Nonstationary Bandit Learning via Predictive Sampling

Yueyang Liu
Stanford University

Benjamin Van Roy
Stanford University

Kuang Xu
Stanford University

Abstract

Thompson sampling has proven effective across a wide range of stationary bandit environments. However, as we demonstrate in this paper, it can perform poorly when applied to non-stationary environments. We show that such failures are attributed to the fact that, when exploring, the algorithm does not differentiate actions based on how quickly the information acquired loses its usefulness due to nonstationarity. Building upon this insight, we propose predictive sampling, an algorithm that deprioritizes acquiring information that quickly loses usefulness. Theoretical guarantee on the performance of predictive sampling is established through a Bayesian regret bound. We provide versions of predictive sampling for which computations tractably scale to complex bandit environments of practical interest. Through numerical simulation, we demonstrate that predictive sampling outperforms Thompson sampling in all non-stationary environments examined.

1 INTRODUCTION

Thompson sampling (Thompson, 1933) operates by sampling at each timestep statistically plausible mean rewards and selecting the action that maximizes the mean reward among them. Its efficacy in stationary environments has been established through empirical and theoretical analyses (Agrawal and Goyal, 2012; Chapelle and Li, 2011; Russo and Van Roy, 2014).

However, as we demonstrate in this paper, Thompson sampling is not suited for non-stationary environments. In particular, we illustrate through a didactic example that Thompson sampling can suffer near worst case performance in some non-stationary bandit environments (Section 3). A key observation is that Thompson sampling fails to account for nonstationarity of future dynamics when selecting an action, and as such attains unstatisfactory performance.

Nonstationarity should impact how an agent selects its current action to balance between exploration and exploitation. In particular, when investing in acquiring information about an action, an agent should consider the duration over and the extent to which that information will remain useful. Unlike in a stationary environment, where information about an action remains equally useful across future timesteps, the usefulness of information in a non-stationary environment often decay as the reward distribution fluctuates randomly. If the information will quickly become less useful, the agent ought to be less inclined to invest in acquiring it.

This insight motivates our proposal of predictive sampling as an algorithm for non-stationary bandit learning (the design principle is covered in Section 5). We establish a Bayesian regret bound satisfied by predictive sampling. The bound grows linearly in $\sqrt{T\Delta}$, where T is the horizon, and Δ measures the total amount of useful information that can be acquired by an agent. When the environment is stationary, Δ is independent of T, and we recover a bound that is linear in $\sqrt{T}$. When Δ is linear in T, as is typically the case in non-stationary environments that continually produce new information to evolve reward distributions, the bound is linear in T.

We specialize the bound to a class of non-stationary Bernoulli bandit environments that generalize the per-arm abrupt switching model with a constant switching rate (Mellor and Shapiro, 2013). In such environments, upper and lower bounds each grow linearly in T, suggesting that the linear dependence is fundamental. In addition, the upper bound exhibits a graceful dependence on environment parameters, which demonstrates the effectiveness of predictive sampling across a range of such environments.

To quantify the advantage of predictive sampling over Thompson sampling, we develop efficient procedures to execute them in a class of non-stationary Gaussian bandits and conduct experiments. The results demonstrate that predictive sampling outperforms Thompson sampling across

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

all non-stationary bandit environments that we examine.

In addition, as an important step towards designing practical algorithms, we develop variations of Thompson sampling and predictive sampling and develop efficient procedures to execute them in a class of non-stationary logistic bandits. This illustrates how computationally efficient variations of predictive sampling can be designed in an manner analogous to Thompson sampling. A number of existing non-stationary bandit learning algorithms (Ghatak, 2021; Gupta et al., 2011; Mellor and Shapiro, 2013; Raj and Kalyani, 2017; Trovo et al., 2020; Viappiani, 2013) can be viewed as variations of Thompson sampling, and our approach serves to amplify the value of that work.

Primary Contributions The three primary contributions of this paper are: (1) evidence that conveys how and why Thompson sampling is unsuitable for non-stationary bandit environments, (2) a proposal and analysis of predictive sampling for non-stationary bandit learning, and (3) versions of predictive sampling that are computationally tractable and apply to a broad range of complex bandit environments of practical interest.

Structure of The Paper The remainder of the paper proceeds as follows. Section 3 provides an didactic example to illustrate how and why Thompson sampling attains unsatisfactory performance in some non-stationary bandit environments. Section 4 contains the general formulation of a bandit environment. Section 5 formally introduces predictive sampling. Sections 6 and 7 include the regret analysis and numerical experiments. The probabilistic framework, information-theoretic notation and concepts, together with technical proofs are presented in supplementary materials.

2 RELATED WORK

We consider non-stationary bandits as 'restless bandits' introduced by (Whittle, 1988). Many algorithms were since proposed in the non-stationary bandit learning literature. These algorithms focus on heuristics on how to make better inferences about the current mean reward. Popular heuristics for inference include using a fixed-length sliding-window, weighing data by recency, and periodic restarts. Given what is inferred, these algorithms apply action-selection schemes that are designed for stationary bandits, for example, TS, upper-confidence-bound methods (Lai and Robbins, 1985), and exponential-weight algorithms (Auer et al., 2002; Freund and Schapire, 1997). These algorithms, which include discounted TS (Raj and Kalyani, 2017), sliding-window TS (Trovo et al., 2020), reset-aware TS (Viappiani, 2013), Change-Point TS (Mellor and Shapiro, 2013), dynamic TS (Gupta et al., 2011), discounted UCB (Kocsis and Szepesvári, 2006; Garivier and Moulines, 2008), sliding-window UCB (Cheung et al., 2019; Garivier and Moulines, 2008), GLR-klUCB (Besson and Kaufmann, 2019), adapt-EvE (Hartland et al., 2006), and Rexp3 (Besbes et al., 2019), can be considered as variations of stationary bandit learning algorithms.

These algorithms have proven to work much more effectively in non-stationary bandits than stationary bandit learning algorithms. However, by applying the action-selection schemes designed for stationary bandits, these algorithms implicitly assume that future dynamics is stationary, and therefore do not account for the future value of information. We complement this literature by proposing PS, which intelligently accounts for the future when selecting actions.

In terms of algorithmic design, most closely related to ours is (Min et al., 2019), which proposes a family of information relaxation sampling algorithms for stationary bandits. When the actions are independent, PS is equivalent to an extreme point of a class of such algorithms. While Min et al. (2019) focuses on stationary bandits, we propose PS for and analyze its performance in non-stationary bandits.

Our work is also closely related to the body of literature on information-theoretic regret analyses in the stationary bandit learning literature (Bubeck et al., 2015; Dong and Van Roy, 2018; Hao et al., 2022; Lattimore and Szepesvári, 2019; Lu et al., 2021; Russo and Van Roy, 2014, 2016). This area of the literature introduces the notion of an information ratio and bounds the regret of an agent in terms of its information ratio.

Our work adds to this stream of research by extending the information-theoretic framework to non-stationary bandits. Critical to our framework is a new notion of information ratio that is better suited for non-stationary bandits and a concept of predictive information. We bound the regret of an agent in terms of the amount of cumulative predictive information and the agent's information ratio. We also develop useful tools to bound the amount of cumulative predictive information by parameters of the bandits.

3 MOTIVATION

This section provides insight on why Thompson sampling is not suitable for non-stationary bandit learning and why it is crucial to deprioritize information that more quickly loses usefulness. To illustrate these ideas, let us focus on a specific example of a non-stationary bandit environment.

Example 1 (A two-armed non-stationary Bernoulli bandit). *Consider a Bernoulli bandit with a set of two actions $\mathcal{A} = \{1, 2\}$. We use $\theta_{t,a}$ to denote the mean reward associated with action a at timestep t, and $(\theta_t : t \in \mathbb{Z}_+)$ to denote the sequence of mean rewards associated with both actions at all timesteps. We take the first action to be a known benchmark: let $\theta_{t,1} = 1 - \epsilon$ for all $t \in \mathbb{Z}_+$. As for action 2, we let $\theta_{0,2} \sim \text{unif}\{0, 1\}$. The mean reward*

associated with action 2 can vary over time. In particular,

$$\theta_{t+1,2} = \begin{cases} \theta_{t+1,\text{new}} \sim \text{unif}\{0,1\}, & \text{with prob } q = 1-\epsilon, \\ \theta_{t,2}, & \text{otherwise}, \end{cases}$$

where $\epsilon = 1/100$. That is, the mean reward for action 2 is "redrawn" with probability q at each timestep. We use $R_{t+1,a}$ to denote the reward that an agent receives upon executing action a at timestep t. Conditioning on $(\theta_t : t \in \mathbb{Z}_+)$, $R_{t+1,a}$ is Bernoulli distributed with mean $\theta_{t,a}$, independent of the rewards associated with other times or actions.

With this bandit environment, at timestep $t \in \mathbb{Z}_+$, selecting action 1 gives an expected reward of $\theta_{t,1} = 1-\epsilon$, selecting action 2 gives an expected reward of at most $1 - \frac{1}{2}q = \frac{1}{2}(1+\epsilon)$. So an optimal agent would select action 1 at each timestep.

At each timestep $t \in \mathbb{Z}_+$, a Thompson sampling agent samples a mean reward estimate $\hat{\theta}^{\pi_{\text{TS}}}_{t,a}$ for each action $a \in \mathcal{A}$, and executes an action $A^{\pi_{\text{TS}}}_t$ that maximizes the estimate. With this bandit environment, for all $t \in \mathbb{Z}_+$, since the mean reward associated with the first action is known and fixed at $1-\epsilon$, TS samples $\hat{\theta}^{\pi_{\text{TS}}}_{t,1} = 1-\epsilon$. Observe that the mean reward associated with the second action is "redrawn" at each timestep with probability q, and when it is "redrawn", it takes value 1 with probability $\frac{1}{2}$. So the posterior distribution of $\theta_{t,2}$ always has at least $\frac{1}{2}q$ mass on 1. As such, TS samples $\hat{\theta}^{\pi_{\text{TS}}}_{t,2} = 1$ with a probability that is at least $\frac{1}{2}q = \frac{1}{2}(1-\epsilon)$. So a Thompson sampling agent would select action 2 at each timestep with a probability that is at least $\frac{1}{2}(1-\epsilon)$. Here, although action 2 is associated with a smaller expected reward compared to action 1, a Thompson sampling agent selects action 2 with a positive probability because it wishes to acquire more information about this action.

Unfortunately, such information gathering is almost fruitless due to nonstationarity: with $q = 1-\epsilon = \frac{99}{100}$, the mean reward associated with action 2 is likely to be redrawn in each timestep, so any information acquired about it rapidly becomes meaningless. By failing to deprioritize acquiring such information, a Thompson sampling agent would select action 2 more than it should, and as a result collects a reward that is at least $\frac{1}{5}$ less than that collected by an optimal agent in expectation.

Following a similar argument, we establish the following result: there exists a non-stationary Bernoulli bandit environment in which Thompson sampling obtains nearly 0 reward; a different agent (predictive sampling agent, as we show in Section 5.2) obtains almost 1 per timestep. Since rewards in a Bernoulli bandit are binary-valued, this indicates that Thompson sampling performs almost as badly as the worst-possible agent for this environment.

Proposition 1. *For all $\epsilon > 0$, there exists a non-stationary Bernoulli bandit environment and a policy π that selects A^π_t at timestep t such that for all $T \in \mathbb{Z}_{++}$, $\mathbb{E}[\sum_{t=0}^{T-1} R_{t+1,A^\pi_t}] \geq (1-\epsilon)T$ and $\mathbb{E}[\sum_{t=0}^{T-1} R_{t+1,A^{\pi_{\text{TS}}}_t}] \leq \epsilon T$.*

It is worth mentioning that a number of non-stationary bandit learning algorithms (Ghatak, 2021; Gupta et al., 2011; Mellor and Shapiro, 2013; Raj and Kalyani, 2017; Trovo et al., 2020; Viappiani, 2013) use various hueristics to estimate the current mean reward and then take Thompson sampling as a subroutine to select action at each timestep. Our argument on Thompson sampling with Example 1 applies and Proposition 1 extends to each of these algorithms.

4 BANDIT ENVIRONMENTS

This section introduces the general formulation of a bandit environment.

4.1 Bandit Environments

We first formalize the concept of a bandit environment, in which all random quantities are defined with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 1 (bandit environment). *Let $\mathcal{A}$ be a finite set, and $(R_{t+1} : t \in \mathbb{Z})$ be a random sequence of vectors in $\mathbb{R}^{|\mathcal{A}|}$. The sequence $(R_{t+1} : t \in \mathbb{Z})$ is a bandit environment with a finite set $\mathcal{A}$ of actions if there exists a random sequence $(P_{t,a} : t \in \mathbb{Z})$ of probability distributions over $\mathbb{R}$ for each $a \in \mathcal{A}$, such that for all $t \in \mathbb{Z}$ and $a \in \mathcal{A}$,*

$$\mathbb{P}(R_{t+1,a} \in \cdot | P_{-\infty:\infty}, R_{\neg(t+1)}, R_{t+1,\neg a}) = P_{t,a},$$

where $R_{\neg(t+1)}$ denotes the sequence of all reward vectors excluding R_{t+1}, $R_{t+1,\neg a}$ denotes the vector consisting of all components of R_{t+1} except for the ath, P_t denotes the collection of $P_{t,a}$ for $a \in \mathcal{A}$, and $P_{-\infty:\infty}$ denotes the collection of P_t for $t \in \mathbb{Z}$; we say that the sequence $P_{-\infty:\infty}$ generates the bandit environment $(R_{t+1} : t \in \mathbb{Z})$.

In a bandit environment $(R_{t+1} : t \in \mathbb{Z})$, each $R_{t+1,a}$ represents the reward that will be realized if an agent executes action a at time t. By definition, each $R_{t+1,a}$ can be viewed as sampled from $P_{t,a}$, independently from other rewards associated with other times or actions. While components of R_{t+1} are independent conditioning on P_t, they are not necessarily independent unconditionally. We assume that each reward has a well-defined and finite expectation.

It is worth noting that the reward distribution sequence $P_{-\infty:\infty}$ may not be unique for a bandit environment; multiple such sequences may generate the same environment. We say that a bandit environment is *stationary* if there exists a sequence $P_{-\infty:\infty}$, such that $P_t = P_0$ for all $t \in \mathbb{Z}$, that generates the environment; we also say that P_0 generates the environment. A bandit environment is called *non-stationary* if it is not stationary.

4.2 Policy

Let $\mathcal{H}$ denote the set of all sequences of a finite number of action-reward pairs. We refer to the elements of $\mathcal{H}$ as *histories*. Below we provide a formal definition of a *policy*.

Definition 2 (policy). *A policy π is a function that maps a history in $\mathcal{H}$ to a probability distribution over $\mathcal{A}$.*

So a policy π assigns, for each realization of history $h \in \mathcal{H}$, a probability $\pi(a|h)$ of choosing an action a for all $a \in \mathcal{A}$. Let H_0^π be the empty history and for each time $t \in \mathbb{Z}_{++}$ let $H_t^\pi = (A_0^\pi, R_{1,A_0^\pi}, \ldots, A_{t-1}^\pi, R_{t,A_{t-1}^\pi})$, where $\mathbb{P}(A_t^\pi \in \cdot|H_t^\pi) = \pi(\cdot|H_t^\pi)$. Then H_t^π represents the history generated as an agent executes policy π by sampling each action A_t^π from $\pi(\cdot|H_t^\pi)$ and receives the reward R_{t+1,A_t^π}.

Much of the work presented in this paper studies an agent that executes a specific policy, i.e., predictive sampling. When it is clear from context, we suppress superscripts that indicate this. For example, we use H_t for the history generated as an agent executes predictive sampling, and A_t for the action.

5 PREDICTIVE SAMPLING

This section introduces predictive sampling, which is designed around the insight that an agent should ideally deprioritize acquiring information that quickly loses its usefulness.

5.1 The Algorithm

We first introduce the notion of a learning target, which is central to algorithm design in bandit learning. Many such algorithms are designed to trade off between information acquisition and immediate reward optimization. In such contexts, a learning target is a random variable about which an agent aims to acquire information. With a stationary bandit, a natural learning target is the reward distribution, and many agents trade off between immediate reward and information about this distribution.

However, when the environment is non-stationary, taking the reward distribution as the learning target does not deprioritize information that is losing its usefulness. The algorithm design principle of using $R_{t+2:\infty}$ as the learning target addresses this. To see why, observe that the agent trades off between acquiring more information about $R_{t+2:\infty}$ and optimizing the current reward. Therefore, if selecting an action offers information that more quickly loses usefulness, then selecting this action offers less information about $R_{t+2:\infty}$ and is thus deprioritized.

Based on this design principle, we propose predictive sampling (PS), which differs from Thompson sampling in using $R_{t+2:\infty}$ instead of the reward distribution as the learning target. As is known, a Thompson sampling agent samples a statistically plausible model of the reward distribution from its posterior at each timestep t and acts optimally by pretending that the sample is the true model. Aiming for a different learning target of $R_{t+2:\infty}$, a predictive sampling agent instead samples a statistically plausible sequence of future rewards $\hat{R}_{t+2:\infty}^{(t)}$ at each timestep t, and acts optimally by pretending that $\hat{R}_{t+2:\infty}^{(t)}$ is the true sequence of future rewards.

In characterizing the predictive sampling algorithm, we introduce the following change-of-measure notation. Consider random variables X and Y and a conditional probability $\mathbb{P}(Y \in \cdot|X = x)$ for all x in the image of X. Let $f(x) \equiv \mathbb{P}(Y \in \cdot|X = x)$. Given a random variable Z with the same image as X, $f(Z)$ is a random variable. We use the notation $\mathbb{P}(Y \in \cdot|X \leftarrow Z)$ for $f(Z)$.

Algorithm 5 formally introduces predictive sampling. At each timestep t, a predictive sampling agent:

1. (Step 2) samples an infinite sequence of future rewards $\hat{R}_{t+2:\infty}^{(t)}$ from its posterior,
2. (Step 3) derives expected mean rewards $\hat{\theta}_t$ by pretending that $\hat{R}_{t+2:\infty}^{(t)}$ will indeed be realized, and
3. (Step 4) selects the action that maximizes $\hat{\theta}_{t,a}$.

For ease of comparison, we also present Thompson sampling in Algorithm 2.

While Steps 2 and 3 of Algorithm 5 seem intractable, Lemma 1 below suggests that these steps are equivalent to sampling $\hat{\theta}_t$ from $\mathbb{P}(\mathbb{E}[R_{t+1}|H_t, R_{t+2:\infty}] \in \cdot|H_t)$. This gives an alternative representation of the algorithm that can be computationally tractable, as we will show.

Lemma 1. *For all $t \in \mathbb{Z}_+$, $\mathbb{P}(\hat{\theta}_t \in \cdot|H_t) = \mathbb{P}(\mathbb{E}[R_{t+1}|H_t, R_{t+2:\infty}] \in \cdot|H_t)$.*

Algorithm 1: predictive sampling (PS)

for $t = 0, 1, \ldots, T-1$ **do**
- **sample**: $\hat{R}_{t+2:\infty}^{(t)} \sim \mathbb{P}(R_{t+2:\infty} \in \cdot|H_t)$
- **estimate**: $\hat{\theta}_t = \mathbb{E}[R_{t+1}|H_t, R_{t+2:\infty} \leftarrow \hat{R}_{t+2:\infty}^{(t)}]$
- **select**: $A_t \in \arg\max_{a \in \mathcal{A}} \hat{\theta}_{t,a}$
- **observe**: R_{t+1,A_t}

Algorithm 2: Thompson sampling (TS)

for $t = 0, 1, \ldots, T-1$ **do**
- **sample**: $\hat{P}_t^{\pi_{\text{TS}}} \sim \mathbb{P}(P_t^{\pi_{\text{TS}}} \in \cdot|H_t^{\pi_{TS}})$
- **estimate**: $\hat{\theta}_t^{\pi_{\text{TS}}} = \mathbb{E}[R_{t+1}|P_t^{\pi_{\text{TS}}} \leftarrow \hat{P}_t^{\pi_{\text{TS}}}]$
- **select**: $A_t^{\pi_{\text{TS}}} \in \arg\max_{a \in \mathcal{A}} \hat{\theta}_{t,a}^{\pi_{\text{TS}}}$
- **observe**: $R_{t+1,A_t^{\pi_{\text{TS}}}}$

It is helpful to first consider predictive sampling's application to stationary bandit environments. Theorem 1 shows

that in such contexts, predictive sampling executes the same policy as Thompson sampling when the Thompson sampling agent takes $P_t^{\pi_{\mathrm{TS}}} = P$ in its execution (see Algorithm 2), where P generates the stationary bandit.

Theorem 1. *Let $(R_{t+1} : t \in \mathbb{Z})$ be a stationary bandit environment generated by a distribution P. A predictive sampling agent and a Thompson sampling agent that takes $P_t^{\pi_{\mathrm{TS}}} = P$ for all $t \in \mathbb{Z}_+$ execute the same policy.*

In light of this equivalence relation, it is clear that predictive sampling is suited to the range of stationary bandit environments for which Thompson sampling succeeds. Such environments have been the subject of much research in the bandit learning literature (Lattimore and Szepesvári, 2020; Russo et al., 2018). Moreover, recall that multiple sequences $P_{-\infty:\infty}$ may generate the same bandit environment, this equivalence relation also reveals that predictive sampling is more robust compared to Thompson sampling because its execution does not depend on parametrization.

5.2 Back to Example 1

Now that we have introduced predictive sampling, below we apply it to Example 1 to illustrate how predictive sampling achieves optimal behavior of selecting action 1 at each timetsep in this bandit environment.

First, observe that $\mathbb{E}[R_{t+1,2}|H_t, R_{t+2:\infty}] \leq 1 - \frac{1}{2}q^2 < \frac{119}{200}$ a.s. Thus, by Lemma 1, a predictive sampling agent samples $\hat{\theta}_{t,1} = \frac{99}{100}$ and $\hat{\theta}_{t,2} < \frac{119}{200}$ following the alternative representation of Steps 2 and 3, and selects action 1 following Step 4 of Algorithm 5 at each timestep t.

Besides attaining the optimal behavior in this example, predictive sampling in addition performs almost as well as the best possible agent for the environments we constructed in Proposition 1. Below we formally establish the result. The results suggests that by suitably deprioritizing information, predictive sampling succeeds in those environments where we observe Thompson sampling fails.

Proposition 2. *For all $\epsilon > 0$, there exists a non-stationary Bernoulli bandit environment such that for all $T \in \mathbb{Z}_{++}$, $\mathbb{E}[\sum_{t=0}^{T-1} R_{t+1,A_t}] \geq (1-\epsilon)T$ and $\mathbb{E}[\sum_{t=0}^{T-1} R_{t+1,A_t^{\pi_{\mathrm{TS}}}}] \leq \epsilon T$.*

5.3 Examples and Variations of TS and PS

We present computationally tractable examples and variations of Thompson sampling and predictive sampling. For general environments, variations of the algorithms can be designed and efficient procedures to execute them can be constructed through a finite-sample approximation, which we will discuss, and hidden Markov models (HMM). Here, we focus on a class of Gaussian bandits to characterize the advantage of predictive sampling over Thompson sampling and a class of logistic bandits as an important step towards designing practical algorithms.

We first introduce the bandit environments that we consider. We refer to them as the AR(1) bandits and the AR(1) logistic bandits because each of them is determined by a sequence $(\alpha_t : t \in \mathbb{Z}_+)$ and that each $(\alpha_{t,a} : t \in \mathbb{Z}_+)$ independently transitions according to a first-order autoregressive (AR(1)) process. non-stationary bandit environments that are similar to the AR(1) bandits have been considered in (Gupta et al., 2011) and (Slivkins and Upfal, 2008).

Example 2 (AR(1) bandit). *In an AR(1) bandit, each reward distribution $P_{t,a}$ is Gaussian with a random mean $\theta_{t,a} = \alpha_{t,a} = \mathbb{E}[R_{t+1,a}|P_t]$ and a deterministic variance σ_a^2. Each realized reward can be interpreted as a sum $R_{t+1,a} = \theta_{t,a} + Z_{t+1,a}$, where $Z_{t+1,a}$ is independent zero-mean noise with deterministic variance σ_a^2. The variable $\alpha_{t,a}$ changes over time, evolving according to*

$$\alpha_{t+1,a} = (1-\gamma_a)c_a + \gamma_a \alpha_{t,a} + W_{t+1,a},$$

for each action $a \in \mathcal{A}$. The coefficients c_a and γ_a are deterministic, and each takes value in $\mathbb{R}$ and $[0,1]$, respectively; $W_{t+1,a}$ is independent zero-mean Gaussian noise with deterministic variance δ_a^2, where $\delta_a \in \mathbb{R}_+$. When $\gamma_a = 1$, we require that $\delta_a = 0$. We assume that the sequence $(\alpha_{t,a} : t \in \mathbb{Z}_+)$ is in steady-state: when $\gamma_a < 1$, this steady-state distribution is $\mathcal{N}(c_a, \delta_a^2/(1-\gamma_a^2))$.

Example 3 (AR(1) logistic bandit). *In an AR(1) logistic bandit, each reward distribution $P_{t,a} = \mathbb{P}(R_{t+1,a} \in \cdot|P_t)$ is Bernoulli. The mean reward $\mathbb{E}[R_{t+1,a}|P_t] = \frac{\exp(\alpha_t^\top \phi_a)}{1+\exp(\alpha_t^\top \phi_a)}$, where $\alpha_t \in \mathbb{R}^d$ with a known $d \in \mathbb{Z}_{++}$ and $\phi_a \in \mathbb{R}^d$ denotes a known feature vector associated with action $a \in \mathcal{A}$. The variable $\alpha_{t,a}$ is defined exactly as in Example 2, transitioning following an AR(1) process.*

While we have made no explicit assumptions on the prior knowledge of the agents, with AR(1) bandits, AR(1) logistic bandits, or modulated Bernoulli banadits which we will introduce, the agents could have already learned the environment parameters in the long run. Since we focus on the asymptotic behavior and performance, without loss of generality, we assume that the environment parameters, c_a, γ_a, δ_a, q_a, $\mathbb{P}(\theta_{0,a} \in \cdot)$, σ_a for all $a \in \mathcal{A}$, are known to both agents a priori. Such assumptions similarly appear in (Slivkins and Upfal, 2008) and (Mellor and Shapiro, 2013) and techniques to learn the environment parameters have been discussed in (Wilson et al., 2010; Turner et al., 2009).

5.3.1 Examples of TS and PS in AR(1) Bandits

We first focus on Thompson sampling. Below we introduce a lemma, which gives an alternative representation of Thompson sampling: Steps 2 and 3 of Algorithm 2 are equivalent to sampling $\hat{\theta}_t^{\pi_{\mathrm{TS}}}$ from $\mathbb{P}(\mathbb{E}[R_{t+1}|P_t^{\pi_{\mathrm{TS}}}] \in \cdot|H_t^{\pi_{\mathrm{TS}}})$.

Lemma 2. *For all $t \in \mathbb{Z}_+$, $\mathbb{P}(\hat{\theta}_t^{\pi_{\mathrm{TS}}} \in \cdot|H_t^{\pi_{\mathrm{TS}}}) = \mathbb{P}(\mathbb{E}[R_{t+1}|P_t^{\pi_{\mathrm{PS}}}] \in \cdot|H_t^{\pi_{\mathrm{TS}}})$.*

Suppose that a Thompson sampling agent maintains $\hat{P}^{\pi^{\text{TS}}}_{-\infty:\infty} = P_{-\infty:\infty}$. By Lemma 2, the Thompson sampling agent samples $\hat{\theta}^{\pi^{\text{TS}}}_t$ from $\mathbb{P}(\theta_t \in \cdot | H^{\pi^{\text{TS}}}_t)$, and selects an action that maximizes $\hat{\theta}^{\pi^{\text{TS}}}_{t,a}$.

Recall that in an AR(1) bandit, $\mathbb{P}(\theta_0 \in \cdot)$ is Gaussian distributed. We use μ_0 and Σ_0 to denote its mean and covariance. When action a is selected at timestep t, the agent observes $R_{t+1,a} \sim \mathcal{N}(\theta_{t,a}, \sigma_a^2)$, where σ_a is deterministic and known. So $\mathbb{P}(\theta_t \in \cdot | H^{\pi^{\text{TS}}}_t)$ is Gaussian. We use $\mu^{\pi^{\text{TS}}}_t$ and $\Sigma^{\pi^{\text{TS}}}_t$ to denote its mean and covariance. Algorithm 3 provides an example of Thompson sampling in an AR(1) bandit, where Step 5 can be derived using Kalman filter.

As for predictive sampling, first observe that $\mathbb{P}(\theta_t \in \cdot | H_t)$ is Gaussian. We use μ_t and Σ_t to denote its mean and variance. Since the actions are independent, Σ_t is diagonal, and we use $\sigma^2_{t,a}$ to denote the a-th entry along its diagonal.

Recall that Lemma 2 suggests that predictive sampling samples $\hat{\theta}_t$ from $\mathbb{P}(\hat{\theta}_t | H_t) = \mathbb{P}(\mathbb{E}[R_{t+1} | H_t, R_{t+2:\infty}] \in \cdot | H_t)$. Proposition 3 shows that this posterior is Gaussian with mean $\tilde{\mu}_t$ and diagonal covariance matrix $\tilde{\Sigma}_t$ (we use $\tilde{\sigma}_{t,a}$ to denote the a-th entry along its diagonal); $\tilde{\mu}_t$ and $\tilde{\Sigma}_t$ can be derived from μ_t, Σ_t and latent variables of the AR(1) bandit. Algorithm 3 provides an example of predictive sampling in an AR(1) bandit, where Step 5 can be derived by updating μ_t and Σ_t using Kalman filter, and updating $\tilde{\mu}_t$ and $\tilde{\Sigma}_t$ based on μ_t and Σ_t using Proposition 3.

Algorithm 3: PS in an AR(1) bandit

1 **for** $t = 0, 1, \ldots, T-1$ **do**
2 **sample**: $\hat{\theta}_t \sim \mathcal{N}(\tilde{\mu}_t, \tilde{\Sigma}_t)$
3 **execute**: $A_t \in \arg\max_{a \in \mathcal{A}} \hat{\theta}_{t,a}$
4 **observe**: R_{t+1,A_t}
5 **update**: $\tilde{\mu}_{t+1} \leftarrow \mathbb{E}[\hat{\theta}_{t+1} | H_{t+1}], \tilde{\Sigma}_{t+1} \leftarrow \mathbb{V}(\hat{\theta}_{t+1} | H_{t+1})$

Algorithm 4: TS in an AR(1) bandit

1 **for** $t = 0, 1, \ldots, T-1$ **do**
2 **sample**: $\hat{\theta}^{\pi^{\text{TS}}}_t \sim \mathcal{N}(\mu^{\pi^{\text{TS}}}_t, \Sigma^{\pi^{\text{TS}}}_t)$
3 **execute**: $A^{\pi^{\text{TS}}}_t \in \arg\max_{a \in \mathcal{A}} \hat{\theta}^{\pi^{\text{TS}}}_{t,a}$
4 **observe**: $R_{t+1,A^{\pi^{\text{TS}}}_t}$
5 **update**: $\mu^{\pi^{\text{TS}}}_{t+1} \leftarrow \mathbb{E}[\theta_{t+1} | H^{\pi^{\text{TS}}}_{t+1}], \Sigma^{\pi^{\text{TS}}}_{t+1} \leftarrow \mathbb{V}(\theta_{t+1} | H^{\pi^{\text{TS}}}_{t+1})$

Proposition 3. *In an AR(1) bandit, for all $t \in \mathbb{Z}_+$ and $a \in \mathcal{A}$, conditioned on H_t, $\hat{\theta}_{t,a}$ is Gaussian with mean and variance $\tilde{\mu}_{t,a} = \mu_{t,a}$ and $\tilde{\sigma}^2_{t,a} = \frac{\gamma_a^2 \sigma^4_{t,a}}{\gamma_a^2 \sigma^2_{t,a} + x^*_a}$, where $x^*_a = \frac{1}{2}(\delta_a^2 + \sigma_a^2 - \gamma_a^2 \sigma_a^2 + \sqrt{(\delta_a^2 + \sigma_a^2 - \gamma_a^2 \sigma_a^2)^2 + 4\gamma_a^2 \delta_a^2 \sigma_a^2})$.*

5.3.2 Variations in AR(1) Logistic Bandits

We introduce two techniques, which we call incremental Laplace approximation and Gaussian imagination, respectively, in designing computationally tractable variations of Thompson sampling and predictive sampling in AR(1) logistic bandits.

Incremental Laplace Approximation Laplace approximation (Laplace, 1986) is a standard practice in the literature and has been popular with stationary logistic bandits. The key idea is to approximate the posterior of a variable using a Gaussian distribution centered at the maximum a posteriori (MAP) of the variable, with a variance that is equal to the inverse of the Hessian of the log-posterior.

However, applying the method to approximate $\mathbb{P}(\alpha_t | H^\pi_t)$ is computationally onerous due to nonstationarity. To reduce the computational complexity, we propose what we call incremental Laplace approximation: the practice of approximating the posterior incrementally at each timestep using Laplace approximation. We demonstrate in supplementary materials that incremental Laplace approximation is comparable with the standard Laplace approximation in stationary logistic bandits, but can be efficiently carried out in non-stationary ones.

Gaussian Imagination In executing predictive sampling, an agent needs to sample at each timestep t an infinite number of rewards $\hat{R}^{(t)}_{t+2:\infty}$ and derive an conditional expectation $\mathbb{E}[R_{t+1} | H_t, R_{t+2:\infty} \leftarrow \hat{R}^{(t)}_{t+2:\infty}]$. This derivation is usually intractable. The agent can derive an approximation pretending that the rewards are Gaussian. This practice is called Gaussian imagination (Liu et al., 2022).

Finite-sample Approximation An additional technique that is useful in constructing computationally tractable variations of predictive sampling is to sample a finite number of rewards $\hat{R}^{(t)}_{t+2:t+n+1}$, where $n \in \mathbb{Z}_{++}$ instead of $\hat{R}^{(t)}_{t+2:\infty}$ and proceed with the inference.

Variations of Thompson sampling and predictive sampling can be constructed using the aforementioned techniques. We provide detailed steps in the supplementary materials.

6 REGRET ANALYSIS

This section provides theoretical guarantees on the performance of predictive sampling.

6.1 Performance and Regret

To measure the performance of an agent, it can be helpful to consider benchmarking it against an oracle who acts optimally with respect to the full knowledge of the learning target; we define the regret as the gap between the expected rewards accumulated by the agent and that accumulated by the oracle. Indeed, when the environment is stationary, a natural learning target is the reward distribution P. As such, the regret is defined as $\sum_{t=0}^{T-1} \mathbb{E}[R_* - R_{t+1,A^\pi_t}]$, with

$R_* = \max_{a\in\mathcal{A}} \mathbb{E}[R_{t+1,a}|P]$, where $\mathbb{E}[R_{t+1}|P]$ is the mean reward vector which does not depend on t.

In a non-stationary environment, as we have discussed in Section 3, an agent should aim for a different learning target of $R_{t+2:\infty}$ at each timestep t. To measure the performance of such an agent, we define the regret associated with policy π over T timesteps below:

$$\mathrm{Regret}(T;\pi) = \sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,*} - R_{t+1,A_t^\pi}\right], \quad (1)$$

where $R_{t+1,*} = \max_{a\in\mathcal{A}} \mathbb{E}[R_{t+1,a}|R_{t+2:\infty}]$ is the reward accumulated by an oracle that acts optimally with respect ot the full knowledge of $R_{t+2:\infty}$ at timestep t. Since much of the work presented in this paper studies a predictive sampling agent, we use $\mathrm{Regret}(T)$ to denote its regret.

It is worth highlighting that in stationary environments, the two oracles execute the same policy, because $\mathbb{E}[R_{t+1}|P] = \mathbb{E}[R_{t+1}|R_{t+2:\infty}]$. Thus, the regret (1) extends the same notion in stationary bandits.

We establish in supplementary materials that the regret incurred by any agent is nonnegative in all bandit environments that we consider. Indeed, we establish the nonnegativity under mild conditions, which hold in all stationary environments and a broad class of non-stationary ones.

6.2 General Regret Analysis

Oftentimes, an agent is designed to trade off between acquiring information and optimizing the current reward. Whenever an agent incurs regret, the agent learns something useful. Based on this observation, one can bound the regret through bounding the total amount of useful information that can be acquired by an agent, and how efficient the agent is in acquiring information per unit cost of regret. As such, Bayesian analyses in stationary bandit learning (Bubeck et al., 2015; Lattimore and Szepesvári, 2019; Lu et al., 2021; Russo and Van Roy, 2016) bound the regret.

In a non-stationary bandit environment, we introduce a metric that measures the total amount of useful information, i.e., information about the learning target $R_{t+2:\infty}$. Let $\mathcal{E}_t = \mathbb{P}(R_{t+2:\infty} \in \cdot|R_{-\infty:t+1})$ for all $t \in \mathbb{Z}$; $\mathcal{E}_t$ represents what an agent can learn from past rewards about $R_{t+2:\infty}$. Let $\Delta_0 = \mathbb{I}(R_{2:\infty};\mathcal{E}_0)$ and $\Delta_t = \sup_\pi\{\mathbb{I}(R_{t+2:\infty};\mathcal{E}_t|H_t^\pi) - \mathbb{I}(R_{t+1:\infty};\mathcal{E}_{t-1}|H_t^\pi)\}$ for all $t \in \mathbb{Z}_{++}$. With this definition, Δ_0 measures the amount of useful information in the environment at timestep 0; Δ_t measures the additional amount of useful information that is relevant for learning that arrives at the environment at timestep t. We then use $\Delta = \sum_{t=0}^{T-1}\Delta_t$ to measure the total amount of useful information that can be acquired by an agent.

As a sanity check, we have the following proposition, which shows that when the environment is stationary, $\Delta_t = 0$ for all $t \in \mathbb{Z}_{++}$, and that $\Delta = \Delta_0$.

Proposition 4. *In a stationary environment, $\Delta_t = 0$ for all $t \in \mathbb{Z}_{++}$.*

We next introduce a notion of information ratio. Similar notions have been introduced in stationary bandit learning analyses to measure how ineffective an agent is in acquiring information. The one we introduce differs from these in its choice of learning target $R_{t+2:\infty}$. That is, our information ratio measures the trade-off between a single-timestep regret, defined against an oracle that acts optimally with respect to $R_{t+2:\infty}$, and the information obtained about $R_{t+2:\infty}$: let

$$\Gamma_t^\pi = \frac{\mathbb{E}\left[R_{t+1,*} - R_{t+1,A_t^\pi}\right]^2}{\mathbb{I}\left(R_{t+2:\infty}; A_t^\pi, R_{t+1,A_t^\pi}|H_t\right)}, \quad (2)$$

for all $t \in \mathbb{Z}_+$.

Theorem 2 establishes a general regret bound that can be applied to any agent; it characterizes how the regret is determined by the ineffectiveness of the agent in gathering information in this environment, as measured by Γ_t^π, and the total amount of useful information Δ in the environment.

Theorem 2. *Let $\overline{\Gamma}^\pi = \sup_{t\in\mathbb{Z}_+}\Gamma_t^\pi$. For all $T \in \mathbb{Z}_+$, $\mathrm{Regret}(T;\pi) \le \sqrt{\overline{\Gamma}^\pi T\Delta}$.*

Theorem 3 below presents an upper bound on the regret of predictive sampling, by applying the general bound established by Theorem 2, and bounding the information ratio of predictive sampling.

Theorem 3. *If for all $t \in \mathbb{Z}_+$ and $a \in \mathcal{A}$, $\mathbb{P}(R_{t+1,a} \in \cdot|H_t)$ is almost surely σ_{SG}-sub-Gaussian, then for all $T \in \mathbb{Z}_+$, the regret that a predictive sampling agent incurs $\mathrm{Regret}(T) \le \sqrt{2|\mathcal{A}|\sigma_{SG}^2 T\Delta}$.*

Theorem 3 shows how the regret of predictive sampling depends on the environment via Δ. The theorem provides a foundation for deriving more refined regret bounds in specialized models where Δ can be carefully characterized. As a sanity check, when the environment is stationary, by Proposition 4, $\Delta = \Delta_0$ and $\mathrm{Regret}(T) \le \sqrt{2|\mathcal{A}|\sigma_{\mathrm{SG}}^2 T\Delta_0}$,: a predictive sampling agent incurs a regret that grows at a rate which is at most linear in $\sqrt{T}$.

6.3 Regret in Modulated Bernoulli Bandits

The general analysis enables us to focus on particular important examples of non-stationary bandit environments to further examine the performance of predictive sampling through carefully characterizing Δ and obtaining more refined regret bounds. In particular, we focus on the following class of Bernoulli bandit environments, which generalizes an abrupt switching model introduced in (Mellor and Shapiro, 2013).

Example 4 (modulated Bernoulli bandit). *Consider a Bernoulli bandit with independent actions. Let each mean*

reward be denoted by $\theta_{t,a} = \mathbb{P}(R_{t+1,a} = 1|P_t)$*. A mean reward varies over time, transitioning according to*

$$\theta_{t+1,a} = \begin{cases} \theta^{\text{new}}_{t+1,a} \sim \mathbb{P}(\theta_{0,a} \in \cdot), & \text{with probability } q_a, \\ \theta_{t,a}, & \text{otherwise}, \end{cases}$$

where $q_a \in [0, 1]$ *is deterministic and known. Conditioning on* $(\theta_t : t \in \mathbb{Z})$*,* $R_{t+1,a} \sim$ *Bernoulli*$(\theta_{t,a})$*, independent of the rewards associated with other timesteps or actions.*

6.3.1 Regret Lower Bound

We first establish a lower bound on the regret incurred by any agent in a modulated Bernoulli bandit environment.

Theorem 4. *There exists a modulated Bernoulli bandit environment and a constant* $C \in \mathbb{R}_{++}$*, such that for all policy* π*, and all* $T \in \mathbb{Z}_{++}$*,* $\text{Regret}(T; \pi) \geq CT$*.*

The result suggests that the modulated Bernoulli bandit is challenging and a linear dependence of the regret on T cannot be improved; this lower bound provides a baseline to which upper bounds can be compared.

Below we present the key ideas of the proof. We first construct a modulated Bernoulli bandit environment with $\mathcal{A} = \{1, 2\}$, and $\theta_{0,a} \sim \text{unif}\{0, 1\}$ for each $a \in \mathcal{A}$; we let $q = [1/2, 1]$. With this bandit environment, $q_2 = 1$, so selecting action 2 gives no useful information. If the probability of selecting action 2 is small, then the agent collects a reward that is close to $\mathbb{E}[R_{t,1}] = \frac{1}{2}$, and thus incurs a large regret in the current timestep; if the probability of selecting action 2 is large, then the agent is short in information compared to the oracle who knows $R_{t+2:\infty}$, and thus incurs a large regret on the next timestep. We use this fact to lower-bound the regret across all agents.

6.3.2 Regret Analysis of Predictive Sampling

We specialize the regret bound established in Theorem 3 to a modulated Bernoulli bandit through bounding Δ.

Corollary 1. *For all* $T \in \mathbb{Z}_+$*, the regret that a predictive sampling agent incurs in a modulated Bernoulli bandit is* $\text{Regret}(T) \leq \sqrt{\frac{1}{2}|\mathcal{A}|T\{V_1 + (T-1)\min\{V_1, V_2\}\}}$*, where* $V_1 = \sum_{a\in\mathcal{A}}(1 - q_a)\mathbb{H}(\theta_{0,a})$*, and* $V_2 = \sum_{a\in\mathcal{A}}[q_a\mathbb{H}(\theta_{0,a}) + \mathbb{H}(q_a)]$*.*

First, the bound grows at a rate that is at most linear in T, matching the lower bound established in Theorem 4, which indicates that predictive sampling attains good performance. Moreover, when $q_a = 0$ for all $a \in \mathcal{A}$, the environment is stationary and we recover a bound that is linear in $\sqrt{T}$, which confirms that predictive sampling succeeds in such stationary environments. In addition, the bound exhibits nice dependence on environment parameters. In particular, as $\inf_{a\in\mathcal{A}} q_a \to 1$, $\text{Regret}(T) \to 0$, suggesting that a predictive sampling agent performs well when the mean reward vector transitions very often: in such environments, a predictive sampling agent deprioritizes acquiring information about the mean reward and thus performs well.

7 EXPERIMENTS

To quantify the advantage of predictive sampling over Thompson sampling, we conduct numerical experiments in a range of non-stationary AR(1) bandits and AR(1) logistic bandits. Comparisons of predictive sampling with state-of-the-art algorithms (Besbes et al., 2019; Garivier and Moulines, 2008; Kocsis and Szepesvári, 2006) are presented in the supplementary materials.

7.1 AR(1) Bandits

To examine the advantage of predictive sampling over Thompson sampling extensively, we focus on a sequence of AR(1) bandits with varying parameters: $\mathcal{A} = \{1, 2\}$, the stationary distribution of each arm's mean reward is $\mathcal{N}(0, 1)$, $\gamma_1 \in \{0.1, 0.3, 0.5, 0.9\}$, and $\gamma_2 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, respectively.

Figure 1 plots the average reward collected by a predictive sampling agent and that collected by a Thompson sampling agent. Formally, the figure plots 95% confidence intervals of $\mathbb{E}[\frac{1}{T}\sum_{t=0}^{T-1} R_{t+1,A_t}]$ and that of $\mathbb{E}[\frac{1}{T}\sum_{t=0}^{T-1} R_{t+1,A_t^{\pi_{\text{TS}}}}]$ for $T = 1000$. We observe that predictive sampling consistently outperforms Thompson sampling.

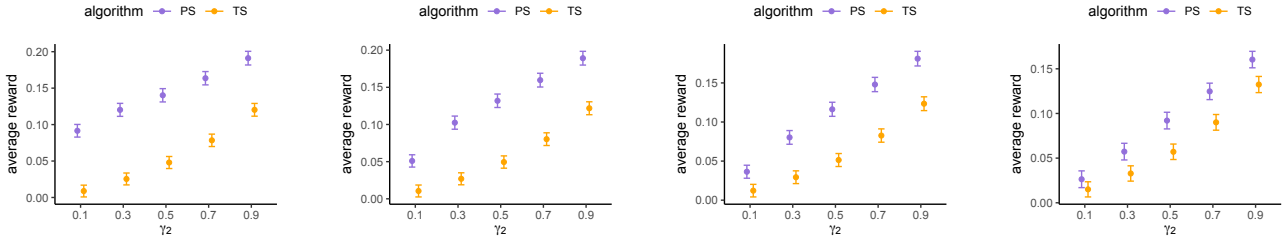
7.2 AR(1) Logistic Bandits

We run experiments with AR(1) logistic bandits of the following specification: $\mathcal{A} = \{1, 2, 3\}$, $\gamma = [x, 0.9, 0.9]$, $\delta^2 = [1 - x^2, 0.19, 0.19]$, where $x \in \{0.1, 0.9\}$. Let $\phi \in \{\phi^{\text{ind}}, \phi^{\text{dep}}\}$, where

$$\phi^{\text{ind}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \phi^{\text{dep}} = \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0 & 0.9 & 0.1 \\ 0.1 & 0 & 0.9 \end{bmatrix}.$$

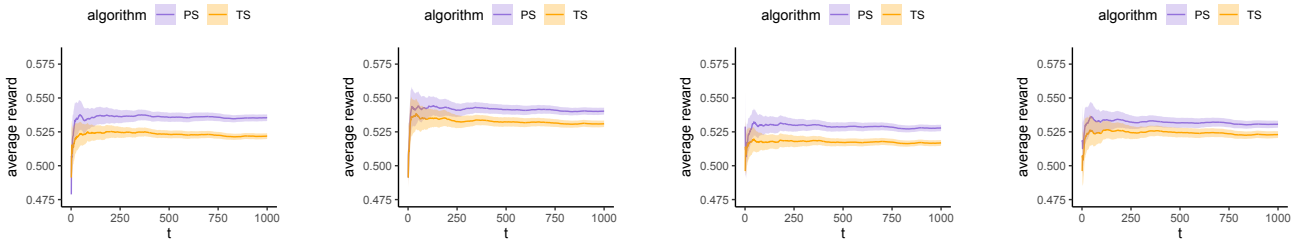
The set of experiments with $\phi = \phi^{\text{ind}}$ corresponds to bandits with independent actions, and the set of experiments with $\phi = \phi^{\text{dep}}$ corresponds to ones with dependent actions.

Figure 2 plots the average rewards collected by a predictive sampling agent and a Thompson sampling agent. Formally, the figure plots 95% confidence intervals of $\mathbb{E}[\frac{1}{t}\sum_{k=0}^{t-1} R_{k+1,A_k}]$ and that of $\mathbb{E}[\frac{1}{t}\sum_{k=0}^{t-1} R_{k+1,A_k^{\pi_{TS}}}]$ for t ranging from 1 to T. Observe that an agent can collect an average reward of 0.5 when taking actions uniformly at random at every timestep. Based on this observation, the plot suggests that predictive sampling consistently outperforms Thompson sampling across all bandit environments that we examine and the advantage is both statistically and practically significant.



(a) $\gamma_1 = 0.1$ (b) $\gamma_1 = 0.3$ (c) $\gamma_1 = 0.5$ (d) $\gamma_1 = 0.9$

Figure 1: Average reward collected by predictive sampling (PS) and Thompson sampling (TS) agents in two-armed AR(1) bandits with varying γ_1 and γ_2, with the error bars representing 95% confidence intervals



(a) $x = 0.1, \phi = \phi^{\text{ind}}$ (b) $x = 0.9, \phi = \phi^{\text{ind}}$ (c) $x = 0.1, \phi = \phi^{\text{dep}}$ (d) $x = 0.9, \phi = \phi^{\text{dep}}$

Figure 2: Average reward collected by a predictive sampling agent and a Thompson sampling agent in non-stationary AR(1) logistic bandits with varying parameters

8 CONCLUDING REMARKS

This paper demonstrates that TS and its variations that were proposed in the literature are often not suited for non-stationary bandit learning, because they fail to intelligently account for the duration of information when selecting actions. To address this, we propose PS, an algorithm that can be viewed as a version of TS that takes the sequence of future rewards to be the learning target. We develop efficient procedures to execute PS in AR(1) bandits and a practical approximation of it in AR(1) logistic bandits. We demonstrate the efficacy of PS through coin-tossing examples, regret bounds, and numerical experiments.

At a high level, our paper illustrates how we can modify an existing algorithm, in this case TS, to construct a new one, in this case PS, that is more suited for non-stationary bandits. As we discussed in Sections 2 and **??**, similar to TS, a number of existing algorithms also do not account for the duration of information when selecting actions. Therefore, a future direction would be to build algorithms more suited for non-stationary bandits by modifying these existing algorithms beyond TS through taking the sequence of future rewards to be the learning target.

A key idea in this paper is taking the sequence of future rewards $R_{t+2:\infty}$ to be the learning target. We believe that this is only a starting point—a future direction would be to investigate other learning targets. For example, by suitably defining the "optimal action" at each timestep, we may alternatively take the sequence of future optimal actions as the learning target. It remains an interesting question whether this learning target or a different learning target enables an agent to more intelligently account for the duration of information when selecting actions.

References

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In Mannor, S., Srebro, N., and Williamson, R. C., editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learn-*

ing Research, pages 39.1–39.26, Edinburgh, Scotland. PMLR.

Anantharam, V., Varaiya, P., and Walrand, J. (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards. *IEEE Transactions on Automatic Control*, 32(11):977–982.

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

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

Besson, L. and Kaufmann, E. (2019). The generalized likelihood ratio test meets KLUCB: an improved algorithm for piece-wise non-stationary bandits. *Proceedings of Machine Learning Research vol XX*, 1:35.

Bubeck, S., Dekel, O., Koren, T., and Peres, Y. (2015). Bandit convex optimization:$\sqrt{T}$ regret in one dimension. In *Conference on Learning Theory*, pages 266–278. PMLR.

Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson sampling. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2019). Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087. PMLR.

Dong, S. and Van Roy, B. (2018). An information-theoretic analysis for thompson sampling with many actions. *Advances in Neural Information Processing Systems*, 31.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

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

Ghatak, G. (2021). A change-detection-based Thompson sampling framework for non-stationary bandits. *IEEE Transactions on Computers*, 70(10):1670–1676.

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media.

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

Hao, B., Lattimore, T., and Qin, C. (2022). Contextual information-directed sampling. *arXiv preprint arXiv:2205.10895*.

Hartland, C., Gelly, S., Baskiotis, N., Teytaud, O., and Sebag, M. (2006). Multi-armed bandit, dynamic environments and meta-bandits.

Kocsis, L. and Szepesvári, C. (2006). Discounted UCB. In *2nd PASCAL Challenges Workshop*, volume 2, pages 51–134.

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statistical science*, 1(3):364–378.

Lattimore, T. and Szepesvári, C. (2019). An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pages 2111–2139. PMLR.

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

Liu, Y., Devraj, A. M., Van Roy, B., and Xu, K. (2022). Gaussian imagination in bandit learning. *arXiv preprint arXiv:2201.01902*.

Lu, X., Van Roy, B., Dwaracherla, V., Ibrahimi, M., Osband, I., and Wen, Z. (2021). Reinforcement learning, bit by bit. *arXiv preprint arXiv:2103.04047*.

Mellor, J. and Shapiro, J. (2013). Thompson sampling in switching environments with Bayesian online change detection. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 442–450, Scottsdale, Arizona, USA. PMLR.

Min, S., Maglaras, C., and Moallemi, C. C. (2019). Thompson sampling with information relaxation penalties. *Advances in Neural Information Processing Systems*, 32.

Ortner, R., Ryabko, D., Auer, P., and Munos, R. (2014). Regret bounds for restless Markov bandits. *Theoretical Computer Science*, 558:62–76.

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

Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4).

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Slivkins, A. and Upfal, E. (2008). Adapting to a changing environment: the Brownian restless bandits. In *COLT*, pages 343–354.

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

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

Turner, R., Saatci, Y., and Rasmussen, C. E. (2009). Adaptive sequential bayesian change point detection. In *Temporal Segmentation Workshop at NIPS*, pages 1–4. Citeseer.

Viappiani, P. (2013). Thompson sampling for bayesian bandits with resets. In *International Conference on Algorithmic Decision Theory*, pages 399–410. Springer.

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

Wilson, R. C., Nassar, M. R., and Gold, J. I. (2010). Bayesian online learning of the hazard rate in change-point problems. *Neural computation*, 22(9):2452–2476.

A PROBABILISTIC FRAMEWORK

Probability theory emerges from an intuitive set of axioms, and this paper builds on that foundation. Statements and arguments we present have precise meaning within the framework of probability theory. However, we often leave out measure-theoretic formalities for the sake of readability. It should be easy for a mathematically-oriented reader to fill in these gaps.

We will define all random quantities with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The probability of an event $F \in \mathcal{F}$ is denoted by $\mathbb{P}(F)$. For any events $F, G \in \mathcal{F}$ with $\mathbb{P}(G) > 0$, the probability of F conditioned on G is denoted by $\mathbb{P}(F|G)$.

A random variable is a function with the set of outcomes Ω as its domain. For any random variable Z, $\mathbb{P}(Z \in \mathcal{Z})$ denotes the probability of the event that Z lies within a set $\mathcal{Z}$. The probability $\mathbb{P}(F|Z = z)$ is of the event F conditioned on the event $Z = z$. When Z takes values in $\mathbb{R}$ and has a density p_Z, though $\mathbb{P}(Z = z) = 0$ for all z, conditional probabilities $\mathbb{P}(F|Z = z)$ are well-defined and denoted by $\mathbb{P}(F|Z = z)$. For fixed F, this is a function of z. We denote the value, evaluated at $z = Z$, by $\mathbb{P}(F|Z)$, which is itself a random variable. Even when $\mathbb{P}(F|Z = z)$ is ill-defined for some z, $\mathbb{P}(F|Z)$ is well-defined because problematic events occur with zero probability.

For each possible realization z, the probability $\mathbb{P}(Z = z)$ that $Z = z$ is a function of z. We denote the value of this function evaluated at Z by $\mathbb{P}(Z)$. Note that $\mathbb{P}(Z)$ is itself a random variable because it depends on Z. For random variables Y and Z and possible realizations y and z, the probability $\mathbb{P}(Y = y|Z = z)$ that $Y = y$ conditioned on $Z = z$ is a function of (y, z). Evaluating this function at (Y, Z) yields a random variable, which we denote by $\mathbb{P}(Y|Z)$.

Particular random variables appear routinely throughout the paper. One is the environment $\mathcal{E}$, a random probability measure over $\mathbb{R}^{\mathcal{A}}$ such that, for all $t \in \mathbb{Z}_+$, $\mathbb{P}(R_{t+1} \in \cdot|\mathcal{E}) = \mathcal{E}(\cdot)$ and $R_{1:\infty}$ is i.i.d. conditioned on $\mathcal{E}$. We often consider probabilities $\mathbb{P}(F|\mathcal{E})$ of events F conditioned on the environment $\mathcal{E}$.

A policy π assigns a probability $\pi(a|h)$ to each action a for each history h. For each policy π, random variables $A_0^\pi, R_{1,A_0^\pi}, A_1^\pi, R_{2,A_1^\pi}, \ldots$, represent a sequence of interactions generated by selecting actions according to π. In particular, with $H_t^\pi = (A_0^\pi, R_{1,A_0^\pi}, \ldots, R_{t,A_{t-1}^\pi})$ denoting the history of interactions through time t, we have $\mathbb{P}(A_t^\pi|H_t^\pi) = \pi(A_t^\pi|H_t^\pi)$. As shorthand, we generally suppress the superscript π and instead indicate the policy through a subscript of $\mathbb{P}$. For example,

$$\mathbb{P}_\pi(A_t|H_t) = \mathbb{P}(A_t^\pi|H_t^\pi) = \pi(A_t^\pi|H_t^\pi).$$

We denote independence of random variables X and Y by $X \perp Y$ and conditional independence, conditioned on another random variable Z, by $X \perp Y|Z$.

When expressing expectations, we use the same subscripting notation as with probabilities. For example, the expectation of a reward R_{t+1,A_t^π} is written as $\mathbb{E}[R_{t+1,A_t^\pi}] = \mathbb{E}_\pi[R_{t+1,A_t}]$.

Much of the paper studies properties of interactions under a specific policy π_{agent}. When it is clear from context, we suppress superscripts and subscripts that indicate this. For example, $H_t = H_t^{\pi_{\text{agent}}}$, $A_t = A_t^{\pi_{\text{agent}}}$, $R_{t+1} = R_{t+1,A_t^{\pi_{\text{agent}}}}$. Further,

$$\mathbb{P}(A_t|H_t) = \mathbb{P}_{\pi_{\text{agent}}}(A_t|H_t) = \pi_{\text{agent}}(A_t|H_t) \qquad \text{and} \qquad \mathbb{E}[R_{t+1,A_t}] = \mathbb{E}_{\pi_{\text{agent}}}[R_{t+1,A_t}].$$

B INFORMATION-THEORETIC CONCEPTS, NOTATIONS, AND RELATIONS

We review some standard information-theoretic concepts and associated notations in this section.

A central concept is the entropy $\mathbb{H}(X)$, which quantifies the information content or, equivalently, the uncertainty of a random variable X. For a random variable X that takes values in a countable set $\mathcal{X}$, we will define the entropy to be $\mathbb{H}(X) = -\mathbb{E}[\ln \mathbb{P}(X)]$, with a convention that $0 \ln 0 = 0$. Note that we are defining entropy here using the natural rather than binary logarithm. As such, our notion of entropy can be interpreted as the expected number of nats – as opposed to bits – required to identify X. The realized conditional entropy $\mathbb{H}(X|Y = y)$ quantifies the uncertainty remaining after observing $Y = y$. If Y takes on values in a countable set $\mathcal{Y}$ then $\mathbb{H}(X|Y = y) = -\mathbb{E}[\ln \mathbb{P}(X|Y)|Y = y]$. This can be viewed as a function $f(y)$ of y, and we write the random variable $f(Y)$ as $\mathbb{H}(X|Y = Y)$. The conditional entropy $\mathbb{H}(X|Y)$ is its expectation $\mathbb{H}(X|Y) = \mathbb{E}[\mathbb{H}(X|Y = Y)]$.

The mutual information $\mathbb{I}(X; Y) = \mathbb{H}(X) - \mathbb{H}(X|Y)$ quantifies information common to random variables X and Y, or equivalently, the information about Y required to identify X. If Z is a random variable taking on values in a countable

set $\mathcal{Z}$ then the realized conditional mutual information $\mathbb{I}(X;Y|Z=z)$ quantifies remaining common information after observing $Z=z$, defined by $\mathbb{I}(X;Y|Z=z)=\mathbb{H}(X|Z=z)-\mathbb{H}(X|Y,Z=z)$. The conditional mutual information $\mathbb{I}(X;Y|Z)$ is its expectation $\mathbb{I}(X;Y|Z)=\mathbb{E}[\mathbb{I}(X;Y|Z=Z)]$.

For random variables X and Y taking on values in (possibly uncountable) sets $\mathcal{X}$ and $\mathcal{Y}$, mutual information is defined by $\mathbb{I}(X;Y)=\sup_{f\in\mathcal{F}_{\text{finite}},g\in\mathcal{G}_{\text{finite}}}\mathbb{I}(f(X);g(Y))$, where $\mathcal{F}_{\text{finite}}$ and $\mathcal{G}_{\text{finite}}$ are the sets of functions mapping $\mathcal{X}$ and $\mathcal{Y}$ to finite ranges. Specializing to the case where $\mathcal{X}$ and $\mathcal{Y}$ are countable recovers the previous definition. The generalized notion of entropy is then given by $\mathbb{H}(X)=\mathbb{I}(X;X)$. Conditional counterparts to mutual information and entropy can be defined in a manner similar to the countable case.

One representation of mutual information, which we will use, is in terms of the differential entropy. The differential entropy $\mathbf{h}(X)$ of a random variable X with probability density f is defined by

$$\mathbf{h}(X)=-\int f(x)\ln f(x)dx.$$

The conditional differential entropy $\mathbf{h}(X|Y)$ of X conditioned on Y is evaluated similarly but with a conditional density function. Finally, mutual information can be written as $\mathbb{I}(X;Y)=\mathbf{h}(X)-\mathbf{h}(X|Y)$.

We will also make use of KL-divergence as measures of difference between distributions. We denote KL-divergence by

$$\mathbf{d}_{\mathrm{KL}}(P\|P')=\int P(dx)\ln\frac{dP}{dP'}(x).$$

Gibbs' inequality asserts that $\mathbf{d}_{\mathrm{KL}}(P\|P')\geq 0$, with equality if and only if P and P' agree almost everywhere with respect to P.

The following result is established by Theorem 5.2.1 of (Gray, 2011).

Lemma 3 (Variational form of the KL-divergence)**.** *For any probability distribution P and real-valued random variable X, both defined with respect to a measureable space $(\Omega',\mathbb{F}')$, let $\mathbb{E}_P[X]=\int_{x\in\Re}xP(dx)$. For probability distributions P and P' on a measureable space $(\Omega',\mathbb{F}')$ such that P is absolutely continuous with respect to P',*

$$\mathbf{d}_{\mathrm{KL}}(P\|P')=\sup_X(\mathbb{E}_P[X]-\ln\mathbb{E}_{P'}[\exp(X)]),\tag{3}$$

where the supremum is taken over real-valued random variables on $(\Omega',\mathbb{F}')$ for which $\mathbb{E}_Q[\exp(X)]<\infty$.

Mutual information and KL-divergence are intimately related. For any probability measure $P(\cdot)=\mathbb{P}((X,Y)\in\cdot)$ over a product space $\mathcal{X}\times\mathcal{Y}$ and probability measure P' generated via a product of marginals $P'(dx\times dy)=P(dx)P(dy)$, mutual information can be written in terms of KL-divergence:

$$\mathbb{I}(X;Y)=\mathbf{d}_{\mathrm{KL}}(P\|P').\tag{4}$$

Further, the following lemma presents an alternative representation of mutual information.

Lemma 4 (KL-divergence representation of mutual information)**.** *For any random variables X and Y,*

$$\mathbb{I}(X;Y)=\mathbb{E}[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y\in\cdot|X)\|\mathbb{P}(Y\in\cdot))].\tag{5}$$

In other words, the mutual information between X and Y is the KL-divergence between the distribution of Y with and without conditioning on X.

Mutual information satisfies the chain rule and the data-processing inequality.

Lemma 5 (Chain rule for mutual information)**.**

$$\mathbb{I}(X_1,X_2,...,X_n;Y)=\sum_{i=1}^{n}\mathbb{I}(X_i;Y|X_1,X_2,...,X_{i-1}).$$

Lemma 6 (Data processing inequality for mutual information)**.** *If X and Z are independent conditioning on Y, then*

$$\mathbb{I}(X;Y)\geq\mathbb{I}(X;Z).$$

The following lemma presents one useful property of mutual information.

Lemma 7. *Let A, B, and C be three random variables. If $A \perp C|B$ then*

$$\mathbb{I}(A; B|C) \leq \mathbb{I}(A; B).$$

Proof. We prove for the case where B has finite entropy. For the case where B has infinite entropy, we use differential entropy instead of entropy in the analysis.

$$\begin{aligned}\mathbb{I}(A; B|C) &= \mathbb{H}(A|C) - \mathbb{H}(A|B, C) \\ &= \mathbb{H}(A|C) - \mathbb{H}(A|B) \\ &\leq \mathbb{H}(A) - \mathbb{H}(A|B) \\ &= \mathbb{I}(A; B).\end{aligned}$$

□

C ALTERNATIVE REPRESENTATIONS OF PREDICTIVE SAMPLING AND THOMPSON SAMPLING: PROOFS OF LEMMAS 1 and 2

C.1 Proof of Lemma 1

Proof. With predictive sampling, first observe that for all $t \in \mathbb{Z}_+$, $\mathbb{P}(\hat{R}^{(t)}_{t+2:\infty} \in \cdot|H_t) = \mathbb{P}(R_{t+2:\infty} \in \cdot|H_t)$. Therefore, for all $t \in \mathbb{Z}_+$, $\mathbb{P}(\hat{\theta}_t \in \cdot|H_t) = \mathbb{P}(\mathbb{E}[R_{t+1}|H_t, R_{t+2:\infty} \leftarrow \hat{R}^{(t)}_{t+2:\infty}] \in \cdot|H_t) = \mathbb{P}(\mathbb{E}[R_{t+1}|H_t, R_{t+2:\infty}] \in \cdot|H_t)$. □

C.2 Proof of Lemma 2

Proof. With Thompson sampling, we can first observe that for all $t \in \mathbb{Z}_+$, $\mathbb{P}(\hat{P}^{\pi_{\text{TS}}}_t \in \cdot|H^{\pi_{\text{TS}}}_t) = \mathbb{P}(P^{\pi_{\text{TS}}}_t \in \cdot|H^{\pi_{\text{TS}}}_t)$. Therefore, for all $t \in \mathbb{Z}_+$, $\mathbb{P}(\hat{\theta}^{\pi_{\text{TS}}}_t \in \cdot|H^{\pi_{\text{TS}}}_t) = \mathbb{P}(\mathbb{E}[R_{t+1}|P^{\pi_{\text{TS}}}_t \leftarrow \hat{P}^{\pi_{\text{TS}}}_t] \in \cdot|H^{\pi_{\text{TS}}}_t) = \mathbb{P}(\mathbb{E}[R_{t+1}|P^{\pi_{\text{TS}}}_t] \in \cdot|H^{\pi_{\text{TS}}}_t)$. □

D EQUIVALENCE OF PREDICTIVE SAMPLING TO THOMPSON SAMPLING IN STATIONARY BANDIT ENVIRONMENTS: PROOF OF THEOREM 1

Proof. Observe that it is sufficient to show that for all $t \in \mathbb{Z}_+$,

$$\mathbb{P}\left(\hat{\theta}_t \in \cdot|H_t\right) = \mathbb{P}\left(\hat{\theta}^{\pi_{\text{TS}}}_t \in \cdot|H^{\pi_{\text{TS}}}_t \leftarrow H_t\right). \tag{6}$$

We first show that if H_t and $H^{\pi_{\text{TS}}}_t$ have the same support, then the change of measure is well-defined and (6) holds. Note that, for any $t \in \mathbb{Z}_+$, $\lim_{T\to\infty} \frac{1}{T-t-1} \sum_{k=t+1}^{T-1} R_{k+1} \overset{\text{a.s.}}{=} \mathbb{E}[R_{t+1}|P]$ by the strong law of large numbers. Therefore, for all $t \in \mathbb{Z}_+$,

$$\mathbb{E}[R_{t+1}|H_t, R_{t+2:\infty}] \overset{\text{a.s.}}{=} \mathbb{E}[R_{t+1}|H_t, R_{t+2:\infty}, \mathbb{E}[R_{t+1}|P]] = \mathbb{E}[R_{t+1}|\mathbb{E}[R_{t+1}|P]] = \mathbb{E}[R_{t+1}|P]. \tag{7}$$

These conditional expectations determine how actions are sampled by predictive sampling and Thompson sampling, and the equivalence implies that the two implement the same policy; that is, for all $t \in \mathbb{Z}_+$,

$$\begin{aligned}\mathbb{P}(\hat{\theta}_t \in \cdot|H_t) &\overset{(a)}{=} \mathbb{P}(\mathbb{E}[R_{t+1}|H_t, R_{t+2:\infty}] \in \cdot|H_t) \\ &\overset{(b)}{=} \mathbb{P}(\mathbb{E}[R_{t+1}|P] \in \cdot|H_t) = \mathbb{P}(\mathbb{E}[R_{t+1}|P] \in \cdot|H^{\pi_{\text{TS}}}_t \leftarrow H_t) \overset{(c)}{=} \mathbb{P}(\hat{\theta}^{\pi_{\text{TS}}}_t \in \cdot|H^{\pi_{\text{TS}}}_t \leftarrow H_t),\end{aligned}$$

where (a) follows from Lemma 1, (b) follows from (7), and (c) follows from Lemma 2 and from the fact that the Thompson sampling agent we consider implementing Algorithm 2 using $P^{\pi_{\text{TS}}}_t = P$ for all $t \in \mathbb{Z}_+$.
Note that $H_0 = H^{\pi_{\text{TS}}}_0$, so it is clear that by induction, for all $t \in \mathbb{Z}_+$, H_t and $H^{\pi_{\text{TS}}}_t$ have the same support and (6) holds. □

E MAXIMUM ADVANTAGE OF PREDICTIVE SAMPLING OVER THOMPSON SAMPLING: PROOFS OF PROPOSITIONS 1 and 2

Proof. It suffices to show that for all $\epsilon \in (0,1)$, there exists a nonstationary bandit environment with rewards bounded in $[0,1]$ such that for all $T \in \mathbb{Z}_+$, the following holds:

$$\mathbb{E}\left[\sum_{t=0}^{T-1} R_{t+1,A_t}\right] - \mathbb{E}\left[\sum_{t=0}^{T-1} R_{t+1,A_t^{\pi_{\mathrm{TS}}}}\right] \geq (1-\epsilon)T. \tag{8}$$

Let $\epsilon \in (0,1)$. Consider a modulated Bernoulli bandit (see Example 4 in Section 5.3) with K arms, where

$$K = \left\lceil \log_{1-\frac{\epsilon}{6}}\left(\frac{\frac{\epsilon}{3}}{1-\frac{\epsilon}{3}}\right)\right\rceil + 1.$$

Arm 1 has a deterministic mean of $x = 1 - \frac{\epsilon}{3}$, which does not change with time. Each of arm 2 through K's mean reward takes value 1 with probability $p = \frac{\frac{\epsilon}{6}}{1-\frac{\epsilon}{6}}$ and takes value 0 with probability $1-p$. The probability of transition is $q = (0, b, ..., b)$, where $b = 1 - \frac{\epsilon}{6}$. Note that this bandit environment is nonstationary by Lemma 8.

A predictive sampling agent estimates $\hat{\theta}_t$ at each timestep $t \in \mathbb{Z}_+$. Note that for all $a \in \{2, ..., K\}$,

$$\mathbb{E}\left[R_{t+1,a}|H_t, R_{t+2:\infty}\right] = \mathbb{P}\left(\theta_{t,a} = 1|H_t, R_{t+2:\infty}\right) \leq 1 - b^2(1-p) = 1 - \left(1-\frac{\epsilon}{3}\right)\left(1-\frac{\epsilon}{6}\right),$$

which implies that for all $a \in \{2, ..., K\}$,

$$\hat{\theta}_{t,a} \leq 1 - \left(1-\frac{\epsilon}{3}\right)\left(1-\frac{\epsilon}{6}\right) < 1 - \frac{\epsilon}{3} = x = \hat{\theta}_{t,1}.$$

So a predictive sampling agent selects action 1 with probability one and collects cumulative reward

$$\mathbb{E}\left[\sum_{t=0}^{T-1} R_{t+1,A_t}\right] = xT. \tag{9}$$

A Thompson sampling agent estimates $\hat{\theta}_t^{\pi_{\mathrm{TS}}}$ at each timestep $t \in \mathbb{Z}_+$. Note that $\hat{\theta}_{t,1}^{\pi_{\mathrm{TS}}} = x = 1 - \frac{\epsilon}{3} \in (0,1)$. So a Thompson sampling agent selects action 1 at each timestep $t \in \mathbb{Z}_+$ with probability

$$\mathbb{P}\left(\max_{a\in\{2,...,K\}} \hat{\theta}_{t,a}^{\pi_{\mathrm{TS}}} < \hat{\theta}_{t,1}^{\pi_{\mathrm{TS}}} | H_t^{\pi_{\mathrm{TS}}}\right) = \prod_{a=2}^{K} \mathbb{P}\left(\hat{\theta}_{t,a}^{\pi_{\mathrm{TS}}} = 0 | H_t^{\pi_{\mathrm{TS}}}\right) \leq \left(b(1-p)+1-b\right)^{K-1} = (1-bp)^{K-1}.$$

Note that $x = 1 - \frac{\epsilon}{3} > \frac{\epsilon}{3} = 1 - b + bp$. So a Thompson sampling agent collects cumulative rewards

$$\mathbb{E}\left[\sum_{t=0}^{T-1} R_{t+1,A_t^{\pi_{\mathrm{TS}}}}\right] \leq (1-bp)^{K-1}x + \left[1-(1-bp)^{K-1}\right](1-b+bp). \tag{10}$$

So by (9) and (10), for all $T \in \mathbb{Z}_+$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1} R_{t+1,A_t}\right] - \mathbb{E}\left[\sum_{t=0}^{T-1} R_{t+1,A_t^{\pi_{\mathrm{TS}}}}\right] &\geq \left\{x - (1-bp)^{K-1}x - \left[1-(1-bp)^{K-1}\right](1-b+bp)\right\}T \\
&\geq \left\{x - (1-bp)^{K-1}x - (1-b+bp)\right\}T \\
&\geq (1-\epsilon)T.
\end{aligned}$$

□

Lemma 8. *A modulated Bernoulli bandit is stationary if and only if for all $a \in \mathcal{A}$, $q_a = 0$, $q_a = 1$, or $\mathbb{P}(\theta_{0,a} \in \cdot)$ is a point mass.*

Proof. When the environment is stationary, by definition, there exists a distribution P that generates the environment. If we use $\hat{F}_n$ to denote the empirical distribution of $R_{-n:1}$, then by Glivenko–Cantelli theorem, $\lim_{n\to+\infty} \|\hat{F}_n - P\|_\infty \overset{a.s.}{=} 0$. Therefore,

$$\mathbb{P}(R_2 \in \cdot | R_{-\infty:1}) = \mathbb{P}(R_2 \in \cdot | P) = \mathbb{P}(R_3 \in \cdot | P) = \mathbb{P}(R_3 \in \cdot | R_{-\infty:1}).$$

This implies that a necessary condition for an environment to be stationary is

$$\mathbb{E}[R_2 | R_{-\infty:1}] = \mathbb{E}[R_3 | R_{-\infty:1}].$$

If we restrict our attentions to modulated Bernoulli bandit, a necessary condition for a such environment to be stationary is

$$\mathbb{E}[\theta_{1,a} | R_{-\infty:1}] = \mathbb{E}[\theta_{2,a} | R_{-\infty:1}], \ \forall a \in \mathcal{A}. \tag{11}$$

We first describe an alternative formulation of the modulated Bernoulli bandit environment. For all $a \in \mathcal{A}$, $\theta_{0,a} = X_{0,a}$ and, for all $a \in \mathcal{A}$ and $t \in \mathbb{Z}_+$,

$$\theta_{t+1,a} = \begin{cases} X_{t+1,a} & \text{if } B_{t+1,a} = 1 \\ \theta_{t,a} & \text{if } B_{t+1,a} = 0. \end{cases}$$

where $(B_{t,a} : t \in \mathbb{Z}_{++})$ is an i.i.d. Bernoulli(q_a) process and $(X_{t,a} : t \in \mathbb{Z}_+)$ is an i.i.d. process with discrete range. With this formulation, observe that for all $t \in \mathbb{Z}_+$ and $a \in \mathcal{A}$,

$$\theta_{t+1,a} = (1 - B_{t+1,a})\theta_{t,a} + B_{t+1,a} X_{t+1,a}. \tag{12}$$

Now we proceed to show that a modulated Bernoulli bandit is stationary if and only if for all $a \in \mathcal{A}$, $q_a = 0$, $q_a = 1$, or $\mathbb{P}(\theta_{0,a} \in \cdot)$ is a point mass. We prove the two directions separately.

1. Suppose that the environment is stationary. By (12), for all $a \in \mathcal{A}$,

$$\begin{aligned} \mathbb{E}[\theta_{2,a} | R_{-\infty:1}] &= \mathbb{E}[(1 - B_{2,a})\theta_{1,a} + B_{2,a} X_{2,a} | R_{-\infty,1}] \\ &= (1 - q_a)\mathbb{E}[\theta_{1,a} | R_{-\infty:1}] + q_a \mathbb{E}[X_{1,a}] \\ &= (1 - q_a)\mathbb{E}[\theta_{1,a} | R_{-\infty:1}] + q_a \mathbb{E}[\theta_{1,a}]. \end{aligned}$$

   Therefore, the necessary condition (11) simplifies to

$$\mathbb{E}[\theta_{1,a} | R_{-\infty} : 1] = (1 - q_a)\mathbb{E}[\theta_{1,a} | R_{-\infty:1}] + q_a \mathbb{E}[\theta_{1,a}], \ \forall a \in \mathcal{A}.$$

   This implies that for all $a \in \mathcal{A}$, either $q_a = 0$, or $\mathbb{E}[\theta_{1,a} | R_{-\infty,1}] = \mathbb{E}[\theta_{1,a}]$. The latter implies that $q_a = 1$ or that $\mathbb{P}(\theta_{0,a} \in \cdot)$ is a point mass. Hence, we have proved that if the environment is stationary, then for all $a \in \mathcal{A}$, $q_a = 0$, $q_a = 1$, or $\mathbb{P}(\theta_{0,a} \in \cdot)$ is a point mass.

2. Suppose that $q_a = 0$, $q_a = 1$, or $\mathbb{P}(\theta_{0,a} \in \cdot)$ is a point mass for all $a \in \mathcal{A}$. Below we construct a probability distribution P that generates the bandit environment: for each $a \in \mathcal{A}$,

   (a) if $q_a = 0$, then we let $P_a \sim \text{Bernoulli}(\theta_{0,a})$;
   (b) otherwise, we let $P_a \sim \text{Bernoulli}(\mathbb{E}[\theta_{0,a}])$.

   It is clear that P generates the environment, so by definition, the environment is stationary. Hence, we have shown that if $q_a = 0$, $q_a = 1$, or $\mathbb{P}(\theta_{0,a} \in \cdot)$ is a point mass for all $a \in \mathcal{A}$, the bandit environment is stationary.

□

F AR(1) BANDITS: PROOF OF PROPOSITION 3

Proof. The analysis is done for predictive sampling and an arbitrary arm $a \in \mathcal{A}$. We drop the the subscript a from most of the random variables.

For all $t \in \mathbb{Z}_+$, and $n \in \mathbb{Z}_+$, $n \geq 2$, let

$$\overline{\theta}_t^H(n) = \mathbb{E}[R_{t+1}|H_t, R_{t+2:t+n+1}].$$

We define

$$\tilde{R}_{t+2} = R_{t+2}, \text{ and } \tilde{R}_{t+i} = R_{t+i} - \gamma R_{t+i-1} - (1-\gamma)c \text{ for all } i \in \{3, ..., n+1\}.$$

Then we can rewrite $\overline{\theta}_t^H(n)$ as follows:

$$\overline{\theta}_t^H(n) = \mathbb{E}\left[R_{t+1}|H_t, R_{t+2:t+n+1}\right] = \mathbb{E}\left[R_{t+1}|H_t, \tilde{R}_{t+2:t+n+1}\right].$$

Conditioned on H_t, for all $n \geq 2$, $R_{t+1:t+n+1}$ is Gaussian, so the vector constructed by stacking R_{t+1} and $\tilde{R}_{t+2:t+n+1}$ is also Gaussian. We use μ_n and Σ_n to denote its mean and variance. In particular, we view μ_n as a block matrix with blocks $\mu_{n1} \in \mathbb{R}$ and $\mu_{n2} \in \mathbb{R}^n$ and Σ_n a block matrix with blocks $\Sigma_{n11} \in \mathbb{R}$, $\Sigma_{n12} \in \mathbb{R}^{1\times n}$, $\Sigma_{n21} \in \mathbb{R}^{n\times 1}$ and $\Sigma_{n22} \in \mathbb{R}^{n\times n}$. Observe that for all $n \geq 2$, conditioned on H_t,

$$\overline{\theta}_t^H(n) = \mathbb{E}\left[R_{t+1}|H_t, \tilde{R}_{t+2:t+n+1}\right] = \mu_{n1} + \Sigma_{n12}\Sigma_{n22}^{-1}\left(\tilde{R}_{t+2:t+n+1} - \mu_{n2}\right).$$

Then, conditioned on H_t, $\overline{\theta}_t^H(n)$ is Gaussian with mean

$$\mathbb{E}\left[\overline{\theta}_t^H(n)\middle|H_t\right] = \mu_{n1} = \mu_{t,a}$$

and variance

$$\mathbb{V}\left(\overline{\theta}_t^H(n)\middle|H_t\right) = \Sigma_{n12}\Sigma_{n22}^{-1}\Sigma_{n22}\Sigma_{n22}^{-1}\Sigma_{n21} = \Sigma_{n12}\Sigma_{n22}^{-1}\Sigma_{n21}.$$

Observe that for all $t \in \mathbb{Z}_+$ and $k \in \mathbb{Z}_+$, $k \geq 2$:

$$\tilde{R}_{t+k+1} = R_{t+k+1} - \gamma R_{t+k} - (1-\gamma)c = W_{t+k} + Z_{t+k+1} - \gamma Z_{t+k}.$$

Then we have for all $t \in \mathbb{Z}_+$:

(i) $\mathbb{V}(R_{t+1}, \tilde{R}_{t+2}|H_t) = \mathbb{V}(R_{t+1}, R_{t+2}|H_t) = \mathbb{V}(\theta_t + Z_{t+1}, \gamma\theta_t + W_{t+1} + Z_{t+2}|H_t) = \gamma\mathbb{V}(\theta_t|H_t) = \gamma\sigma_t^2$;

(ii) $\mathbb{V}(R_{t+1}, \tilde{R}_{t+k+1}|H_t) = \mathbb{V}(\theta_t + Z_{t+1}, W_{t+k} + Z_{t+k+1} - \gamma Z_{t+k}|H_t) = 0$ for all $k \geq 2$, $k \in \mathbb{Z}_+$;

(iii) $\mathbb{V}(\tilde{R}_{t+2}|H_t) = \mathbb{V}(R_{t+2}|H_t) = \mathbb{V}(\gamma\theta_t + W_{t+1} + Z_{t+2}|H_t) = \gamma^2\sigma_t^2 + \delta^2 + \sigma^2$;

(iv) $\mathbb{V}(\tilde{R}_{t+k+1}|H_t) = \mathbb{V}(W_{t+k} + Z_{t+k+1} - \gamma Z_{t+k}|H_t) = \delta^2 + \sigma^2 + \gamma^2\sigma^2$, for all $k \geq 2$, $k \in \mathbb{Z}_+$;

(v) $\mathbb{V}(\tilde{R}_{t+i+1}, \tilde{R}_{t+k+1}|H_t) = \mathbb{V}(W_{t+i} + Z_{t+i+1} - \gamma Z_{t+i}, W_{t+k} + Z_{t+k+1} - \gamma Z_{t+k}|H_t)$ for all $k > i \geq 1$. Hence, $\mathbb{V}(\tilde{R}_{t+i+1}, \tilde{R}_{t+k+1}|H_t) = \gamma\sigma^2$ for $k = i+1$, $i \geq 1$, and $\mathbb{V}(\tilde{R}_{t+i+1}, \tilde{R}_{t+k+1}|H_t) = 0$ for $k \geq i+2$, $i \geq 1$.

Based on these derivations,

$$\Sigma_{n12} = \gamma\sigma_t^2\begin{bmatrix}1 & 0 & 0 & ... & 0\end{bmatrix},$$

and

$$\Sigma_{n22} = \begin{bmatrix}\gamma^2\sigma_t^2 + \delta^2 + \sigma^2 & \gamma\sigma^2 & 0 & ... & 0 \\ \gamma\sigma^2 & \delta^2 + (1+\gamma^2)\sigma^2 & \gamma\sigma^2 & ... & 0 \\ 0 & \gamma\sigma^2 & \delta^2 + (1+\gamma^2)\sigma^2 & ... & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & ... & \delta^2 + (1+\gamma^2)\sigma^2\end{bmatrix}.$$

We use Gaussian elimination to compute the inverse of Σ_{n22}. In particular, for $k = 1, 2, ..., n-1$, we perform row operations on the $(k+1)$-th to last row: Subtract r_k times k-th to the last row from the $(k+1)$-th to last row. The sequence

$\{r_k\}$ are such that the matrix becomes lower-triangular after the $n-1$ row operations. If we use d_k to denote the diagonal entry of the matrix on the k-th to last row after these row operations. Then the sequence $\{d_k\}$ satisfies the following recurrence:

$$\begin{aligned}
d_1 &= \delta^2 + (1+\gamma^2)\sigma^2, \\
d_k &= \delta^2 + (1+\gamma^2)\sigma^2 - \frac{\gamma^2\sigma^4}{d_{k-1}},\ k = 2, ..., n-1, \\
d_n &= \gamma^2(\sigma_t^2 - \sigma^2) + \delta^2 + (1+\gamma^2)\sigma^2 - \frac{\gamma^2\sigma^4}{d_{n-1}}.
\end{aligned}$$

Note that the recurrence induces the following fixed-point equation:

$$d_* = \delta^2 + (1+\gamma^2)\sigma^2 - \frac{\gamma^2\sigma^4}{d_*}.$$

Solving for d_*, we have

$$\begin{aligned}
d_* &= \frac{1}{2}\left(\gamma^2\sigma^2 + \sigma^2 + \delta^2 \pm \sqrt{(\gamma^2\sigma^2 + \sigma^2 + \delta^2)^2 - 4\gamma^2\sigma^4}\right) \\
&= \frac{1}{2}\left(\gamma^2\sigma^2 + \sigma^2 + \delta^2 \pm \sqrt{(\delta^2 + \sigma^2 - \gamma^2\sigma^2)^2 + 4\gamma^2\delta^2\sigma^2}\right).
\end{aligned}$$

Then for all $t \in \mathbb{Z}_+$, the variance

$$\begin{aligned}
\mathbb{V}\left(\hat{\theta}_{t,a}|H_t\right) &= \mathbb{V}\left(\overline{\theta}^H_{t,a}|H_t\right) = \lim_{n\to\infty} \mathbb{V}\left(\overline{\theta}^H_{t,a}(n)|H_t\right) \\
&= \lim_{n\to\infty} \Sigma_{n12}\Sigma_{n22}^{-1}\Sigma_{n21} = \frac{\gamma_a^2\sigma_{t,a}^4}{d_* + \gamma_a^2(\sigma_{t,a}^2 - \sigma^2)} = \frac{\gamma_a^2\sigma_{t,a}^4}{\gamma_a^2\sigma_{t,a}^2 + x_a^*},
\end{aligned}$$

where

$$x_a^* = \frac{1}{2}\left(\delta_a^2 + \sigma_a^2 - \gamma_a^2\sigma_a^2 + \sqrt{(\delta_a^2 + \sigma_a^2 - \gamma_a^2\sigma_a^2)^2 + 4\gamma_a^2\delta_a^2\sigma_a^2}\right).$$

□

G CHARACTERIZING Δ_t IN STATIONRY ENVIRONMENTS: PROOF OF PROPOSITION 4

Proof. By definition of stationary environments, there exists a distribution P such that the environment is generated by P. Observe that for all $t \in \mathbb{Z}$, and $n \in \mathbb{Z}_+$, if we use $\hat{F}_{t,n}$ to denote the empirical distribution of $R_{t+1-n:t+1}$, then by Glivenko–Cantelli theorem, $\lim_{n\to+\infty} \|\hat{F}_{t,n} - P\|_\infty = 0$. Therefore, for all $t \in \mathbb{Z}_+$,

$$\mathcal{E}_t = \mathbb{P}(R_{t+2:\infty} \in \cdot | R_{-\infty:t+1}) = \mathbb{P}(R_{t+2:\infty} \in \cdot | P, R_{-\infty:t+1}) = \mathbb{P}(R_{t+2:\infty} \in \cdot | P),$$

which corresponds to an infinite product of P. This implies that

$$\Delta_t = \mathbb{I}(R_{t+2:\infty}; \mathcal{E}_t | H_t) - \mathbb{I}(R_{t+1:\infty}; \mathcal{E}_{t-1} | H_t) = 0$$

for all $t \in \mathbb{Z}_{++}$. □

H GENERAL REGRET BOUND: PROOF OF THEOREM 2

Proof. For all policy π and $T \in \mathbb{Z}_+$,

$$\begin{aligned}\mathrm{Regret}(T;\pi) &= \sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,*} - R_{t+1,A_t^\pi}\right] \\ &\le \sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,*} - R_{t+1,A_t^\pi}\right]_+ \\ &\overset{(a)}{\le} \sum_{t=0}^{T-1} \sqrt{\Gamma_t^\pi \mathbb{I}(R_{t+2:\infty}; A_t^\pi, R_{t+1,A_t^\pi}|H_t^\pi)} \\ &\overset{(b)}{\le} \sqrt{\sum_{t=0}^{T-1} \mathbb{I}\left(R_{t+2:\infty}; A_t^\pi, R_{t+1,A_t^\pi}|H_t^\pi\right)} \sqrt{\overline{\Gamma}^\pi T}, \end{aligned} \tag{13}$$

where $[X]_+$ denotes the positive part of a random variable X, (a) follows from the definition of the information ratio, and (b) follows from the Cauchy-Bunyakovsky-Schwarz inequality. Recall that $\mathcal{E}_t = \mathbb{P}(R_{t+2:\infty} \in \cdot | R_{-\infty:t+1})$, so $R_{t+2:\infty} \perp R_{-\infty:t+1}|\mathcal{E}_t$, which implies that $R_{t+2:\infty} \perp H_{t+1}^\pi|\mathcal{E}_t$. Hence, for all policy π and $t \in \mathbb{Z}_+$,

$$\mathbb{I}(R_{t+2:\infty}; A_t^\pi, R_{t+1,A_t^\pi}|H_t^\pi) = \mathbb{I}(R_{t+2:\infty}; \mathcal{E}_t|H_t^\pi) - \mathbb{I}(R_{t+2:\infty}; \mathcal{E}_t|H_{t+1}^\pi).$$

Therefore, for all $T \in \mathbb{Z}_+$,

$$\begin{aligned}\sum_{t=0}^{T-1} \mathbb{I}(R_{t+2:\infty}; A_t^\pi, R_{t+1,A_t^\pi}|H_t^\pi) &= \sum_{t=0}^{T-1} \left(\mathbb{I}(R_{t+2:\infty}; \mathcal{E}_t|H_t^\pi) - \mathbb{I}(R_{t+2:\infty}; \mathcal{E}_t|H_{t+1}^\pi)\right) \\ &\le \mathbb{I}(R_{2:\infty}; \mathcal{E}_0) + \sum_{t=1}^{T-1} \left[\mathbb{I}(R_{t+2:\infty}; \mathcal{E}_t|H_t^\pi) - \mathbb{I}(R_{t+1:\infty}; \mathcal{E}_{t-1}|H_t^\pi)\right] \\ &= \sum_{t=0}^{T-1} \Delta_t = \Delta. \end{aligned} \tag{14}$$

Incorporating (13) and (14), we have for all $T \in \mathbb{Z}_+$,

$$\mathrm{Regret}(T;\pi) \le \sqrt{\sum_{t=0}^{T-1} \mathbb{I}\left(R_{t+2:\infty}; A_t^\pi, R_{t+1,A_t^\pi}|H_t^\pi\right)} \sqrt{\overline{\Gamma}^\pi T} \le \sqrt{\overline{\Gamma}^\pi T \Delta}.$$

□

I PREDICTIVE SAMPLING REGRET BOUND: PROOF OF THEOREM 3

We first establish the following lemma that upper-bounds the information ratio of predictive sampling. Then the proof of Theorem 3 follows directly from the lemma and the general regret bound estalished by Theorem 2.

Lemma 9. *If for all $t \in \mathbb{Z}_+$ and $a \in \mathcal{A}$, $\mathbb{P}(R_{t+1,a} \in \cdot | H_t)$ is almost surely σ_{SG}-sub-Gaussian, then for all $t \in \mathbb{Z}_+$, the information ratio associated with a predictive sampling agent is*

$$\Gamma_t \le 2|\mathcal{A}|\sigma_{SG}^2.$$

Proof. For all $t \in \mathbb{Z}_+$, let

$$\overline{\theta}_t^H = \mathbb{E}\left[R_{t+1} \middle| H_t, R_{t+2:\infty}\right],$$

and $A_{t,*}^H \in \arg\max_{a \in \mathcal{A}} \overline{\theta}_{t,a}^H$ and $R_{t+1,*}^H = R_{t+1,A_{t,*}^H}$. Then for all $t \in \mathbb{Z}_+$, we have

$$\mathbb{P}(A_{t,*}^H \in \cdot | H_t) = \mathbb{P}(A_t \in \cdot | H_t)$$

and $A^H_{t,*} \perp A_t | H_t$.

We begin by establishing a relation using KL-divergence. For all $a, a' \in \mathcal{A}$, and $\lambda \in \mathbb{R}_+$, it follows from the variational form of KL-divergence (Lemma 3 of Appendix B) with $X = \lambda(R_{t+1,a} - \mathbb{E}[R_{t+1,a}|H_t])$ that for all $t \in \mathbb{Z}_+$ and $h \in \mathcal{H}_t$,

$$\begin{aligned}
&\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(R_{t+1,a} \in \cdot | A^H_{t,*} = a', H_t = h) \middle\| \mathbb{P}(R_{t+1,a} \in \cdot | H_t = h)\right) \\
&\geq \mathbb{E}\left[X | H_t = h, A^H_{t,*} = a'\right] - \ln \mathbb{E}[\exp(X) | H_t = h] \\
&\geq \lambda \mathbb{E}\left[R_{t+1,a} - \mathbb{E}[R_{t+1,a}|H_t = h] | H_t = h, A^H_{t,*} = a'\right] - \frac{1}{2}\lambda^2 \sigma^2_{\mathrm{SG}}.
\end{aligned}$$

By maximizing over λ, we obtain

$$\begin{aligned}
&\left(\mathbb{E}\left[R_{t+1,a} | A^H_{t,*} = a', H_t = h\right] - \mathbb{E}\left[R_{t+1,a} | H_t = h\right]\right)^2 \\
&\leq 2\sigma^2_{\mathrm{SG}} \mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(R_{t+1,a} \in \cdot | A^H_{t,*} = a', H_t = h\right) \middle\| \mathbb{P}\left(R_{t+1,a} \in \cdot | H_t = h\right)\right).
\end{aligned} \tag{15}$$

We next establish a relation between this KL-divergence and mutual information. In particular,

$$\begin{aligned}
&\mathbb{I}\left(A^H_{t,*}; A_t, R_{t+1,A_t} | H_t = h\right) \\
&= \mathbb{I}\left(A^H_{t,*}; A_t | H_t = h\right) + \mathbb{I}\left(A^H_{t,*}; R_{t+1,A_t} | A_t, H_t = h\right) \\
&\overset{(a)}{=} \mathbb{I}\left(A^H_{t,*}; R_{t+1,A_t} | A_t, H_t = h\right) \\
&= \sum_{a \in \mathcal{A}} \mathbb{P}(A_t = a | H_t = h) \mathbb{I}\left(A^H_{t,*}; R_{t+1,A_t} | A_t = a, H_t = h\right) \\
&= \sum_{a \in \mathcal{A}} \mathbb{P}(A_t = a | H_t = h) \mathbb{I}\left(A^H_{t,*}; R_{t+1,a} | A_t = a, H_t = h\right) \\
&\overset{(b)}{=} \sum_{a \in \mathcal{A}} \mathbb{P}(A_t = a | H_t = h) \mathbb{I}\left(A^H_{t,*}; R_{t+1,a} | H_t = h\right) \\
&\overset{(c)}{=} \sum_{a \in \mathcal{A}} \mathbb{P}(A_t = a | H_t = h) \\
&\quad \left[\sum_{a' \in \mathcal{A}} \mathbb{P}\left(A^H_{t,*} = a' | H_t = h\right) \mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(R_{t+1,a} \in \cdot | A^H_{t,*} = a', H_t = h\right) \middle\| \mathbb{P}(R_{t+1,a} \in \cdot | H_t = h)\right)\right] \\
&\overset{(d)}{=} \sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \mathbb{P}\left(A^H_{t,*} = a | H_t = h\right) \mathbb{P}\left(A^H_{t,*} = a' | H_t = h\right) \\
&\quad \mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(R_{t+1,a} \in \cdot | A^H_{t,*} = a', H_t = h\right) \middle\| \mathbb{P}(R_{t+1,a} \in \cdot | H_t = h)\right)
\end{aligned} \tag{16}$$

where (a) follows from the fact that $A_t \perp A^H_{t,*} | H_t$, (b) follows from $A_t \perp (A^H_{t,*}, R_{t+1,a}) | H_t$, (c) follows from the KL-divergence representation of mutual information (Lemma 4 of Appendix B), and (d) follows from $\mathbb{P}(A_t \in \cdot | H_t = h) = \mathbb{P}(A^H_{t,*} \in \cdot | H_t = h)$ for all $t \in \mathbb{Z}_+$ and $h \in \mathcal{H}_t$.

Next, we bound the difference between $R_{t+1,A^H_{t,*}}$ and R_{t+1,A_t}. For all $t \in \mathbb{Z}_+$ and $h \in \mathcal{H}_t$, we have

$$
\begin{aligned}
&\mathbb{E}\left[R_{t+1,A^H_{t,*}} - R_{t+1,A_t}\middle|H_t = h\right]^2 \\
&\overset{(a)}{=} \left[\sum_{a\in\mathcal{A}} \mathbb{P}\left(A^H_{t,*} = a\middle|H_t = h\right)\left(\mathbb{E}\left[R_{t+1,a}\middle|A^H_{t,*} = a, H_t = h\right] - \mathbb{E}[R_{t+1,a}|H_t = h]\right)\right]^2 \\
&\overset{(b)}{\le} |\mathcal{A}| \sum_{a\in\mathcal{A}} \mathbb{P}\left(A^H_{t,*} = a|H_t = h\right)^2 \left(\mathbb{E}\left[R_{t+1,a}|A^H_{t,*} = a, H_t = h\right] - \mathbb{E}\left[R_{t+1,a}|H_t = h\right]\right)^2 \\
&\le |\mathcal{A}| \sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}} \mathbb{P}\left(A^H_{t,*} = a|H_t = h\right)\mathbb{P}\left(A^H_{t,*} = a'|H_t = h\right)\left(\mathbb{E}\left[R_{t+1,a}|A^H_{t,*} = a', H_t = h\right] - \mathbb{E}\left[R_{t+1,a}|H_t = h\right]\right)^2 \\
&\overset{(c)}{\le} 2|\mathcal{A}|\sigma^2_{\mathrm{SG}} \sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}} \mathbb{P}\left(A^H_{t,*} = a|H_t = h\right)\mathbb{P}\left(A^H_{t,*} = a'|H_t = h\right) \\
&\quad \mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(R_{t+1,a}\in\cdot|A^H_{t,*} = a', H_t = h\right)\|\mathbb{P}\left(R_{t+1,a}\in\cdot|H_t = h\right)\right) \\
&\overset{(d)}{=} 2|\mathcal{A}|\sigma^2_{\mathrm{SG}}\mathbb{I}\left(A^H_{t,*}; A_t, R_{t+1,A_t}|H_t = h\right),
\end{aligned} \tag{17}
$$

where (a) follows from $A_t \perp R_{t+1,a}|H_t$ and $\mathbb{P}(A_t \in \cdot|H_t = h) = \mathbb{P}(A^H_{t,*} \in \cdot|H_t = h)$, (b) follows from the Cauchy-Bunyakovsky-Schwartz inequality, (c) follows from Equation (15), and (d) follows from Equation (16). Hence,

$$
\begin{aligned}
\mathbb{E}\left[R_{t+1,A^H_{t,*}} - R_{t+1,A_t}\right]^2 &= \mathbb{E}\left[\mathbb{E}\left[R_{t+1,A^H_{t,*}} - R_{t+1,A_t}|H_t\right]\right]^2 \\
&\overset{(a)}{\le} \mathbb{E}\left[\mathbb{E}\left[R_{t+1,A^H_{t,*}} - R_{t+1,A_t}|H_t\right]^2\right] \\
&\overset{(b)}{\le} \mathbb{E}\left[2|\mathcal{A}|\sigma^2_{\mathrm{SG}}\mathbb{I}\left(A^H_{t,*}; A_t, R_{t+1,A_t}|H_t = H_t\right)\right] \\
&= 2|\mathcal{A}|\sigma^2_{\mathrm{SG}}\mathbb{I}\left(A^H_{t,*}; A_t, R_{t+1,A_t}|H_t\right),
\end{aligned} \tag{18}
$$

where (a) follows from Jensen's Inequality and (b) follows from (17).

In addition, for all $t \in \mathbb{Z}_+$,

$$
\begin{aligned}
\mathbb{E}[R_{t+1,*}] &= \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}\left[R_{t+1,a}|R_{t+2:\infty}\right]\right] \\
&= \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}\left[\mathbb{E}\left[R_{t+1,a}|H_t, R_{t+2:\infty}\right]|R_{t+2:\infty}\right]\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}\left[R_{t+1,a}|H_t, R_{t+2:\infty}\right]\middle|R_{t+2:\infty}\right]\right] \\
&= \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}\left[R_{t+1,a}|H_t, R_{t+2:\infty}\right]\right] \\
&= \mathbb{E}\left[R_{t+1,A^H_{t,*}}\right].
\end{aligned} \tag{19}
$$

By the data-processing inequality of mutual information (Lemma 6 of Appendix B), we have for all $t \in \mathbb{Z}_+$,

$$
\mathbb{I}\left(R_{t+2:\infty}; A_t, R_{t+1,A_t}|H_t\right) \ge \mathbb{I}\left(A^H_{t,*}; A_t, R_{t+1,A_t}|H_t\right). \tag{20}
$$

Then it follows from (18), (19) and (20) that for all $t \in \mathbb{Z}_+$,

$$
\Gamma_t = \frac{\mathbb{E}\left[R_{t+1,*} - R_{t+1,A_t}\right]^2_+}{\mathbb{I}\left(R_{t+2:\infty}; A_t, R_{t+1,A_t}|H_t\right)} \le \frac{\mathbb{E}\left[R_{t+1,A^H_{t,*}} - R_{t+1,A_t}\right]^2_+}{\mathbb{I}\left(R_{t+2:\infty}; A_t, R_{t+1,A_t}|H_t\right)} \le \frac{\mathbb{E}\left[R_{t+1,A^H_{t,*}} - R_{t+1,A_t}\right]^2}{\mathbb{I}\left(A^H_{t,*}; A_t, R_{t+1,A_t}|H_t\right)} \le 2|\mathcal{A}|\sigma^2_{\mathrm{SG}}.
$$

□

J REGRET LOWER BOUND: PROOF OF THEOREM 4

Proof. We introduce a modulated Bernoulli bandit (see Example 4 in Section 4.1) with a set of two actions $\mathcal{A} = \{1, 2\}$, and for each $a \in \mathcal{A}$,

$$\theta_{0,a} = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2. \end{cases}$$

We let $q = [1/2, 1]$. Then for all $t \in \mathbb{Z}_+$, the baseline at time t is

$$\begin{aligned}
\mathbb{E}[R_{t+1,*}] &= \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[R_{t+1,a}|R_{t+2:\infty}]\right] \\
&= \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[R_{t+1,a}|R_{-\infty:t}]\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[\theta_{t,a}|\theta_{t-1}]\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[\theta_{t,a}|\theta_{t-1,1}]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[\theta_{t,a}|\theta_{t-1,1}]\middle| H^\pi_{t-1}\right]\right] \\
&\overset{(c)}{=} \mathbb{E}\left[\sum_{a'\in\mathcal{A}} \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[\theta_{t,a}|\theta_{t-1,1}]\middle| H^\pi_{t-1}\right] \mathbb{P}(A^\pi_{t-1} = a'|H^\pi_{t-1})\right],
\end{aligned} \tag{21}$$

where (a) follows from that $(\theta_t : t \in \mathbb{Z})$ follows a Markov process, and that $R_{t+1} = \theta_t$, (b) follows from $q_2 = 1$, and (c) from that A^π_{t-1} is independent of θ_{t-1} conditioned on H^π_{t-1}.

For all policy π and all $t \in \mathbb{Z}_+$, the reward collected at time t is upper-bounded by

$$\begin{aligned}
\mathbb{E}\left[R_{t+1,A^\pi_t}\right] &= \mathbb{E}\left[\mathbb{E}\left[R_{t+1,A^\pi_t}|H^\pi_t\right]\right] \\
&= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{E}\left[R_{t+1,a}|H^\pi_t\right] \mathbb{P}(A^\pi_t = a|H^\pi_t)\right] \\
&\leq \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}\left[R_{t+1,a}|H^\pi_t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[R_{t+1,a}|H^\pi_t]\middle| H^\pi_{t-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[R_{t+1,a}|H^\pi_{t-1}, A^\pi_{t-1}, R_{t,A^\pi_{t-1}}]\middle| H^\pi_{t-1}\right]\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{a'\in\mathcal{A}} \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[R_{t+1,a}|H^\pi_{t-1}, R_{t,a'}]\middle| H^\pi_{t-1}\right] \mathbb{P}(A^\pi_{t-1} = a'|H^\pi_{t-1})\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{a'\in\mathcal{A}} \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[\theta_{t,a}|H^\pi_{t-1}, \theta_{t-1,a'}]\middle| H^\pi_{t-1}\right] \mathbb{P}(A^\pi_{t-1} = a'|H^\pi_{t-1})\right],
\end{aligned} \tag{22}$$

where (a) follows from that A^π_{t-1} is independent of R_t conditioned on H^π_{t-1}, and (b) from $R_{t+1} = \theta_t$. Observe that for all $t \in \mathbb{Z}_+$, the term $\mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[\theta_{t,a}|H^\pi_{t-1}, \theta_{t-1,a'}]\middle| H^\pi_{t-1}\right]$ in (22) for each of $a' \in \mathcal{A} = \{1, 2\}$ can be derived or upper-bounded as follows:

$$\mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[\theta_{t,a}|H^\pi_{t-1}, \theta_{t-1,1}]\middle| H^\pi_{t-1}\right] = \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[\theta_{t,a}|\theta_{t-1,1}]\middle| H^\pi_{t-1}\right], \tag{23}$$

and

$$\begin{aligned}\mathbb{E}\left[\max_{a\in\mathcal{A}}\mathbb{E}[\theta_{t,a}|H^\pi_{t-1},\theta_{t-1,2}]\middle|H^\pi_{t-1}\right] &\overset{(a)}{=} \mathbb{E}\left[\max_{a\in\mathcal{A}}\mathbb{E}[\theta_{t,a}|H^\pi_{t-1}]\middle|H^\pi_{t-1}\right]\\ &= \max_{a\in\mathcal{A}}\mathbb{E}[\theta_{t,a}|H^\pi_{t-1}]\\ &\overset{(b)}{=} \max_{a\in\mathcal{A}}\mathbb{E}\left[\mathbb{E}\left[\theta_{t,a}|\theta_{t-2}\right]|H^\pi_{t-1}\right]\\ &\overset{(c)}{\leq} \mathbb{E}\left[\max_{a\in\mathcal{A}}\mathbb{E}\left[\theta_{t,a}|\theta_{t-2}\right]\middle|H^\pi_{t-1}\right]\\ &\overset{(d)}{=} \mathbb{E}\left[\max_{a\in\mathcal{A}}\mathbb{E}\left[\theta_{t,a}|\theta_{t-2,1}\right]\middle|H^\pi_{t-1}\right],\end{aligned} \tag{24}$$

where (a) follows from $q_2 = 1$, (b) follows from that θ_t is independent of H^π_{t-1} conditioned on θ_{t-2} (recall that $H^\pi_{t-1} = (A^\pi_0, R_{1,A^\pi_0}, ...A^\pi_{t-2}, R_{t-1,A^\pi_{t-2}}) = (A^\pi_0, \theta_{0,A^\pi_0}, ...A^\pi_{t-2}, \theta_{t-2,A^\pi_{t-2}})$), (c) from Jensen's inequality, and (d) again from $q_2 = 1$. Subtracting (22) from (21), we establish a lower bound on the instantaneous regret:

$$\begin{aligned}\mathbb{E}[R_{t+1,*} - R_{t+1,A^\pi_t}] &\geq \mathbb{E}\left[\sum_{a'\in\mathcal{A}}\mathbb{E}\left[\max_{a\in\mathcal{A}}\mathbb{E}[\theta_{t,a}|\theta_{t-1,1}] - \max_{a\in\mathcal{A}}\mathbb{E}[\theta_{t,a}|H^\pi_{t-1},\theta_{t-1,a'}]\middle|H^\pi_{t-1}\right]\mathbb{P}(A^\pi_{t-1}=a'|H^\pi_{t-1})\right]\\ &\overset{(a)}{=} \mathbb{E}\left[\mathbb{E}\left[\max_{a\in\mathcal{A}}\mathbb{E}[\theta_{t,a}|\theta_{t-1,1}] - \max_{a\in\mathcal{A}}\mathbb{E}[\theta_{t,a}|\theta_{t-2,1}]\middle|H^\pi_{t-1}\right]\mathbb{P}(A^\pi_{t-1}=2|H^\pi_{t-1})\right]\\ &\overset{(b)}{=} \mathbb{E}\left[\frac{1}{16}\mathbb{P}(A^\pi_{t-1}=2|H^\pi_{t-1})\right]\\ &= \frac{1}{16}\mathbb{P}(A^\pi_{t-1}=2),\end{aligned} \tag{25}$$

where (a) follows from (23) and (24), and (b) from computing the conditional expectation, which turns out to be independent of H^π_{t-1}.

Below we derive another lower bound on the instantaneous regret. First, observe that for all policy π and all $t \in \mathbb{Z}_+$, the reward collected at time t is upper-bounded by:

$$\begin{aligned}\mathbb{E}\left[R_{t+1,A^\pi_t}\right] &= \mathbb{E}\left[R_{t+1,A^\pi_t}|A^\pi_t=1\right]\mathbb{P}\left(A^\pi_t=1\right) + \mathbb{E}\left[R_{t+1,A^\pi_t}|A^\pi_t=2\right]\mathbb{P}\left(A^\pi_t=2\right)\\ &\leq \mathbb{E}\left[R_{t+1,1}|A^\pi_t=1\right]\mathbb{P}\left(A^\pi_t=1\right) + \mathbb{P}\left(A^\pi_t=2\right)\\ &\leq \mathbb{E}\left[R_{t+1,1}\right] + \mathbb{P}\left(A^\pi_t=2\right),\end{aligned}$$

where both inequalities follow from that rewards are bounded in $[0,1]$. Therefore, for all policy π and all $t \in \mathbb{Z}_+$, the instantaneous regret can be lower-bounded as follows:

$$\begin{aligned}\mathbb{E}\left[R_{t+1,*} - R_{t+1,A^\pi_t}\right] &\geq \mathbb{E}\left[R_{t+1,*}\right] - \mathbb{E}\left[R_{t+1,1}\right] - \mathbb{P}\left(A^\pi_t=2\right)\\ &= \frac{5}{8} - \frac{1}{2} - \mathbb{P}(A^\pi_t=2) = \frac{1}{8} - \mathbb{P}(A^\pi_t=2).\end{aligned} \tag{26}$$

Incorporating the two lower bounds on instantaneous regret established in (25) and (26), respectively, we derive a lower bound on the cumulative regret: for all policy π, and $T \in \mathbb{Z}_{++}$, $T \geq 2$,

$$\begin{aligned}\mathrm{Regret}(T;\pi) &\geq \max\left\{\sum_{t=0}^{T-2}\frac{1}{16}\mathbb{P}(A^\pi_t=2), \sum_{t=0}^{T-2}\left[\frac{1}{8} - \mathbb{P}(A^\pi_t=2)\right]\right\}\\ &\geq \frac{16}{17}\sum_{t=0}^{T-2}\frac{1}{16}\mathbb{P}(A^\pi_t=2) + \frac{1}{17}\sum_{t=0}^{T-2}\left[\frac{1}{8} - \mathbb{P}(A^\pi_t=2)\right]\\ &= \frac{1}{136}(T-1)\\ &\geq \frac{1}{272}T.\end{aligned} \tag{27}$$

For all policy π, and $T = 1$,

$$\begin{aligned}\text{Regret}(T;\pi) &= \mathbb{E}[R_{1,*}] - \mathbb{E}[R_{1,A_0^\pi}] \\ &\geq \mathbb{E}[R_{1,*}] - \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[R_{t+1,a}]\right] \\ &= \frac{5}{8} - \frac{1}{2} = \frac{1}{8} \geq \frac{1}{272}T. \end{aligned} \tag{28}$$

Combining (27) and (28), we complete the proof. □

K PREDICTIVE SAMPLING REGRET BOUND IN A MODULATED BERNOULLI BANDIT: PROOF OF COROLLARY 1

We first introduce Lemma 10, which establishes an upper bound on Δ_0 and an upper bound on Δ_t that holds uniformly over all $t \in \mathbb{Z}_{++}$ for a modulated Bernoulli bandit. Corollary 1 follows directly from Lemma 10 and Theorem 3.

Lemma 10. *In a modulated Bernoulli bandit environment,*

$$\begin{aligned}\Delta_0 &\leq \sum_{a\in\mathcal{A}} (1-q_a)\mathbb{H}(\theta_{0,a}), \\ \Delta_t &\leq \min\left\{\sum_{a\in\mathcal{A}} (1-q_a)\mathbb{H}(\theta_{0,a}), \sum_{a\in\mathcal{A}} [q_a\mathbb{H}(\theta_{0,a}) + \mathbb{H}(q_a)]\right\}, \textit{ for all } t \in \mathbb{Z}_{++},\end{aligned}$$

where $\mathbb{H}(q_a)$ denotes the entropy of a Bernoulli(q_a) random variable.

Proof. We first describe an alternative formulation of the modulated Bernoulli bandit environment. For all $a \in \mathcal{A}$, $\theta_{0,a} = X_{0,a}$ and, for all $a \in \mathcal{A}$ and $t \in \mathbb{Z}_+$,

$$\theta_{t+1,a} = \begin{cases} X_{t+1,a} & \text{if } B_{t+1,a} = 1 \\ \theta_{t,a} & \text{if } B_{t+1,a} = 0. \end{cases}$$

where $(B_{t,a} : t \in \mathbb{Z}_{++})$ is an i.i.d. Bernoulli(q_a) process and $(X_{t,a} : t \in \mathbb{Z}_+)$ is an i.i.d. process with discrete range. With this formulation, observe that for all $t \in \mathbb{Z}_+$ and $a \in \mathcal{A}$,

$$\theta_{t+1,a} = (1 - B_{t+1,a})\theta_{t,a} + B_{t+1,a}X_{t+1,a}. \tag{29}$$

We derive the mutual information between θ_1 and θ_0:

$$\begin{aligned}\mathbb{I}(\theta_1;\theta_0) &= \sum_{a\in\mathcal{A}} \mathbb{I}(\theta_{1,a};\theta_{0,a}) \\ &= \sum_{a\in\mathcal{A}} (\mathbb{H}(\theta_{1,a}) - \mathbb{H}(\theta_{1,a}|\theta_{0,a})) \\ &\leq \sum_{a\in\mathcal{A}} (\mathbb{H}(\theta_{0,a}) - \mathbb{H}(\theta_{1,a}|\theta_{0,a}, B_{1,a})) \\ &\overset{(a)}{=} \sum_{a\in\mathcal{A}} (\mathbb{H}(\theta_{0,a}) - \mathbb{H}((1-B_{1,a})\theta_{0,a} + B_{1,a}X_{1,a}|\theta_{0,a}, B_{1,a})) \\ &= \sum_{a\in\mathcal{A}} (\mathbb{H}(\theta_{0,a}) - \mathbb{H}(B_{1,a}X_{1,a}|\theta_{0,a}, B_{1,a})) \\ &= \sum_{a\in\mathcal{A}} (\mathbb{H}(\theta_{0,a}) - \mathbb{H}(B_{1,a}X_{1,a}|B_{1,a})) \\ &= \sum_{a\in\mathcal{A}} (\mathbb{H}(\theta_{0,a}) - q_a\mathbb{H}(X_{1,a})) \\ &= \sum_{a\in\mathcal{A}} (\mathbb{H}(\theta_{0,a}) - q_a\mathbb{H}(\theta_{0,a})) \\ &= \sum_{a\in\mathcal{A}} (1-q_a)\mathbb{H}(\theta_{0,a}), \end{aligned} \tag{30}$$

where (a) follows from (29).

Now we derive Δ_0:

$$\Delta_0 = \mathbb{I}(R_{2:\infty}; \mathcal{E}_0) \overset{(a)}{\leq} \mathbb{I}(R_{2:\infty}; \theta_0) \leq \mathbb{I}(\theta_1; \theta_0) \overset{(b)}{=} \sum_{a \in \mathcal{A}} (1 - q_a) \mathbb{H}(\theta_{0,a}),$$

where (a) follows from $R_{2:\infty} \perp \mathcal{E}_0 | \theta_0$, where $\mathcal{E}_0 = \mathbb{P}(R_{2:\infty} \in \cdot | R_{-\infty:1})$, and (b) follows from (30).

For all $t \in \mathbb{Z}_+$,

$$\begin{aligned}
\Delta_t &= \mathbb{I}(R_{t+2:\infty}; \mathcal{E}_t | H_t) - \mathbb{I}(R_{t+1:\infty}; \mathcal{E}_{t-1} | H_t) \\
&\overset{(a)}{\leq} \mathbb{I}(R_{t+2:\infty}; \mathcal{E}_t) \\
&\overset{(b)}{\leq} \mathbb{I}(R_{t+2:\infty}; \theta_t) \\
&\leq \mathbb{I}(\theta_{t+1}; \theta_t) \\
&\overset{(c)}{=} \mathbb{I}(\theta_1; \theta_0) \\
&\overset{(d)}{\leq} \sum_{a \in \mathcal{A}} (1 - q_a) \mathbb{H}(\theta_{0,a}),
\end{aligned}$$

where (a) follows from Lemma 7 of Appendix B and that $R_{t+2:\infty} \perp H_t | \mathcal{E}_t$, where $\mathcal{E}_t = \mathbb{P}(R_{t+2:\infty} \in \cdot | R_{-\infty:t+1})$, (b) from $R_{t+2:\infty} \perp \mathcal{E}_t | \theta_t$, (c) from that $(\theta_t : t \in \mathbb{Z})$ is stationary, and (d) from (30).

In addition, for all $t \in \mathbb{Z}_{++}$,

$$\begin{aligned}
\Delta_t &= \mathbb{I}(R_{t+2:\infty}; \mathcal{E}_t | H_t) - \mathbb{I}(R_{t+1:\infty}; \mathcal{E}_{t-1} | H_t) \\
&\overset{(a)}{=} \mathbb{I}(R_{t+2:\infty}; \theta_t | H_t) - \mathbb{I}(R_{t+2:\infty}; \theta_t | H_t, \mathcal{E}_t) - \mathbb{I}(R_{t+1:\infty}; \theta_{t-1} | H_t) + \mathbb{I}(R_{t+1:\infty}; \theta_{t-1} | H_t, \mathcal{E}_{t-1}) \\
&\overset{(b)}{=} \mathbb{I}(R_{t+2:\infty}; \theta_t | H_t) - \mathbb{I}(R_{t+2:\infty}; \theta_t | \mathcal{E}_t) - \mathbb{I}(R_{t+1:\infty}; \theta_{t-1} | H_t) + \mathbb{I}(R_{t+1:\infty}; \theta_{t-1} | \mathcal{E}_{t-1}) \\
&\overset{(c)}{=} \mathbb{I}(R_{t+2:\infty}; \theta_t | H_t) - \mathbb{I}(R_{t+1:\infty}; \theta_{t-1} | H_t),
\end{aligned}$$

where (a) follows from $R_{t+2:\infty} \perp (H_t, \mathcal{E}_t) | \theta_t$ and $R_{t+1:\infty} \perp (H_t, \mathcal{E}_{t-1}) | \theta_{t-1}$, (b) from $(R_{t+2:\infty}, \theta_t) \perp H_t | \mathcal{E}_t$ and $(R_{t+1:\infty}, \theta_{t-1}) \perp H_t | \mathcal{E}_{t-1}$, and (c) from that $(\theta_t : t \in \mathbb{Z})$ is stationary.

By (29), for all $t \in \mathbb{Z}_{++}$,

$$\begin{aligned}
\Delta_t &\leq \mathbb{I}(R_{t+2:\infty}; \theta_t | H_t) - \mathbb{I}(R_{t+1:\infty}; \theta_{t-1} | H_t) \\
&\leq \mathbb{I}(R_{t+1:\infty}; \theta_{t-1}, B_t, B_t X_t | H_t) - \mathbb{I}(R_{t+1:\infty}; \theta_{t-1} | H_t) \\
&\leq \mathbb{I}(R_{t+1:\infty}; B_t, B_t X_t | H_t, \theta_{t-1}) \\
&\leq \mathbb{H}(B_t, B_t X_t | H_t, \theta_{t-1}) \\
&\leq \mathbb{H}(B_t, B_t X_t) \\
&= \sum_{a \in \mathcal{A}} \left(\mathbb{H}(B_{t+1,a}) + \mathbb{H}(B_{t+1,a} X_{t+1,a} | B_{t+1,a}) \right) \\
&= \sum_{a \in \mathcal{A}} \left(\mathbb{H}(B_{1,a}) + \mathbb{H}(B_{t+1,a} X_{t+1,a} | B_{t+1,a}) \right) \\
&= \sum_{a \in \mathcal{A}} \left(\mathbb{H}(B_{1,a}) + q_a \mathbb{H}(X_{t+1,a}) \right) \\
&= \sum_{a \in \mathcal{A}} \left(\mathbb{H}(q_a) + q_a \mathbb{H}(\theta_{0,a}) \right).
\end{aligned}$$

□

L NONNEGATIVITY OF REGRET

It is natural to require that the regret is nonnegative. The following theorem establishes that the regret is nonnegative under mild conditions.

Theorem 5. *Let $(R_{t+1} : t \in \mathbb{Z})$ be an environment generated by $P_{-\infty:\infty}$. If for all $t, t' \in \mathbb{Z}$, and $n \in \mathbb{Z}_+$, $\mathbb{P}(P_{t:t+n} \in \cdot) = \mathbb{P}(P_{t':t'+n} \in \cdot)$, and $\mathbb{P}(P_{t:t+n} \in \cdot) = \mathbb{P}(P_{t:t-n} \in \cdot)$, then for all policies π, and $T \in \mathbb{Z}_+$,*

$$\mathrm{Regret}(T; \pi) \geq 0.$$

Proof. We have for all policies π, and $T \in \mathbb{Z}_+$,

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,A_t^\pi}\right] &= \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbb{E}\left[R_{t+1,A_t^\pi} | R_{1:t}, A_{0:t-1}^\pi\right]\right] \\
&\overset{(a)}{=} \sum_{t=0}^{T-1} \mathbb{E}\left[\sum_{a \in \mathcal{A}} \mathbb{E}\left[R_{t+1,a} | R_{1:t}, A_{0:t-1}^\pi\right] \mathbb{P}(A_t^\pi = a | R_{1:t}, A_{0:t-1}^\pi)\right] \\
&\leq \sum_{t=0}^{T-1} \mathbb{E}\left[\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{t+1,a} | R_{1:t}, A_{0:t-1}^\pi\right]\right] \\
&\overset{(b)}{=} \sum_{t=0}^{T-1} \mathbb{E}\left[\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{t+1,a} | R_{1:t}\right]\right], \qquad (31)
\end{aligned}$$

where (a) follows from the fact that A_t^π is independent of any other random variables conditioning on $H_t^\pi = (A_0^\pi, R_{1,A_0^\pi}, ..., A_{t-1}^\pi, R_{t,A_{t-1}^\pi})$, and (b) follows from recursively applying the same reasoning. By the conditions stated in the theorem, we have

$$\mathbb{P}(R_{1:T} \in \cdot) = \mathbb{P}(R_{T:2T-1} \in \cdot) = \mathbb{P}(R_{T:1} \in \cdot),$$

which implies that $\mathbb{P}(R_{1:t+1} \in \cdot) = \mathbb{P}(R_{T:T-t} \in \cdot)$ for all $t, T \in \mathbb{Z}_+$, $t < T$. Therefore,

$$\mathbb{P}(\mathbb{E}\left[R_{t+1} | R_{1:t}\right] \in \cdot) = \mathbb{P}(\mathbb{E}\left[R_{T-t} | R_{T:T-t+1}\right] \in \cdot)$$

for all $t, T \in \mathbb{Z}_+$, $t < T$. Hence,

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{t+1,a} | R_{1:t}\right]\right] = \sum_{t=0}^{T-1} \mathbb{E}\left[\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{T-t,a} | R_{T:T-t+1}\right]\right] = \sum_{t=0}^{T-1} \mathbb{E}\left[\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{t+1,a} | R_{t+2:T}\right]\right]. \qquad (32)$$

By Jensen's Inequality, we have for all $T \in \mathbb{Z}_+$, and $t \in \mathbb{Z}_+$, $t < T$,

$$\begin{aligned}
\mathbb{E}\left[\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{t+1,a} | R_{t+2:T}\right]\right] &= \mathbb{E}\left[\max_{a \in \mathcal{A}} \mathbb{E}\left[\mathbb{E}\left[R_{t+1,a} | R_{t+2:\infty}\right] | R_{t+2:T}\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{t+1,a} | R_{t+2:\infty}\right] \middle| R_{t+2:T}\right]\right] \\
&= \mathbb{E}\left[\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{t+1,a} | R_{t+2:\infty}\right]\right] \\
&= \mathbb{E}\left[R_{t+1,*}\right]. \qquad (33)
\end{aligned}$$

Combining (31), (32), and (33), we conclude that for all policies π, and $T \in \mathbb{Z}_+$,

$$\mathrm{Regret}(T; \pi) = \sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,*}\right] - \sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,A_t^\pi}\right] \geq 0.$$

□

Remark. The conditions stated in the theorem find close analogues in the theory of Markov chains. Indeed, the conditions on the sequence $P_{-\infty:\infty}$ are referred to as "stationarity" and "time-reversibility" in the Markov chain theory.

That said, the conditions hold in all stationary bandit environments and a wide range of nonstationary bandit environments: in particular, with a Markov bandit (Anantharam et al., 1987; Ortner et al., 2014), P_t transitions following a Markov process—and thus, the assumptions hold if the Markov process is both stationary and time-reversible. The modulated Bernoulli bandit environments (Example 4 in Section 4.1; (Mellor and Shapiro, 2013)) and AR(1) bandit environments (Example 2 in Section 4.1) are two such examples, because P_t is determined by θ_t, and $(\theta_{t,a} : t \in \mathbb{Z})$ is a time-reversible and time-homogenous Markov chain in steady-state. The AR(1) logistic bandit environments serves as an additional example, with which P_t is determined by α_t, and $(\alpha_{t,a} : t \in \mathbb{Z})$ is a time-reversible and time-homogenous Markov chain in steady-state.

It is worth noting that the first condition requires that the distribution of a sequence of P_t's is invariant when the sequence is shifted in time. Note that although the distribution of P_t is the same for all $t \in \mathbb{Z}_+$, P_t can be different across time and the environment can thus be nonstationary.

M MORE ON THE VARIATIONS OF PS AND TS IN AR(1) LOGISTIC BANDITS

M.1 Standard Laplace Approximation and Incremental Laplace Approximation in Stationary Logistic Bandits

This section presents numerical experiments we conduct to show that incremental Laplace approximation is comparable with standard Laplace approximation in stationary logistic bandits. To compare the incremental Laplace approximation with the standard Laplace approximation, we conduct experiments on Thompson sampling agents interacting with a stationary logistic bandit introduced below:

Example 5 (stationary logistic bandit). *In a stationary logistic bandit, each reward distribution $P_a = \mathbb{P}(R_{t+1,a} \in \cdot|P)$ is Bernoulli. Its mean $\mathbb{E}[R_{t+1,a}|P]$ is determined by a random variable $\alpha \in \mathbb{R}^d$, where $d \in \mathbb{Z}_{++}$ is known, and $\phi_a \in \mathbb{R}^d$, a known feature vector associated with action $a \in \mathcal{A}$. In particular, the mean reward $\mathbb{E}[R_{t+1,a}|P] = \frac{\exp(\alpha^\top \phi_a)}{1+\exp(\alpha^\top \phi_a)}$. The variable α has a Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$, where $\mu_0 \in \mathbb{R}^d$ and $\Sigma_0 \in \mathcal{S}_+^d$. Here we use $\mathcal{S}_+^d$ to denote the set of all $d \times d$ positive semi-definite matrices.*

We compare the performance of the agents in using incremental Laplace approximation and the standard Laplace approximation, respectively, in approximating the posterior distribution of α at each timestep t. We run three sets of experiments, in which each coordinate of α is standard Gaussian and independent of the rest of the coordinates, and the number of actions are 2, 3, and 3, with $\phi \in \{\phi^1, \phi^2, \phi^3\}$, where

$$\phi^1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \phi^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ and } \phi^3 = \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0 & 0.9 & 0.1 \\ 0.1 & 0 & 0.9 \end{bmatrix},$$

respectively. Here, we use ϕ to denote the matrix where the a-th row corresponds to the row feature vector $\phi_a^\top$, for all $a \in \mathcal{A}$.

Figure 3 plots the cumulative regret over 1000 timesteps, averaged over 200 simulations, incurred by Thompson sampling agents using standard Laplace approximation and incremental Laplace approximation, respectively, in approximating the posteriors. We observe that the performances of the two methods are comparable across all experiments.

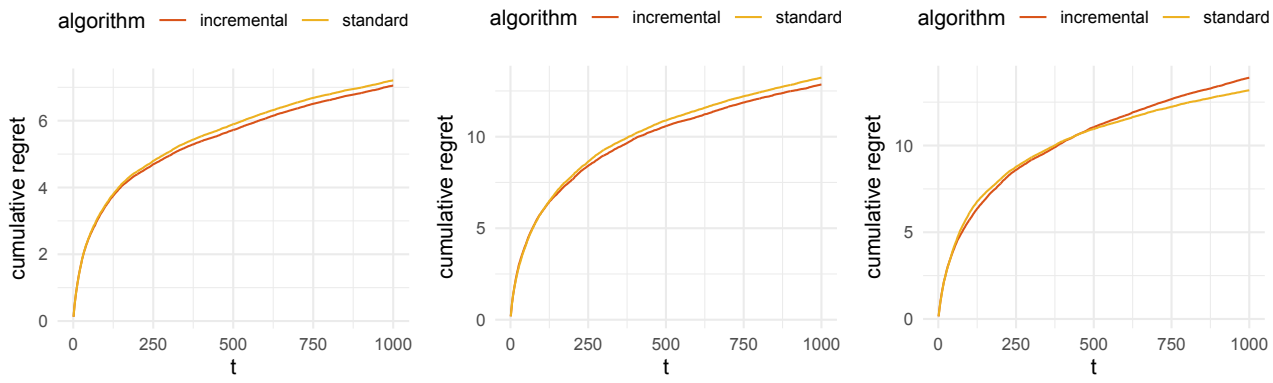
M.2 Incremental Laplace Approximation in Nonstationary Logistic Bandits

Below we present how we can efficiently implement incremental Laplace approximation in approximating $\mathbb{P}(\alpha_t \in \cdot|H_t^\pi)$ using $\mathcal{N}(\mu_t, \Sigma_t)$ in nonstationary AR(1) logistic bandits. At each timestep, the mean μ_t minimizes the following objective:

$$\mu_t \leftarrow \min_\alpha \left\{ \frac{1}{2}(\alpha - \mu_{t-1})^\top \Sigma_{t-1}^{-1}(\alpha - \mu_{t-1}) - R_{t+1,A_t}\phi_{A_t}^\top \alpha + \log\left(1 + \exp\left(\phi_{A_t}^\top \alpha\right)\right) \right\}, \tag{34}$$

and the variance can be derived as follows:

$$\Sigma_t \leftarrow \left[\Sigma_{t-1}^{-1} + \frac{\exp(\phi_{A_t}^\top \mu_t)}{\left(1 + \exp\left(\phi_{A_t}^\top \mu_t\right)\right)^2} \phi_{A_t}\phi_{A_t}^\top \right]^{-1}, \tag{35}$$



(a) bandit with feature matrix ϕ^1 (b) bandit with feature matrix ϕ^2 (c) bandit with feature matrix ϕ^3

Figure 3: Cumulative regret incurred by Thompson sampling agents using standard Laplace approximation and incremental Laplace approximation.

then an additional step is carried out:

$$\mu_t \leftarrow A\mu_t \text{ and } \Sigma_t \leftarrow A\Sigma_t A^\top + V,$$

where A is a diagonal matrix whose a-th entry along its diagonal is γ_a, and V is a diagonal matrix whose a-th entry along its diagonal is δ_a^2.

M.3 Detailed Derivation of Predictive Sampling

We provide detailed procedures to execute a variation of predictive sampling in AR(1) logistic bandits. Recall that $R_{t+1} \sim \text{Bernoulli}\left(\frac{\exp(\phi\alpha_t)}{1+\exp(\phi\alpha_t)}\right)$. We instead pretend that the rewards are Gaussian distributed according to: $\tilde{R}_{t+1} \sim \mathcal{N}\left(\frac{1}{2}e + \frac{1}{4}\phi\alpha_t, \frac{1}{16}I\right)$, where e is an all-one vector and ϕ is the matrix whose a-th row is $\phi_a^\top$. We present the following algorithm for nonstationary logistic bandits, that is designed based on incremental Laplace approximation and Gaussian imagination.

Algorithm 5: predictive sampling (PS)

for $t = 0, 1, \ldots, T-1$ **do**
  sample: $\hat{R}^{(t)}_{t+2:\infty} \sim \mathbb{P}(R_{t+2:\infty} \in \cdot | H_t)$
  estimate: $\hat{\theta}_t = \mathbb{E}[R_{t+1} | H_t, R_{t+2:\infty} \leftarrow \hat{R}^{(t)}_{t+2:\infty}]$
  select: $A_t \in \arg\max_{a \in \mathcal{A}} \hat{\theta}_{t,a}$
  observe: R_{t+1,A_t}

Input. $\phi, A, V, \mu_0, \Sigma_0, n$.

Step 1. At each timestep t, derive the mean μ_t that minimizes the following objective:

$$\mu_t \leftarrow \min_{\alpha} \left\{ \frac{1}{2}(\alpha - \mu_{t-1})^\top \Sigma_{t-1}^{-1}(\alpha - \mu_{t-1}) - R_{t+1,A_t}\phi_{A_t}^\top \alpha + \log\left(1 + \exp\left(\phi_{A_t}^\top \alpha\right)\right) \right\}, \tag{36}$$

and the variance:

$$\Sigma_t \leftarrow \left[\Sigma_{t-1}^{-1} + \frac{\exp(\phi_{A_t}^\top \mu_t)}{\left(1 + \exp\left(\phi_{A_t}^\top \mu_t\right)\right)^2} \phi_{A_t}\phi_{A_t}^\top \right]^{-1}. \tag{37}$$

Step 2. Update

$$\mu_t \leftarrow A\mu_t \text{ and } \Sigma_t \leftarrow A\Sigma_t A^\top + V.$$

Step 3. Sample $\hat{\alpha}_t$ from $\mathcal{N}(\mu_t', \Sigma_t')$, where $\mu_t' = \mu_t$, $\Sigma_t' = \Sigma_{12}^{(t)}\Sigma_{22}^{(t)-1}\Sigma_{21}^{(t)}$, and

$$\Sigma_{21} = \frac{1}{4}\begin{bmatrix} \phi A\Sigma_t \\ \phi A^2\Sigma_t \\ ... \\ \phi A^n\Sigma_t \end{bmatrix}, \Sigma_{12} = \Sigma_{21}^\top,$$

$$\Sigma_{22} = \frac{1}{16}\begin{bmatrix} \phi\tilde{\Sigma}_{t+1}^{(t)}\phi^\top + I & \phi\tilde{\Sigma}_{t+1}^{(t)}A^\top\phi^\top & \phi\tilde{\Sigma}_{t+1}^{(t)}A^{2\top}\phi^\top & \dots & \phi\tilde{\Sigma}_{t+1}^{(t)}A^{n-1\top}\phi^\top \\ \phi A\tilde{\Sigma}_{t+1}^{(t)}\phi^\top & \phi\tilde{\Sigma}_{t+2}^{(t)}\phi^\top + I & \phi\tilde{\Sigma}_{t+2}^{(t)}A^\top\phi^\top & \dots & \phi\tilde{\Sigma}_{t+2}^{(t)}A^{n-2\top}\phi^\top \\ \phi A^2\tilde{\Sigma}_{t+1}^{(t)}\phi^\top & \phi A\tilde{\Sigma}_{t+2}^{(t)}\phi^\top & \phi\tilde{\Sigma}_{t+3}^{(t)}\phi^\top + I & \dots & \phi\tilde{\Sigma}_{t+3}^{(t)}A^{n-3\top}\phi^\top \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi A^{n-1}\tilde{\Sigma}_{t+1}^{(t)}\phi & \phi A^{n-2}\tilde{\Sigma}_{t+2}^{(t)}\phi & \phi A^{n-3}\tilde{\Sigma}_{t+3}^{(t)}\phi & \dots & \phi\tilde{\Sigma}_{t+n}^{(t)}\phi^\top + I \end{bmatrix}.$$

The matrices $\tilde{\Sigma}_t^{(t)}, \tilde{\Sigma}_{t+1}^{(t)}, ..., \tilde{\Sigma}_{t+n}^{(t)}$ satisfy the following recursion:

$$\tilde{\Sigma}_t^{(t)} = \Sigma_t,$$
$$\tilde{\Sigma}_{t+i+1}^{(t)} = A\tilde{\Sigma}_{t+i}^{(t)}A^\top + V,\ i = 0, 1, 2, ..., n-1.$$

Step 4. Estimate $\hat{\theta}_t = \frac{1}{2}e + \frac{1}{4}\phi\hat{\alpha}_t$.

Step 5. Select $A_t \in \arg\max_{a\in\mathcal{A}} \hat{\theta}_{t,a}$.

N COMPARISON WITH STATE-OF-THE-ART ALGORITHMS

This section presents experiments we conduct to compare the performance of predictive sampling with state-of-the-art algorithms designed for nonstationary bandit environments, including Rexp3 (Besbes et al., 2019), discounted UCB (Garivier and Moulines, 2008; Kocsis and Szepesvári, 2006), and sliding-window UCB (Garivier and Moulines, 2008).

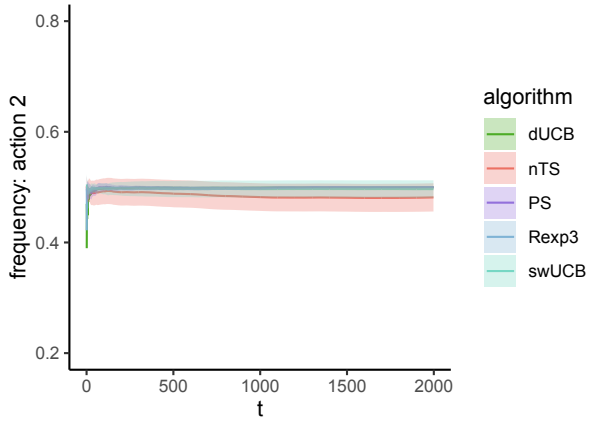
We first introduce the environment in which we conduct the experiments. We design a set of bandit environments that differ from AR(1) bandits only in that the rewards are truncated to $[0, 1]$. The set of environments are designed such that how quickly the information acquired by selecting an action $a \in \mathcal{A}$ loses relevance is determined by the variable γ_a. In addition, they are designed such that the rewards are bounded, the same as most of the experiment settings in frequentist nonstationary bandit learning literature.

We next introduce a set of state-of-the-art algorithms with which we compare predictive sampling. Similar to Thompson sampling and its variations, a large segment of state-of-the-art algorithms focus on heuristics on how the nonstationarity of past rewards affects the inference on current reward distribution and ignores future nonstationarity. Popular examples of such heuristics include using a fixed-length sliding-window, weighing data by recency, and periodic restarts. We choose one algorithm focusing on each of the three heuristics. In addition, we use a naive Thompson sampling agent, who pretends that the environment is stationary, as a baseline. Below we briefly describe each of the aforementioned algorithms:

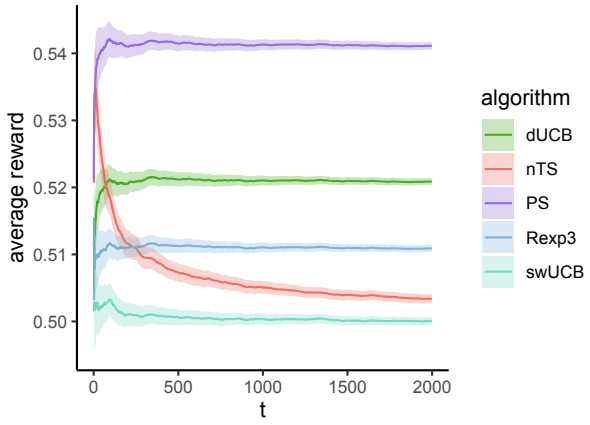
- Rexp3 uses Exp3 as a subroutine and restarts it periodically;
- discounted UCB uses UCB1 as a subroutine and discounts the effect of past rewards on estimating current reward distribution by weighing past data according to recency;
- sliding-window UCB maintains a sliding-window of fixed size and uses UCB1 as a subroutine;
- naive TS pretends that the environment is stationary and proceeds with inference.

We run Rexp3, discounted UCB and sliding-window UCB and naive TS as they are, and run predictive sampling pretending that the environment is the AR(1) bandit before truncation. The parameters of Rexp3 are chosen according to Theorem 2 of (Besbes et al., 2019), where the "variation budget" is assumed to be known in advance for each simulation; the parameters in discounted UCB and sliding-window UCB are chosen according to Remark 3 and Remark 9 of (Garivier and Moulines, 2008), respectively, where "the number of breakpoints" is assumed to be known in advance.

Average reward and action frequency We conduct two sets of experiments. The first environment is a two-armed bandit with $\mathcal{A} = \{1, 2\}$, $c = [0.5, 0.5]$, $\gamma = (0.85, 0.85)$, $\delta_a = 0.15(1 - \gamma_a^2)$ for $a \in \mathcal{A}$, and $\sigma = [0.1, 0.1]$. The two actions can be thought of as "changing equally quickly". Figure 4a and 4b plot the average frequency of selecting action 1, and the average reward collected: although all agents select each action half of the time in the long run, the predictive sampling agent collects more rewards because it explores less accounting for future nonstationarity of the environment.



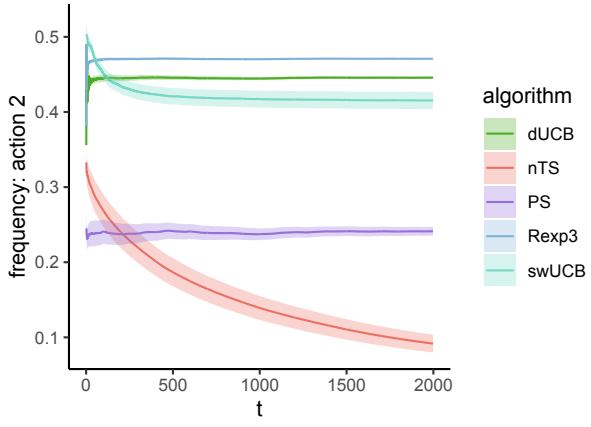
(a) Frequencies of selecting action 2



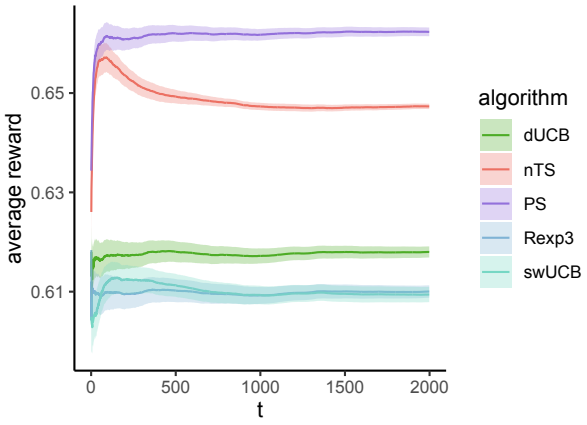
(b) Average reward collected by each agent

Figure 4: Predictive sampling and state-of-the-art algorithms in a two-armed bandit with $c = [0.5, 0.5]$, $\gamma = (0.85, 0.85)$, $\delta_a = 0.15(1 - \gamma_a^2)$ for $a \in \{1, 2\}$, and $\sigma = [0.1, 0.1]$

The second environment is a two-armed bandit with $\mathcal{A} = \{1, 2\}$, $c = [0.65, 0.55]$, $\gamma = (0.1, 0.99)$, $\delta_a = 0.15(1 - \gamma_a^2)$ for $a \in \mathcal{A}$, and $\sigma = [0.1, 0.1]$. The mean reward associated with action 2 can be thought of as "changing more slowly" compared to that associated with action 1. Figure 5a plots the average action selection frequency. We observe that the average frequency of selecting action 2 by naive Thompson sampling converges to zero because action 1 is associated with a smaller c_a. The average frequency of selecting action 1 by predictive sampling is smaller compared to all other algorithms, suggesting that predictive sampling agent deprioritize acquiring information that loses relevance more quickly. Because of this, a predictive sampling agent is accumulating more rewards compared with other agents, as shown in Figure 5b.



(a) Frequencies of selecting action 2



(b) Average rewards collected by each agent

Figure 5: Predictive sampling and state-of-the-art algorithms in a two-armed bandit with $c = [0.65, 0.55]$, $\gamma = (0.1, 0.99)$, $\delta_a = 0.15(1 - \gamma_a^2)$ for $a \in \{1, 2\}$, and $\sigma = [0.1, 0.1]$