follows from Equation (9). Since there are at least $\lfloor T/(n+1) \rfloor$ renewals until time T, $\mathbb{E}\left[\sum_{t=1}^{T}(R_{t,A_t^*}-R_{t,A_t})\right] \geq \lfloor T/(n+1) \rfloor \sqrt{nk}/8$, and therefore,

$$\bar{\Delta}_\infty(\pi) = \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(R_{t,A_t^*}-R_{t,A_t})\right] \geq \frac{\sqrt{nk}}{8(n+1)}.$$

Since $n+1 \leq 2n$ and $n = \lfloor \frac{k-1}{k}\tau_{\text{eff}} \rfloor \leq \tau_{\text{eff}}$, we have $\bar{\Delta}_\infty(\pi) \geq \frac{1}{16}\sqrt{\frac{k}{\tau_{\text{eff}}}}$. □

B. Numerical Experiment in Detail

We here illustrate the detailed procedure of the numerical experiment conducted in Section 1.1.

Generative model. We say a stochastic process $(X_t)_{t\in\mathbb{N}} \sim \mathcal{GP}(\sigma_X^2, \tau_X)$ if $(X_1, \dots, X_t)$ follows a multivariate normal distribution satisfying

$$\mathbb{E}[X_i] = 0, \quad \text{Cov}(X_i, X_j) = \sigma_X^2 \exp\left(-\frac{1}{2}\left(\frac{i-j}{\tau_X}\right)^2\right),$$

for any $i, j \in [t]$ and any $t \in \mathbb{N}$. Note that this process is stationary, and given horizon T a sample path $(X_1, \ldots, X_T)$ can be generated by randomly drawing a multivariate normal variable from the distribution specified by σ_X^2 and τ_X.

As described in Section 1.1, we consider a nonstionary two-arm Gaussian bandit with unit noise variance:

$$R_{t,a} = \underbrace{\theta_t^{\mathrm{cm}} + \theta_{t,a}^{\mathrm{id}}}_{=:\mu_{t,a}} + \epsilon_{t,a}, \quad \forall a \in \{1, 2\}, t \in \mathbb{N},$$

where $\epsilon_{t,a}$'s are i.i.d. noises $\sim \mathcal{N}(0, 1^2)$, $\left(\theta_t^{\mathrm{cm}}\right)_{t\in\mathbb{N}}$ is the common variation process $\sim \mathcal{GP}(1^2, \tau^{\mathrm{cm}})$, and $\left(\theta_{t,a}^{\mathrm{id}}\right)_{t\in\mathbb{N}}$ is arm a's idiosyncratic variation process $\sim \mathcal{GP}(1^2, \tau^{\mathrm{id}})$.

Note that the optimal action process is completely determined by θ^{id}:

$$A_t^* = \begin{cases} 1 & \text{if } \theta_{t,1}^{\mathrm{id}} \geq \theta_{t,2}^{\mathrm{id}} \\ 2 & \text{if } \theta_{t,1}^{\mathrm{id}} < \theta_{t,2}^{\mathrm{id}} \end{cases},$$

and the optimal action switches more frequently when τ^{id} is small (compare Figure 3 with Figure 1).

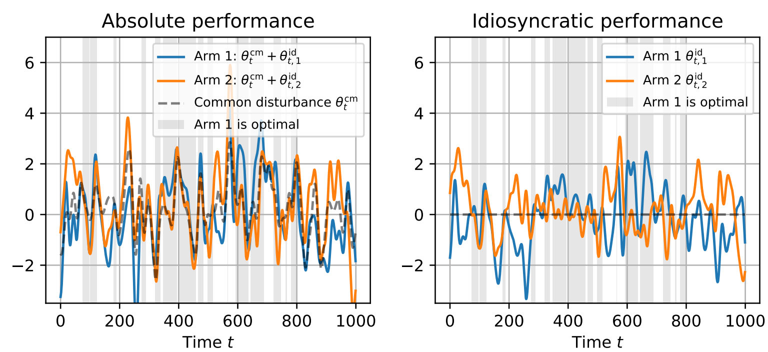

Figure 3. A sample path generated with $\tau^{\mathrm{cm}} = \tau^{\mathrm{id}} = 10$ (cf., Figure 1 was generated with $\tau^{\mathrm{cm}} = 10$ and $\tau^{\mathrm{id}} = 50$).

Consequently, the problem's effective horizon $\tau_{\text{eff}} := 1/\mathbb{P}(A_t^* \neq A_{t-1}^*)$, defined in (5), depends only on τ^{id}. To visualize this relationship, we estimate τ_{eff} using the sample average of the number of switches occurred over $T = 1000$ periods (averaged across 100 sample paths), while varying τ^{id} from 1 to 100. See Figure 4. As expected, τ_{eff} is linear in τ^{id} (more specifically, $\tau_{\text{eff}} \approx 3.0 \times \tau^{\mathrm{id}}$).

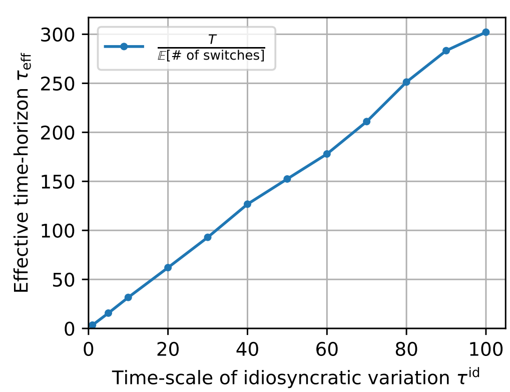

Figure 4. The effective time horizon τ_{eff} as a function of $\tau^{\mathrm{id}} \in \{1, 5, 10, \ldots, 100\}$, estimated from 100 sample paths randomly generated.

Tested bandit algorithms. Given the generative model described above, we evaluate four algorithms – Thompson sampling (TS), Sliding-Window TS (SW-TS; Trovo et al. (2020)), Sliding-Window Upper-Confidence-Bound (SW-UCB; Garivier & Moulines (2008)), and Uniform.

Thompson sampling (TS) is assumed to know the dynamics of latent state processes as well as the exact values of τ^{cm} and τ^{id} (i.e., no model/prior misspecification). More specifically, in each period t, π^{TS} draws a random sample $(\tilde{\theta}^{\mathrm{id}}_{t,1}, \tilde{\theta}^{\mathrm{id}}_{t,2})$ of the latent state $(\theta^{\mathrm{id}}_{t,1}, \theta^{\mathrm{id}}_{t,2})$ from its posterior distribution, and then selects the arm $A_t \leftarrow \arg\max_{a\in\{1,2\}} \tilde{\theta}^{\mathrm{id}}_{t,a}$. Here, the posterior distribution of $(\theta^{\mathrm{id}}_{t,1}, \theta^{\mathrm{id}}_{t,2})$ given the history $\mathcal{F}_{t-1} = (A_1, R_1, \ldots, A_{t-1}, R_{t-1})$ is a multivariate normal distribution that can be computed as follows. Given the past action sequence $(A_1, \ldots, A_{t-1}) \in \mathcal{A}^{t-1}$, the (conditional) distribution of $(R_1, \ldots, R_{t-1}, \theta^{\mathrm{id}}_{t,1}, \theta^{\mathrm{id}}_{t,2})$ is given by

$$\left[\begin{array}{c} R_1 \\ \vdots \\ R_{t-1} \\ \theta^{\mathrm{id}}_{t,1} \\ \theta^{\mathrm{id}}_{t,2} \end{array}\right] \Bigg| (A_1, \ldots, A_{t-1}) \sim \mathcal{N}\left(0 \in \mathbb{R}^{(t-1)+2}, \left[\begin{array}{cc} \Sigma_{t,RR} \in \mathbb{R}^{(t-1)\times(t-1)} & \Sigma_{t,\theta R}^{\top} \in \mathbb{R}^{(t-1)\times 2} \\ \Sigma_{t,\theta R} \in \mathbb{R}^{2\times(t-1)} & \Sigma_{t,\theta\theta} \in \mathbb{R}^{2\times 2} \end{array}\right]\right),$$

where the pairwise covariances are given by $(\Sigma_{t,RR})_{ij} := \mathrm{Cov}(R_i, R_j | A_i, A_j) = \mathbb{I}\{i=j\} \cdot \mathrm{Var}(\epsilon_{i,a}) + \mathrm{Cov}(\theta^{\mathrm{cm}}_i, \theta^{\mathrm{cm}}_j) + \mathbb{I}\{A_i = A_j\} \cdot \mathrm{Cov}(\theta^{\mathrm{id}}_{i,a}, \theta^{\mathrm{id}}_{j,a})$ for $i, j \in [t-1]$, $(\Sigma_{t,\theta R})_{ai} := \mathrm{Cov}(R_i, \theta^{\mathrm{id}}_{t,a} | A_i) = \mathbb{I}\{A_i = a\} \cdot \mathrm{Cov}(\theta^{\mathrm{id}}_{i,a}, \theta^{\mathrm{id}}_{t,a})$ for $a \in [2]$ and $i \in [t-1]$, and $\Sigma_{t,\theta\theta}$ is the identity matrix. Additionally given the reward realizations,

$$\left[\begin{array}{c} \theta^{\mathrm{id}}_{t,1} \\ \theta^{\mathrm{id}}_{t,2} \end{array}\right] \Bigg| \mathcal{F}_{t-1} \sim \mathcal{N}(\hat{\mu}_t \in \mathbb{R}^2, \hat{\Sigma}_t \in \mathbb{R}^{2\times 2}), \quad \text{where} \quad \hat{\mu}_t := \Sigma_{t,\theta R} \Sigma_{t,RR}^{-1} \left[\begin{array}{c} R_1 \\ \vdots \\ R_{t-1} \end{array}\right], \quad \hat{\Sigma}_t := \Sigma_{t,\theta\theta} - \Sigma_{t,\theta R} \Sigma_{t,RR}^{-1} \Sigma_{t,\theta R}^{\top}.$$

Sliding-Window TS is a simple modification of stationary Thompson sampling such that behaves as if the environment is stationary but discards all observations revealed before L periods ago. More specifically, in each period t, $\pi^{\mathrm{SW-TS}}$ with window length L draws a random sample of mean rewards $\tilde{\mu}_{t,a}$ from $\mathcal{N}(\hat{\mu}_{t,a}, \hat{\sigma}^2_{t,a})$ where

$$\hat{\mu}_{t,a} := \frac{\sum_{s=\max\{1,t-L\}}^{t-1} R_s \mathbb{I}\{A_s = a\}}{1 + N_{t,a}}, \quad \hat{\sigma}^2_{t,a} := \frac{1}{1 + N_{t,a}}, \quad N_{t,a} := \sum_{s=\max\{1,t-L\}}^{t-1} \mathbb{I}\{A_s = a\},$$

and then selects the arm $A_t \leftarrow \arg\max_{a\in\mathcal{A}} \tilde{\mu}_{t,a}$. Here, L is a control parameter determining the degree of adaptivity.

Similarly, *Sliding-Window UCB* implements a simple modification of UCB such that computes UCB indices defined as

$$U_{t,a} := \hat{\mu}_{t,a} + \beta \frac{1}{\sqrt{N_{t,a}}}, \quad \hat{\mu}_{t,a} := \frac{\sum_{s=\max\{1,t-L\}}^{t-1} R_s \mathbb{I}\{A_s = a\}}{N_{t,a}}, \quad N_{t,a} := \sum_{s=\max\{1,t-L\}}^{t-1} \mathbb{I}\{A_s = a\},$$

and then selects the arm $A_t \leftarrow \arg\max_{a\in\mathcal{A}} U_{t,a}$. Here, L is a control parameter determining the degree of adaptivity, and β is a control parameter determining the degree of exploration.

Uniform is a naïve benchmark policy that always selects one of two arms uniformly at random. One can easily show that $\bar{\Delta}_T(\pi^{\mathrm{Uniform}}) \approx 0.57$ in our setup, regardless of the choice of τ^{cm} and τ^{id}.

Simulation results. Given a sample path specified by θ, we measure the (pathwise) Cesàro average regret of an action sequence A as

$$\bar{\Delta}_T(A; \theta) := \frac{1}{T} \sum_{t=1}^{T} (\mu_{t,A_t^*} - \mu_{t,A_t}), \quad \text{where} \quad \mu_{t,a} := \theta^{\mathrm{cm}}_t + \theta^{\mathrm{id}}_{t,a}.$$

Given an environment specified by $(\tau^{\mathrm{cm}}, \tau^{\mathrm{id}})$, we estimate the per-period regret of an algorithm π using S sample paths:

$$\hat{\Delta}_T(\pi; \tau^{\mathrm{cm}}, \tau^{\mathrm{id}}) := \frac{1}{S} \sum_{s=1}^{S} \bar{\Delta}_T(A^{\pi,(s)}; \theta^{(s)}),$$

where $\theta^{(s)}$ is the s^{th} sample path (that is shared by all algorithms) and $A^{\pi,(s)}$ is the action sequence taken by π along this sample path. In all experiments, we use $T = 1000$ and $S = 1000$.

We first report convergence of the instantaneous regret in Figure 5: we observe that $\mathbb{E}[\mu_{t,A_t^*} - \mu_{t,A_t}]$ quickly converges to a constant after some initial transient periods, numerically verifying the conjecture made in Remark 2.1. While not reported here, we also observe that the Cesàro average $\bar{\Delta}_T(A;\theta)$ converges to the same limit value as $T \to \infty$ in every sample path, suggesting the ergodicity of the entire system.

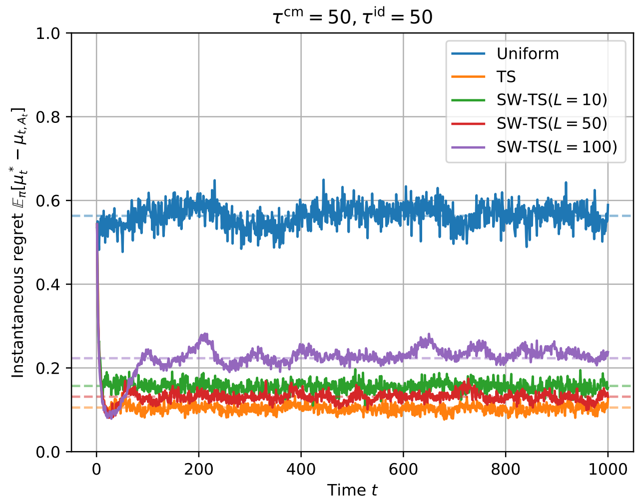

Figure 5. Convergence of instantaneous regret $\mathbb{E}[\mu_{t,A_t^*} - \mu_{t,A_t}]$ in the case of $\tau^{\rm cm} = \tau^{\rm id} = 50$. The solid lines report the instantaneous regret of the algorithms, averaged across $S = 1000$ sample paths, and the dashed horizontal lines represent the estimated per-period regret.

We next examine the effect of $\tau^{\rm id}$ and $\tau^{\rm cm}$ on the performance of algorithms, and provide the detailed simulation results that complement Figure 2 of Section 1.1. While varying $\tau^{\rm id}$ and $\tau^{\rm cm}$, we measure the per-period regret $\hat{\Delta}_T(\pi;\tau^{\rm cm},\tau^{\rm id})$ of algorithms according to the procedure described above. We observe from Figure 6 that for every algorithm its performance is mainly determined by $\tau^{\rm id}$, independent of $\tau^{\rm cm}$, numerically confirming our main claim – the difficulty of problem can be sufficiently characterized by the entropy rate of optimal action sequence, $\bar{H}_\infty(A^*)$, which depends only on $\tau^{\rm id}$. We additionally visualize the upper bound on TS's regret that our analysis predicts[7], assuming that the effective horizon is given by $\tau_{\rm eff} = 3.0 \times \tau^{\rm id}$ (the value 3.0 is obtained from Figure 4). This upper bound seems fairly tight. We also observe that TS performs best across all settings, perhaps because TS exploits the prior knowledge about nonstationarity of the environment, whereas SW-TS or SW-UCB performs well when the window length roughly matches $\tau^{\rm id}$.

[7]Corollary 4.1 and Theorem 4.4 state that $\bar{\Delta}_\infty(\pi^{\rm TS}) \le \sigma\sqrt{2\cdot|\mathcal{A}|\cdot\bar{H}_\infty(A^*)} \le \sigma\sqrt{2\cdot|\mathcal{A}|\cdot\frac{1+\log(\tau_{\rm eff})+H(A_t^*|A_t^*\neq A_{t-1}^*)}{\tau_{\rm eff}}}$, where we have $\sigma = 1$, $\mathcal{A} = 2$, $H(A_t^*|A_t^* \neq A_{t-1}^*) = 0$ in this setting.

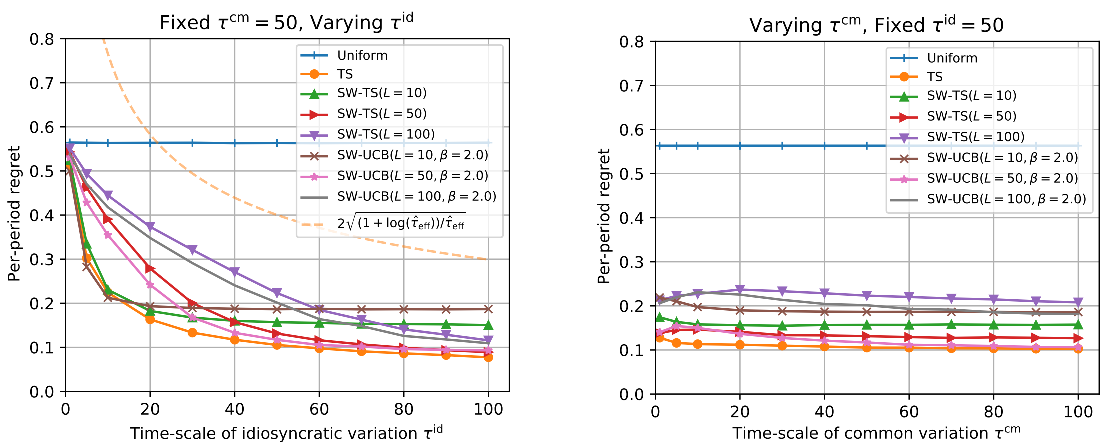

Figure 6. The effect of τ^{id} (left) and τ^{cm} (right) on the performance of algorithms. The dashed line in the left plot represents the upper bound on $\bar{\Delta}_\infty(\pi^{\mathrm{TS}})$, implied by Corollary 4.1 and Theorem 4.4, with estimate $\hat{\tau}_{\mathrm{eff}} := 3.0 \times \tau^{\mathrm{id}}$.