Future Information-Directed Sampling for Bayesian Nonstationary Bandits

Yichen Song
Boston University
ycs@bu.edu

Alessio Russo
Boston University
arusso2@bu.edu

Aldo Pacchiano
Boston University
Broad Institute of MIT and Harvard
pacchian@bu.edu

Abstract

Exploration–exploitation is a central trade-off in bandit learning. While classical algorithms such as upper confidence bound methods and Thompson Sampling effectively balance this trade-off in stationary environments, their exploration strategies mainly reduce uncertainty about the current optimal arm, which can be insufficient in nonstationary settings where future optimal arms may differ substantially from current ones. In this paper, we propose Future Information-Directed Sampling (FIDS), a new algorithm for Bayesian nonstationary bandits that explicitly explores to gather information about future optimal arms. We show that FIDS achieves regret comparable to Thompson Sampling up to a small constant factor, while being able to exploit predictive information structures that conventional exploration objectives fail to capture. To address the practical difficulty of posterior inference, we further propose a supervised-learning-based approximation framework that learns the FIDS policy from offline data, and demonstrate its effectiveness on synthetic benchmarks.

1 Introduction

In this paper, we study multi-armed bandits (MABs) in nonstationary environments. MAB problems have been extensively studied over the past decades as a fundamental sequential decision-making framework [10, 11, 13], primarily under the assumption that the underlying environment is stationary. However, this assumption is often unrealistic in real-world applications such as recommendation systems [4] and online advertising [19], where user preferences and reward distributions may evolve over time. Motivated by these applications, we consider a nonstationary setting in which the environments appearing over time are generated according to a known prior distribution. We refer to this setting as Bayesian nonstationary bandits.

A core challenge in nonstationary bandits is balancing the exploration-exploitation tradeoff [13], as the environment changes over time. Nevertheless, it is possible to build an intuition by looking at the literature of stationary MABs: in stationary Bayesian bandit problems, this trade-off can be tackled by using methods like Thompson Sampling (TS) [25], which achieves both strong theoretical guarantees and empirical performance [23]. An information-theoretic analysis of TS is given by [22], who show that the ratio of instantaneous regret to mutual information gained about the optimal arm is bounded uniformly over time. In stationary environments, gathering information about the current optimal arm is a natural exploration objective, since the optimal arm remains informative for future

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

decisions. However, this intuition breaks down in nonstationary environments, where the optimal arm may change over time. As a result, exploration strategies that focus only on reducing uncertainty about the current optimal arm may fail to gather information that is useful for future decision-making.

Motivated by this observation, we propose a new exploration objective that explicitly gathers information about future optimal arms, while retaining instantaneous expected reward maximization as the exploitation objective. To balance this new exploration–exploitation trade-off, we introduce Future Information-Directed Sampling (FIDS). Our main contributions are summarized as follows:

1. We prove in Theorem 3.2 that FIDS achieves the same regret upper bound as Thompson Sampling in Bayesian nonstationary environments, up to a small constant factor.
2. Through several analytic examples, we demonstrate that FIDS can exploit predictive information structures that conventional exploration objectives fail to capture, leading to significantly improved performance over baseline methods.
3. In practice, the prior distribution may be unknown and posterior inference may be computationally intractable. To sidestep these difficulties, we propose a supervised-learning framework that trains an approximate FIDS policy from offline data, inspired by DPT [14], which similarly trains a Thompson Sampling policy via supervision. Empirically, the learned FIDS policy outperforms DPT and other baselines by effectively exploiting the information structure of the environment.

2 Preliminaries and Problem Formulation

In this section we introduce the problem formulation, and provide background knowledge on Thompson Sampling and Information Directed Sampling.

We consider a finite MAB problem with horizon n and an action set $\mathcal{A}$ with $|\mathcal{A}| = K$ arms. We model the non-stationary environment by a random variable ν drawn from a prior distribution known to the learner. Here $\nu = (\nu_{t,a})_{t\in[n],a\in\mathcal{A}}$ where each $\nu_{t,a}$ specifies the reward distribution of arm a at round t. Consider $R_{t,a}$ to be the reward one will see if they play arm a at round t. We assume that conditional on $\nu_{t,a}$, the reward $R_{t,a}$ is sampled independently as $R_{t,a} \sim \nu_{t,a}$. As a special case, if for every $a \in \mathcal{A}$, we have $\nu_{t,a} = \nu_{s,a}$ for all $t, s \in [n]$ almost surely, then the model reduces to the standard stationary Bayesian bandit setting. We also make the following assumption on the reward distributions.

Assumption 2.1. *There exists a constant $\sigma \in \mathbb{R}$, such that for all $a \in \mathcal{A}$, conditioned on $\mathcal{F}_t$, $R_{t,a} - \mathbb{E}[R_{t,a} \mid \mathcal{F}_t]$ is σ sub-Gaussian. And we assume σ is known to the learner.*

In each round t, the learner interacts with the environment by selecting an action A_t according to a policy π_t, which maps the history $\mathcal{F}_{t-1} = (A_1, R_1, \ldots, A_{t-1}, R_{t-1})$ to a distribution over $\mathcal{A}$, and observes reward $R_t \triangleq R_{t,A_t}$. Let $\pi = (\pi_t)_{t=1}^n$. We evaluate π using the following Bayesian regret:

$$\text{Regret}(\pi, n) = \mathbb{E}\left[\sum_{t=1}^{n} \left(R_{t,A_t^\star} - R_{t,A_t}\right)\right], \tag{1}$$

where $A_t^\star \in \arg\max_{a\in\mathcal{A}} \mu(\nu_{t,a})$. And $\mu(\nu_{t,a})$ denotes the mean of the distribution $\nu_{t,a}$.

Notation. We introduce the following shorthand notations that will be used throughout the analysis. For $i, j \in [n]$ with $i \leqslant j$, let $A^\star_{a:b} \triangleq \left[A^\star_i, \ldots, A^\star_j\right]$. Let $\mathbb{P}_t(\cdot) \triangleq \mathbb{P}(\cdot \mid \mathcal{F}_{t-1})$ denote the posterior measure given the history up to round $t-1$. For random variables X and Y, we denote $\mathbb{E}_t[X]$, $H_t(X)$, $H_t(X \mid Y)$, and $I_t(X;Y)$ to be the expectation, entropy, conditional entropy, and mutual information under the posterior $\mathbb{P}_t$, respectively. Formal definitions of these information-theoretic quantities can be found in Appendix C. Note that $\mathbb{E}_t[X]$, $H_t(X)$, $H_t(X \mid Y)$, and $I_t(X;Y)$ are all random variables, since they are functions of the history $\mathcal{F}_{t-1}$ which is a random variable.

2.1 Preliminaries about TS and IDS

We now recall a key result from the information-theoretic analysis of TS by [22], which underpins both the design of FIDS and its regret analysis.

Proposition 2.1. *(Corollary 1 in [22]) Denote the Thompson sampling policy by $\pi^{TS} = \left(\pi_t^{TS}\right)_{t=1}^n$. Suppose that conditioned on $\mathcal{F}_t$, $R_{t,a} - \mathbb{E}\left[R_{t,a} \mid \mathcal{F}_t\right]$ is σ sub-Gaussian, where σ is a constant number in $\mathbb{R}$. Then,*

$$\frac{\sum_{a\in\mathcal{A}} \pi_t^{TS}(a)\mathbb{E}_t\left[R_{t,A_t^\star} - R_{t,a}\right]}{\sqrt{\sum_{a\in\mathcal{A}} \pi_t^{TS}(a) I_t\left(A_t^\star; R_{t,a}\right)}} \leqslant \sigma\sqrt{2K}, \quad \forall t \in [n], \text{ a.s.} \tag{2}$$

The left-hand side of Equation (2) is known as the *information ratio*. Using this bound, [22] show that the Bayesian regret of TS is at most $\sigma\sqrt{2nKH(A^\star)}$, where $A^\star \triangleq A_1^\star = \cdots = A_n^\star$ since the environment is stationary.

In addition, IDS [21] selects a policy π_t^{IDS} that explicitly minimizes the information ratio at each round t, i.e.,

$$\pi_t^{\text{IDS}} \in \operatorname*{argmin}_{\pi_t\in\Delta(\mathcal{A})} \frac{\sum_{a\in\mathcal{A}} \pi_t(a)\mathbb{E}_t\left[R_{t,A_t^\star} - R_{t,a}\right]}{\sqrt{\sum_{a\in\mathcal{A}} \pi_t(a) I_t\left(A_t^\star; R_{t,a}\right)}}, \tag{3}$$

IDS enjoys the same worst-case regret bound as TS, but can achieve significantly improved performance in certain scenarios (see [21] for more details).

Using the same proof technique, [18] analyze TS in the nonstationary setting of Section 2 and obtain a regret bound of $O(\sigma\sqrt{2nKH([A_1^\star, \ldots, A_n^\star])})$. This subsumes the stationary result: when the environment is stationary, $A_1^\star = \cdots = A_n^\star = A^\star$) almost surely, so $H([A_1^\star, \ldots, A_n^\star]) = H(A^\star)$.

3 FIDS Algorithm

In this section, we develop FIDS by addressing a key limitation of TS and IDS in nonstationary environments. In stationary settings, both algorithms achieve low regret by bounding the information ratio

$$\frac{\sum_{a\in\mathcal{A}} \pi_t(a), \mathbb{E}t[R_{t,A_t^\star} - R_{t,a}]}{\sqrt{\sum_{a\in\mathcal{A}} \pi_t(a), I_t(A_t^\star; R_{t,a})}},$$

which controls the trade-off between instantaneous regret and information gained about the optimal arm $A_t^\star$. This works because $A_t^\star$ is fixed, so reducing uncertainty about it directly improves future decisions. In nonstationary environments, however, $A_t^\star$ may vary with t, so information gained about the current optimum need not reduce uncertainty about future optimal arms. FIDS addresses this by replacing the per-round information term with one that targets the sequence of optimal arms.

To address the exploration-exploration problem in a meaningful way, we introduce *Forward-looking Information-Directed Sampling* (FIDS), which balances instantaneous regret against information gained about future optimal arms. Under Assumption 2.1, FIDS is defined by the policy sequence $\pi^{\text{FIDS}} = (\pi_t^{\text{FIDS}})_{t=1}^n$, where π_t^{FIDS} is any policy that solves the following optimization problem

$$\pi_t^{\text{FIDS}} \in \operatorname*{argmin}_{\pi_t\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} \pi_t(a)\mathbb{E}_t[R_{t,A_t^\star} - R_{t,a}] - \sigma\sqrt{2K \cdot \sum_{a\in\mathcal{A}} \pi_t(a) I_t\left(A_{t+1:n}^\star; R_{t,a}\right)} \tag{4}$$

Note that solving Equation (4) is equivalent to solving the following Equation (5)

$$\pi_t^{\text{FIDS}} \in \operatorname*{argmax}_{\pi_t\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} \pi_t(a)\mathbb{E}_t[R_{t,a}] + \sigma\sqrt{2K \cdot \sum_{a\in\mathcal{A}} \pi_t(a) I_t(A_{t+1:,n}^\star; R_{t,a})}, \tag{5}$$

as $\sum_{a\in\mathcal{A}} \pi_t(a)\mathbb{E}_t\left[R_{t,A_t^\star}\right] = \mathbb{E}_t\left[R_{t,A_t^\star}\right]$ is a constant with respect to π_t.

A useful property of Equation (5) is that the optimization in Equation (5) always admits an optimal policy supported on at most two arms, as shown in the following lemma (proof deferred to the appendix in section F.1).

Lemma 3.1. *Given $\mathbb{E}_t\left[R_{t,a}\right]$ and $I_t\left(\left[A_{t+1}^\star, \ldots, A_n^\star\right]; R_{t,a}\right)$, Equation* (5) *is maximized at some $\pi^\star$ which has at most 2 non-zero elements.*

Intuitively, the objective in (5) is concave in π_t, so at any optimum every arm in the support must have the same marginal contribution. This equal-gradient condition forces a linear relationship between the reward and information values of the supported arms, and in the proof we show that any optimal mixture can be replicated by a two-point mixture of the arms in the support.

Before proceeding with the regret bound of FIDS, we note a subtlety in applying IDS to the nonstationary setting.

Remark 3.1. *In nonstationary environments, $\sum_{a\in\mathcal{A}} \pi_t(a), I_t(A_t^\star; R_{t,a})$ can equal zero: even if the current optimal arm $A_t^\star$ is known with certainty, the future optimal arms $A_{t+1}^\star, \ldots, A_n^\star$ may remain unknown, so the problem is still nontrivial. Because the standard ratio form of IDS is undefined when this information term vanishes, we adopt the following additive variant for nonstationary environments:*

$$\pi_t^{IDS(ns)} \in \operatorname*{argmax}_{\pi_t\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} \pi_t(a)\mathbb{E}_t[R_{t,a}] + \sigma\sqrt{2K \cdot \sum_{a\in\mathcal{A}} \pi_t(a) I_t(A_t^\star; R_{t,a})} \tag{6}$$

All experiments involving IDS use this formulation.

We now show that FIDS achieves a regret bound comparable to that of TS in nonstationary environments.

Theorem 3.2. *Under Assumption 2.1,*

$$\mathrm{Regret}(\pi^{\mathrm{FIDS}}, n) \leqslant 2\sigma\sqrt{2K \cdot n \cdot H([A_1^\star, \ldots, A_n^\star])} \tag{7}$$

This bound matches the TS regret bound of [18] up to a constant factor of 2. Despite this worst-case similarity, FIDS can exploit additional structure in the environment; in Section 3.1, we show that it achieves strictly better performance in several settings. Here is the proof of the theorem.

Proof. Here we outline the proof with several supporting lemmas. The proof of lemmas can be found in Section D. For ease of the notation, in this section, we use $\Delta_t(\pi_t) := \sum_{a\in\mathcal{A}} \pi_t(a)\mathbb{E}_t[R_{t,A_t^\star} - R_{t,a}]$, $\tilde{g}_t(\pi_t) := \sum_{a\in\mathcal{A}} \pi_t(a) I_t([A_{t+1}^\star, \ldots, A_n^\star]; R_{t,a})$ and $g_t(\pi_t) := \sum_{a\in\mathcal{A}} \pi_t(a) I_t(A_t^\star; R_{t,a})$. When an action A_t is sampled from π_t, we slightly abuse the notation by letting $\Delta_t(A_t) := \mathbb{E}_t\left[R_{t,A_t^\star} - R_{t,A_t}\right], \tilde{g}_t(A_t) := I_t\left(\left[A_{t+1}^\star, \ldots, A_n^\star\right]; (A_t, R_{t,A_t})\right)$ and $g_t(A_t) := I_t(A_t^\star; (A_t, R_{t,A_t}))$, as Lemma D.1 shows that $\Delta_t(A_t) = \Delta_t(\pi_t)$, $\tilde{g}_t(A_t) = \tilde{g}_t(\pi_t)$, and $g_t(A_t) = g_t(\pi_t)$.

The first part of the proof connects π^{FIDS} with π^{TS}. By the Tower Rule, $\mathrm{Regret}\left(\pi^{\mathrm{FIDS}}, n\right) = \mathbb{E}\left[\sum_{t=1}^n \Delta_t\left(\pi_t^{\mathrm{FIDS}}\right)\right]$, then by adding and subtracting $\sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{FIDS}}\right)}$, we get

$$\begin{aligned}
\mathrm{Regret}\left(\pi^{\mathrm{FIDS}}, n\right) &= \mathbb{E}\left[\sum_{t=1}^n \Delta_t\left(\pi_t^{\mathrm{FIDS}}\right) - \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{FIDS}}\right)} + \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{FIDS}}\right)}\right] \\
&\overset{(a)}{\leqslant} \mathbb{E}\left[\sum_{t=1}^n \Delta_t\left(\pi_t^{\mathrm{TS}}\right) - \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{TS}}\right)} + \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{FIDS}}\right)}\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{t=1}^n \Delta_t\left(\pi_t^{\mathrm{TS}}\right) - \sigma\sqrt{2K \cdot g_t\left(\pi_t^{\mathrm{TS}}\right)} - \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{TS}}\right)}\right. \\
&\qquad \left. + \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{FIDS}}\right)} + \sigma\sqrt{2K \cdot g_t\left(\pi_t^{\mathrm{TS}}\right)}\right] \\
&\overset{(c)}{\leqslant} \mathbb{E}\left[\sum_{t=1}^n \sigma\sqrt{2K \cdot g_t\left(\pi_t^{\mathrm{TS}}\right)} - \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{TS}}\right)} + \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{FIDS}}\right)}\right]
\end{aligned}$$

where (a) uses the fact that π_t^{FIDS} is the minimizer of Equation (4); (b) adds and subtracts $\sigma\sqrt{2K \cdot g_t\left(\pi_t^{\mathrm{TS}}\right)}$; (c) holds because of Proposition 2.1. Then by Lemma D.2,

$$\mathbb{E}\left[\sum_{t=1}^n \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{FIDS}}\right)}\right] \leqslant \sigma\sqrt{2n \cdot K \cdot H(A_{1:n}^\star)} \tag{8}$$

by Lemma D.4,

$$\mathbb{E}\left[\sum_{t=1}^{n} \sigma\sqrt{2K \cdot g_t\left(\pi_t^{\mathrm{TS}}\right)} - \sigma\sqrt{2K \cdot \tilde{g}_t\left(\pi_t^{\mathrm{TS}}\right)}\right] \leqslant \sigma\sqrt{2n \cdot K \cdot H(A^\star_{1:n})} \tag{9}$$

Combining Equation (8) and (9) proves the theorem. □

In the next subsection, we illustrate settings in which TS and IDS fail to exploit the information structure of the environment, while FIDS does not.

3.1 Illustrative Examples: When Does Information Structure Matter?

We present two examples that illustrate the role of information structure. The first is an i.i.d. environment in which the optimal arm is redrawn independently each round, so observations carry no information about future optimal arms. The second features a change point where information about the future environment is encoded in a currently suboptimal arm; we complement the analytical results for this last example with numerical experiments.

Example 3.1 (Memoryless environment)**.** *At each round t, the environment ν_t is sampled i.i.d. with $K = 2$, where $(\mu_{t,1}, \mu_{t,2}) = (0.4, 0.2)$ with probability $1/2$ and $(\mu_{t,1}, \mu_{t,2}) = (0.3, 0.4)$ with probability $1/2$. Conditional on ν_t, rewards satisfy $R_{t,a} \sim \mathrm{Bernoulli}(\mu_{t,a})$ for $a \in \{1, 2\}$.*

In this setting, exploration is futile: no observation at round t can reduce uncertainty about future environments. As shown in Proposition F.1, FIDS recognizes this and always selects arm 1, the arm with the highest expected instantaneous reward, achieving cumulative reward $0.35n$. In contrast, TS and IDS allocate probability to unnecessary exploration, achieving only $0.325n$ and $0.3294n$ respectively.

Example 3.1 serves as an environment where no future information can be gathered. We next consider a setting where information about future optimal arms is encoded in a currently suboptimal arm, and FIDS is able to exploit this cross-environment structure.

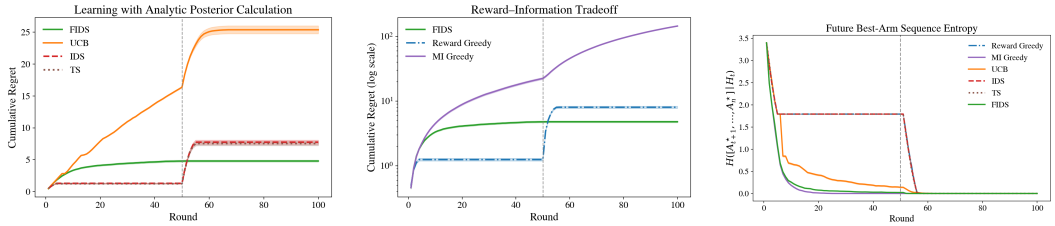
Example 3.2 (Magic arm environment)**.** *We consider a non-stationary 6-arm bandit with horizon $n = 100$ and a single change point at $t = 50$. The environment is **Env 1** for $t \leqslant 50$ and **Env 2** for $t > 50$, and is stationary within each segment.*

- `Env 1`. *Let $Z_1 \sim \mathrm{Unif}(\{1, \ldots, 5\})$. For $a \in \{1, \ldots, 5\}$, $\mu_a^1 = 0.9$ if $a = Z_1$ and $\mu_a^1 = 0.4$ otherwise, with $R_{t,a} \sim \mathcal{N}(\mu_a^1, 0.1^2)$. For $a = 6$, $\mu_6^1 = Z_2/7$ and $R_{t,6} \sim \mathcal{N}(\mu_6^1, 0.1^2)$.*
- `Env 2`. *Let $Z_2 \sim \mathrm{Unif}(\{1, \ldots, 6\})$, independent of Z_1. For all $a \in \mathcal{A}$, $\mu_a^2 = 3.0$ if $a = Z_2$ and $\mu_a^2 = 0.5$ otherwise, with $R_{t,a} \sim \mathcal{N}(\mu_a^2, 0.5^2)$.*

Note that the information about `Env 2` is encoded in the mean reward of arm 6 in `Env 1`. In particular, if $\mu_{1,6}$ is known exactly, then the means of all arms in `Env 2` are immediately determined. FIDS is able to exploit this structure through its objective of gathering information about future optimal arms. In contrast, TS and IDS are designed to gather information only about the currently optimal arm. Since arm 6 is never optimal in `Env 1`, these methods fail to actively explore it and therefore cannot effectively exploit the information it provides about `Env 2`.

Numerical results on the magic arm environment. We evaluate FIDS, UCB [2], IDS and TS on the magic arm environment. The results are shown in Figure 1. In Figure 1a, compared with the other methods, FIDS spends more rounds identifying the optimal arm in *Env 1*, since it allocates part of its sampling probability to the informative but suboptimal arm 6. However, FIDS almost immediately identifies the optimal arm in *Env 2*, resulting in nearly zero regret during the second half of the horizon. This phenomenon is also reflected in Figure 1c: by pulling arm 6 in *Env 1*, FIDS rapidly reduces the entropy of the optimal arms in future rounds, rather than only the entropy of the optimal arm in the current environment.

In Figure 1b, we additionally compare FIDS with two ablated policies: a "pure exploitation" policy that optimizes only the reward term in Equation (5), and a "pure exploration" policy that optimizes only the mutual information term. FIDS outperforms both policies by effectively balancing exploration and exploitation.



(a) (b) (c)

Figure 1: (a) Cumulative regret of FIDS, IDS and TS. (b) Cumulative regret (log scale) of FIDS, Reward Greedy (only optimize for the first reward term of Equation (5)), and MI only (only optimize for the second mutual information term of Equation (5)). (c) The entropy of $\left[A^\star_{t+1}, \ldots, A^\star_n\right]$ conditioned on the history.

Algorithm 1 Infer π^{FIDS} from Dataset $\mathcal{D}$ with prediction window length k: Training and Deployment

// `Training phase - supervised learning with offline dataset` $\mathcal{D}$

Construct training samples $\{z_t^\xi\}_{t=1}^n$ from each data instance $\xi \in \mathcal{D}$; initialize the model M_θ with parameter θ.

for $t = 1$ **to** N_{epochs} **do**

  Calculate loss (10) and do back-propagation to update θ

end

Output the trained model M_θ

// `Deployment phase`

for $t = 1$ **to** n **do**

  For each action a compute $\widehat{\mathbb{E}}_t(R_{t,a})$ and $\widehat{I}_t(A^\star_{t+1:t+k}; R_{t,a})$ using Eqs. (11)-(12)

  Solve

$$(i^\star, j^\star, \alpha^\star) = \underset{i,j\in\mathcal{A},\alpha\in[0,1]}{\operatorname{argmax}} \Big\{\alpha\widehat{\mathbb{E}}_t(R_{t,i}) + (1-\alpha)\widehat{\mathbb{E}}_t(R_{t,j}) + \sigma\sqrt{2\sigma K\Big(\alpha\widehat{I}_t(A^\star_{t+1:t+k}; R_{t,i}) + (1-\alpha)\widehat{I}_t(A^\star_{t+1:t+k}; R_{t,j})\Big)}\Big\}.$$

  Output $\pi_t^{\text{FIDS}}(i^\star) = \alpha^\star$, $\pi_t^{\text{FIDS}}(j^\star) = 1-\alpha^\star$, and $\pi_t^{\text{FIDS}}(a) = 0$ for all $a \notin \{i^\star, j^\star\}$.

end

4 Inferring FIDS policy from Data via Supervised Learning

Computing π^{FIDS} exactly is often impractical: the prior over environments may be unknown, and even when it is known, evaluating Equation (5) requires posterior inference that is generally intractable. A further difficulty is that the mutual information $I_t(A^\star_{t+1:n}; R_{t,a})$ involves the full sequence of future optimal arms, whose support grows exponentially with the horizon.

To address these challenges, we propose a supervised-learning approach (Algorithm 1). Given an offline dataset of environment instances, we train models to estimate the two quantities needed by FIDS: the expected reward $\mathbb{E}_t[R_{t,a}]$ and a *windowed* approximation $I_t(A^\star_{t+1:t+k}; R_{t,a})$, where k is a fixed lookahead horizon that controls the trade-off between fidelity and computational cost. At deployment, these estimates are plugged into Equation (5) to compute π^{FIDS}. We describe the training and deployment phases in detail below.

4.1 On the Training Phase

Algorithm 1 takes an offline dataset $\mathcal{D}$ as input. Each data instance $\xi \in \mathcal{D}$ is generated by first sampling an environment ν from the prior, and then collecting: (i) the full reward table of ν, $\{r^\nu_{t,a}\}_{t\in[n],\, a\in\mathcal{A}}$, (ii) the best-arm sequence of ν, $(a_1^{\star,\nu}, \ldots, a_n^{\star,\nu})$, and (iii) an action sequence $(a_1, \ldots, a_n)$ generated by some behavior policy.

For each $\xi \in \mathcal{D}$ and $t \in [n]$, we construct one training sample: $z_t^\xi := \Big(h_t^\xi,\ a^\star_{t:t+k},\ \{r^\xi_{t,a}\}_{a\in\mathcal{A}}\Big)$, where $h_t^\xi := (a_1, r_1^\nu, \ldots, a_{t-1}, r_{t-1}^\nu)$ denotes the observed history up to round $t-1$, and $r_s^\nu := r^\nu_{s,a_s}$ for $s \in [t]$. $a^{\star,\xi}_{t:t+k} := (a_t^{\star,\nu}, \ldots, a_{t+k}^{\star,\nu})$ denotes the future best-arm sequence.

Our goal is to estimate $\mathbb{E}_t[R_{t,a}]$ and $I_t(A^\star_{t+1:t+k}; R_{t,a})$. To this end, we approximate the following conditional distributions: $\mathbb{P}_t(R_{t,a} \in \cdot)$ for all $a \in \mathcal{A}$, $\mathbb{P}_t(A^\star_{t+1:t+k} \in \cdot)$, $\mathbb{P}_t(A^\star_{t:t+k-1} \in \cdot)$. In practice, we model the reward posterior by a Gaussian, and the best-arm posteriors by categorical distributions. To learn these models, we use a sequential architecture, specifically GPT-2 [20]. The architecture has three heads, and outputs:

$$M_\theta^{\mathrm{rew}}(\cdot \mid h_t^\xi, a), \quad M_\theta^{\mathrm{fut}+}(\cdot \mid h_t^\xi), \quad M_\theta^{\mathrm{fut}}(\cdot \mid h_t^\xi),$$

which model the conditional distributions of $R_{t,a}$, $A^\star_{t+1:t+k}$, and $A^\star_{t:t+k-1}$, respectively. Then, to train the reward posterior, we simply use negative log-likelihood, while the best-arm posteriors are trained via cross-entropy loss. The overall training objective is

$$\mathcal{L}_\theta = \sum_{\xi \in \mathcal{D}} \sum_{t=1}^{n} \Bigg[\sum_{a \in \mathcal{A}} \mathrm{NLL}\big(M_\theta^{\mathrm{rew}}(\cdot \mid h_t^\xi, a), r_{t,a}^\xi\big) + \mathrm{CE}\big(M_\theta^{\mathrm{fut}+}(\cdot \mid h_t^\xi), a^\star_{t+1:t+k}\big) + \mathrm{CE}\big(M_\theta^{\mathrm{fut}}(\cdot \mid h_t^\xi), a^\star_{t:t+k-1}\big) \Bigg]. \tag{10}$$

4.2 On the Deployment Phase

At deployment, we use the inferred posteriors to compute FIDS, which relies on $\mathbb{E}_t\left[R_{t,a}\right]$ and $I_t\left(A^\star_{t+1:t+k}; R_{t,a}\right)$. Given a history h_t, we first estimate the conditional reward mean by the mean of the learned reward posterior:

$$\widehat{\mathbb{E}}_t\left[R_{t,a}\right] = \mathbb{E}_{\widetilde{R}_{t,a} \sim M_\theta^{\mathrm{rew}}(\cdot \mid h_t, a)}\left[\widetilde{R}_{t,a}\right]. \tag{11}$$

Since $M_\theta^{\mathrm{rew}}(\cdot \mid h_t, a)$ is Gaussian, this is simply the predicted Gaussian mean. We next estimate the conditional mutual information via an entropy decomposition

$$I_t(A^\star_{t+1:t+k}; R_{t,a}) = H_t(A^\star_{t+1:t+k}) - H_t(A^\star_{t+1:t+k} \mid R_{t,a}).$$

The first term is estimated directly from the future head: $\widehat{H}_t(A^\star_{t+1:t+k}) = H\big(M_\theta^{\mathrm{fut}+}(\cdot \mid h_t)\big)$. For the conditional entropy term, we draw L Monte Carlo samples: $\widetilde{r}_{t,a}^{(1)}, \ldots, \widetilde{r}_{t,a}^{(L)} \sim M_\theta^{\mathrm{rew}}(\cdot \mid h_t, a)$. For each sampled reward $\widetilde{r}_{t,a}^{(\ell)}$, we form the hypothetical updated history $\widetilde{h}_{t+1}^{(\ell,a)} = (h_t, a, \widetilde{r}_{t,a}^{(\ell)})$. Since $M_\theta^{\mathrm{fut}}(\cdot \mid \widetilde{h}_{t+1}^{(\ell,a)})$ models the posterior distribution of $A^\star_{t+1:t+k}$ after observing action a and reward $\widetilde{r}_{t,a}^{(\ell)}$, we estimate

$$\widehat{H}_t(A^\star_{t+1:t+k} \mid R_{t,a}) = \frac{1}{L} \sum_{\ell=1}^{L} H\Big(M_\theta^{\mathrm{fut}}(\cdot \mid \widetilde{h}_{t+1}^{(\ell,a)})\Big).$$

Therefore, the plug-in Monte Carlo estimator of the conditional mutual information is

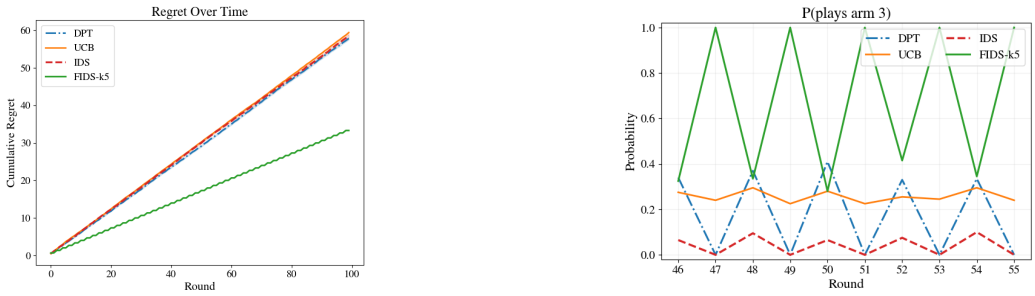
$$\widehat{I}_t(A^\star_{t+1:t+k}; R_{t,a}) = H\big(M_\theta^{\mathrm{fut}+}(\cdot \mid h_t)\big) - \frac{1}{L} \sum_{\ell=1}^{L} H\Big(M_\theta^{\mathrm{fut}}(\cdot \mid \widetilde{h}_{t+1}^{(\ell,a)})\Big). \tag{12}$$

We choose $L = 16$ for all the experiments in Section 5.

Solving Equation (5)**.** With the learned $\mathbb{E}_t\left[R_{t,a}\right]$ and $I_t\left(\left[A^\star_{t+1}, \ldots, A^\star_{t+k}\right]; R_{t,a}\right)$, FIDS policy can be obtained by just plugging in the learned two quantities into Eq. (5) and solve the corresponding optimization problem. With Lemma 3.1, for the plugged-in $\widehat{\mathbb{E}}_t(R_{t,a})$ and $\widehat{I}_t(A^\star_{t+1:n}; R_{t,a})$, one can solve Eq. (5) by first enumerating all pairs of arms and using line search to find the optimal weights assigned on the arms. Details are presented in Alg. 1.

5 Experiments

We evaluate the FIDS policy learned via Algorithm 1 on two environment classes for which exact posterior inference is intractable: the *One-Step Predictive* environment and the *Pair-Revealing* environment. In both environments, information about future optimal arms is encoded in the rewards



(a) (b)

Figure 2: (a) Cumulative Regret of DPT, IDS, UCB and FIDS-k5 in One Step Predictive Environment. (b) The probability of choosing arm 3 in each round.

of a currently suboptimal arm: the first reveals the next-round optimum exactly, while the second only narrows the future best arm to a pair. Further details are provided in the next subsections. Throughout, we set $\sigma = 0.7$ in Equation (5), which is a valid sub-Gaussian parameter by Lemma F.2.

Comparison. We compare the following methods: (i) FIDS with prediction window length $k = 5$; (ii) FIDS with $k = 3, 1$ (used only in the second Pair-Revealing Environment); (iii) DPT; and (iv) IDS. We train IDS using the same procedure as Algorithm 1. The only difference is that for IDS we estimate $I_t(A_t^\star; R_{t,a})$ instead of $I_t(A_{t+1:t+k}^\star; R_{t,a})$.

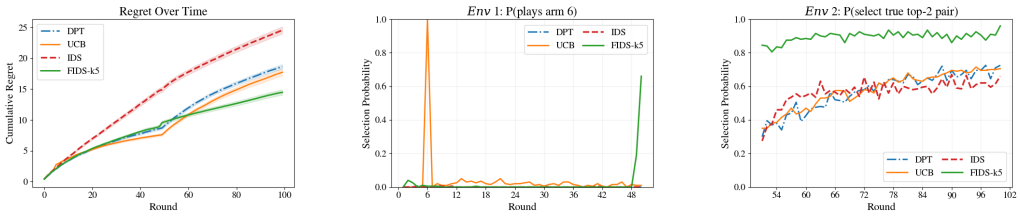
5.1 One-Step Predictive Environment

Environment Description (Next round best arm encoding). We consider a nonstationary 3-arm bandit with horizon $n = 100$, divided into 50 independent windows of length 2. For each window $j \in \{1, \ldots, 50\}$, corresponding to rounds $2j-1$ and $2j$, we independently sample $m_j \sim \mathrm{Unif}(\{1, 2, 3\})$.

At round $2j-1$, we sample $Z_j \sim \mathrm{Unif}(\{1, 2\})$. Arms 1 and 2 have rewards $\mu_{2j-1,a} = \mathbf{1}[a = Z_j]$, while arm 3 has reward $\mu_{2j-1,3} = 1/(2m_j)$. At round $2j$, the best arm is determined by m_j: $\mu_{2j,a} = \mathbf{1}[a = m_j]$ for all $a \in \{1, 2, 3\}$. All rewards are deterministic.

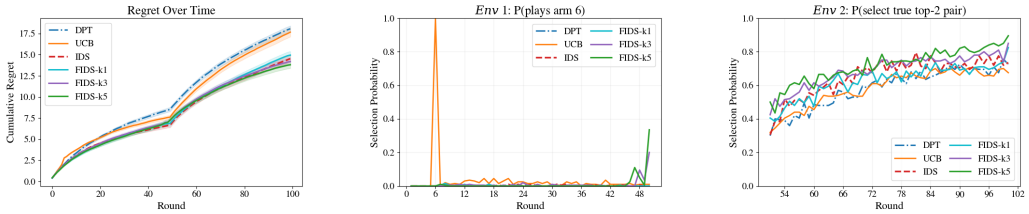
The key feature is that arm 3's reward in odd rounds encodes the identity of the best arm in the following even round, creating a one-step-ahead information channel that a forward-looking policy can exploit.

Results. Figures 2a and 2b summarize the main results (additional plots in Figures 5a–5b in the appendix). Although the windows are independent, each contains a one-step information channel: pulling arm 3 in the first round reveals the optimal arm in the second round. The learned FIDS policy recovers exactly this behavior: it selects arm 3 in the first round of each window and the true optimal arm in the second (Figures 5a, 2b, and 5b). In contrast, TS and IDS never learn to exploit arm 3, since their exploration targets only the current optimal arm.



(a) (b) (c)

Figure 3: Evaluation Results for the Pair-Revealing Environment with $\sigma_6^2 = 0.0$. (a) Cumulative regret. (b) The probability of play arm 6 at each round. (c) The probability of selecting an arm that belongs to the true top-2 set of the underlying environment.



(a) (b) (c)

Figure 4: Evaluation Results for the Pair-Revealing Environment with $\sigma_6^2 = 0.05^2$. (a) Cumulative regret. (b) The probability of play arm 6 at each round. (c) The probability of selecting an arm that belongs to the true top-2 set of the underlying environment.

5.2 Pair-Revealing Environment

Environment Description. Unlike the previous environment, where the informative arm directly identifies the future optimum, here it only reveals the top-2 arm pair in the future environment. We consider a nonstationary 6-arm bandit with horizon $n = 100$ and a single change point at $t = 50$. The environment is `Env 1` for $t \leqslant 50$ and `Env 2` for $t > 50$, stationary within each segment.

- In `Env 1`, let $Z_1 \sim \mathrm{Unif}(\{1, \ldots, 5\})$. For $a \in \{1, \ldots, 5\}$, $\mu_a^1 = 1$ if $a = Z_1$ and $\mu_a^1 = U_a$ otherwise, where $U_a \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0, 1])$, with $R_{t,a} \sim \mathcal{N}(\mu_a^1, 0.5^2)$. Arm 6 has mean $\mu_6^1 = m/30$, where $m \sim \mathrm{Unif}(\{0, \ldots, 14\})$ indexes one of the $\binom{6}{2} = 15$ unordered arm pairs; we denote this pair by $\mathcal{S}(m)$. We consider two noise settings for arm 6: $\sigma_6^2 = 0$ (exact observation) and $\sigma_6^2 = 0.05^2$ (noisy observation).
- In `Env 2`, we sample $V_1, \ldots, V_6 \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0, 1])$ and assign the two largest values to the arms in $\mathcal{S}(m)$, with the remaining values randomly permuted among the other arms, yielding means $\{\mu_a^2\}$. Rewards satisfy $R_{t,a} \sim \mathcal{N}(\mu_a^2, 0.5^2)$.

The information structure is that arm 6 in `Env 1`, while never optimal, encodes which pair of arms will dominate in `Env 2`. The two noise settings let us examine how the precision of this signal affects the benefit of forward-looking exploration.

Results. We evaluate under two noise settings for arm 6.

- ($\sigma_6^2 = 0$). We set the prediction window to $k = 5$. Figure 3a shows that FIDS matches the baselines in `Env 1`, while intentionally pulling arm 6 (Figure 3b) to gather information about `Env 2`. Upon entering `Env 2`, FIDS identifies the candidate top-2 arms and focuses exploration on them rather than the full arm set.
- ($\sigma_6^2 = 0.05^2$). When arm 6's reward is noisy, recovering μ_6^1 requires multiple pulls, making a longer prediction window more valuable. We compare FIDS with $k = 5$ (FIDS-K5) and a shorter window. As shown in Figures 4b and 4c, pulling arm 6 earlier in `Env 1` allows FIDS-K5 to infer the top-2 pair more accurately, leading to faster adaptation after the change point.

Across both environments, the learned FIDS policy consistently exploits informative but suboptimal arms that TS and IDS ignore, confirming that Algorithm 1 successfully recovers forward-looking exploration behavior from data.

6 Conclusion

We introduced FIDS, an algorithm for Bayesian nonstationary bandits that balances instantaneous reward against information gained about future optimal arms. FIDS matches the regret guarantee of Thompson Sampling, yet analytical examples show it can significantly outperform both TS and IDS by exploiting predictive structure across environments. To handle settings where exact posterior inference is intractable, we proposed a supervised-learning framework that approximates the FIDS policy from offline data. Experiments confirm that the learned policy consistently outperforms supervised-learning baselines with alternative exploration objectives. We further discuss the limitation and future work in Appendix A.

References

[1] Yasin Abbasi-Yadkori, András György, and Nevena Lazić. A new look at dynamic regret for non-stationary stochastic bandits. *Journal of Machine Learning Research*, 24(288):1–37, 2023.

[2] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[3] Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on learning theory*, pages 138–158. PMLR, 2019.

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

[5] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[6] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[8] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

[9] Neha Gupta, Ole-Christoffer Granmo, and Ashok Agrawala. 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, 2011.

[10] T. L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, March 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8. URL `https://www.sciencedirect.com/science/article/pii/0196885885900028`.

[11] Tze Leung Lai. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *The Annals of Statistics*, 15(3):1091 – 1114, 1987. doi: 10.1214/aos/1176350495. URL `https://doi.org/10.1214/aos/1176350495`.

[12] Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

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

[14] Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:43057–43083, 2023.

[15] Yueyang Liu, Benjamin Van Roy, and Kuang Xu. Nonstationary bandit learning via predictive sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 6215–6244. PMLR, 2023.

[16] Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:47016–47031, 2023.

[17] Joseph Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change detection. In *Artificial intelligence and statistics*, pages 442–450. PMLR, 2013.

[18] Seungki Min and Daniel Russo. An information-theoretic analysis of nonstationary bandit learning. In *International Conference on Machine Learning*, pages 24831–24849. PMLR, 2023.

[19] Sandeep Pandey, Deepak Agarwal, Deepayan Chakrabarti, and Vanja Josifovski. Bandits for taxonomies: A model-based approach. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 216–227. SIAM, 2007.

[20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

[21] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in neural information processing systems*, 27, 2014.

[22] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

[23] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[24] Joe Suk and Samory Kpotufe. Tracking most significant arm switches in bandits. In *Conference on Learning Theory*, pages 2160–2182. PMLR, 2022.

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

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

A Limitations and Future Work

One limitation of Algorithm 1 is that training requires access to the full reward table $\{R_{t,a}\}_{t\in[n],a\in\mathcal{A}}$ in the offline dataset, which may be unrealistic in many practical settings. In addition, estimating $I_t(A^\star_{t+1:n}; R_{t,a})$ during deployment requires sampling multiple $\widetilde{R}_{t,a}$ values and performing Monte Carlo estimation, which can become computationally expensive when a large number of samples is needed. Developing more scalable implementations of the FIDS principle therefore remains an important direction for future work.

B More Related Work

Nonstationary Bandit Learning. Predictive Sampling [15] considers the same Bayesian nonstationary setting and shares a similar intuition with our work, namely that exploration strategies should account for future information in nonstationary environments. However, our work adopts a stronger regret notion in the theoretical analysis. As a result, our regret guarantees directly imply guarantees under their setting, while the converse does not hold.

In addition, several variants of Thompson Sampling have been proposed for nonstationary environments (e.g., [26, 9, 17]). However, these methods typically rely on specific structural assumptions about the environment and do not provide meaningful guarantees in general Bayesian nonstationary settings.

Nonstationary bandits have also been extensively studied in the frequentist setting, e.g., [8, 3]. More recent results in [24, 1] establish regret bounds of order $O(\sqrt{Ln})$, where L denotes the number of changes of the optimal arm. Interestingly, these bounds share a similar flavor with our result, as both depend on the temporal variation of the optimal arm sequence.

Learning Decision-Making Algorithms from Data. The proposed supervised-learning-based framework for approximating the FIDS policy is primarily inspired by [14], which introduces the Decision-Pretrained Transformer (DPT), a model that theoretically learns the Thompson Sampling policy. More broadly, the idea of learning sequential decision-making algorithms from data is also related in spirit to the meta-RL literature [12, 6, 16, 7].

C Information Theoretic Preliminaries

In this section, X is limited to discrete random variables and Y, Z can be general random variables, under measure $\mathbb{P}$.

Lemma C.1. *(Mutual Information and Entropy, Theorem 2.4.1 in [5])*

$$H(X) = -\sum_x \mathbb{P}(X = x) \log(\mathbb{P}(X = x))$$

$$I(X;Y) = H(X) - H(X|Y)$$

Lemma C.2. *(Chain Rule for Entropy, Theorem 2.5.1 in [5])*

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, X_{i-1})$$

Lemma C.3. *(Chain Rule for Mutual Information, Theorem 2.5.2 in [5])*

$$I([X_1, X_2, \ldots, X_n]; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_1, \ldots, X_{i-1})$$

A direct application of the above lemma gives the following Corollary,

Corollary C.4.

$$I(X;Y) - I(Z;Y) = I(X;Y|Z) - I(Z;Y|X)$$

Proof. The proof is done by simply noticing that $I(X_1;Y) + I(X_2;Y|X_1) = I(X_2;Y) + I(X_1;Y|Z_2) = I([X_1,X_2];Y)$ by Lemma C.3. □

Corollary C.5.

$$H(X|Y,Z) \leqslant H(X|Y)$$

Proof.

$$H(X|Y,Z) \leqslant H(X|Y) \Leftrightarrow H(X) - I(X;[Y,Z]) \leqslant H(X) - I(X;Y) \Leftrightarrow I(X;[Y,Z]) \geqslant I(X;Y)$$

which is true by Lemma C.3. □

Lemma C.6. *(Non-negativity of Mutual information and Entropy)*

$$H(X) \geqslant 0; H(X|Y) \geqslant 0; I(X;Y) \geqslant 0; I(X;Y|Z) \geqslant 0$$

Lemma C.7. *Let $a, b \geqslant 0$, $\sqrt{a} - \sqrt{b} \leqslant \sqrt{|a-b|}$*

Proof. Clearly we only need to consider the case where $a \geqslant b$. In such a case,

$$\sqrt{a} - \sqrt{b} \leqslant \sqrt{|a-b|} \Leftrightarrow a + b - 2\sqrt{ab} \leqslant a - b \Leftrightarrow a \geqslant b$$

□

D Supporting Lemmas in the Proof of Theorem 3.2

Lemma D.1.

$$\begin{aligned}\mathbb{E}_t\left[R_{t,A_t^\star} - R_{t,A_t}\right] &= \sum_{a\in[K]} \mathbb{P}_t(A_t = a)\mathbb{E}_t\left[R_{t,A_t^\star} - R_{t,a} \mid A_t = a\right] \\ &\overset{(a)}{=} \sum_{a\in[K]} \pi_t(a)\mathbb{E}_t\left[R_{t,A_t^\star} - R_{t,a}\right] \end{aligned} \tag{13}$$

(a) uses the fact that conditioned on $\mathcal{F}_{t-1}$, A_t is jointly independent of $R_{t,A_t^\star}$ and $R_{t,a}$.

$$\begin{aligned} I_t\left(A_t^\star;(A_t,R_{t,A_t})\right) &\overset{(a)}{=} I_t\left(A_t^\star;A_t\right) + I_t\left(A_t^\star;R_{t,A_t} \mid A_t\right) \\ &\overset{(b)}{=} I_t\left(A_t^\star;R_{t,A_t} \mid A_t\right) \\ &= \sum_{a\in[K]} \mathbb{P}_t(A_t = a) I_t\left(A_t^\star;R_{t,a} \mid A_t = a\right) \\ &\overset{(c)}{=} \sum_{a\in[K]} \mathbb{P}_t(A_t = a) I_t\left(A_t^\star;R_{t,a}\right) \\ &= \sum_{a\in[K]} \pi_t(a) I_t\left(A_t^\star;R_{t,a}\right) \end{aligned} \tag{14}$$

(a) uses Lemma C.3; (b) uses the fact conditioned on $\mathcal{F}_{t-1}$, A_t is independent of $A_t^\star$, and the mutual information between two independent variables is 0; (c) uses the fact that conditioned on $\mathcal{F}_{t-1}$, A_t is jointly independent of $A_t^\star$ and $R_{t,a}$.

Similarly,

$$I_t\left(\left[A_{t+1}^\star,\ldots,A_n^\star\right];(A_t,R_{t,A_t})\right) = \sum_{a\in[K]} \pi_t(a) I_t\left(\left[A_{t+1}^\star,\ldots,A_n^\star\right];R_{t,a}\right) \tag{15}$$

Lemma D.2.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n}\sigma\sqrt{2K\cdot I_t(A^\star_{t+1:n};(A_t,R_{t,A_t}))}\right] &\overset{(a)}{\leqslant} \sum_{t=1}^{n}\sigma\sqrt{2K\cdot \mathbb{E}\left[I_t(A^\star_{t+1:n};(A_t,R_{t,A_t}))\right]} \\
&= \sum_{t=1}^{n}\sigma\sqrt{2K\cdot I(A^\star_{t+1:n};(A_t,R_{t,A_t})|\mathcal{F}_{t-1})} \\
&\overset{(b)}{\leqslant} \sigma\sqrt{2n\cdot K\cdot \sum_{t=1}^{n} I(A^\star_{t+1:n};(A_t,R_{t,A_t})|\mathcal{F}_{t-1})} \\
&\overset{(c)}{\leqslant} \sigma\sqrt{2n\cdot K\cdot \sum_{t=1}^{n} I(A^\star_{t+1:n};(A_t,R_{t,A_t})|\mathcal{F}_{t-1})} \\
&\overset{(d)}{=} \sigma\sqrt{2n\cdot K\cdot I(A^\star_{1:n};\mathcal{F}_n)} \\
&\overset{(e)}{\leqslant} \sigma\sqrt{2n\cdot K\cdot H(A^\star_{1:n})} \qquad (16)
\end{aligned}$$

(a) uses Jensen's Inequality; (b) uses Cauchy–Schwarz inequality; (c) uses Lemma C.3 and Lemma C.6; (d) uses Lemma C.3; (e) uses Lemma C.1 and C.6.

Lemma D.3.

$$\sigma\sqrt{2K\cdot g_t\left(\pi_t^{TS}\right)} - \sigma\sqrt{2K\cdot \tilde{g}_t\left(\pi_t^{TS}\right)} \leqslant \sigma\sqrt{2K}\cdot\sqrt{I_t(A^\star_t;(\tilde{A}_t,R_{t,\tilde{A}_t})|A^\star_{t+1:n})}$$

Proof. When $I_t(A^\star_t;(\tilde{A}_t,R_{t,\tilde{A}_t})) \leqslant I_t(A^\star_{t+1:n};(\tilde{A}_t,R_{t,\tilde{A}_t}))$. We have term1$(t) \leqslant 0$ directly. When $I_t(A^\star_t;(\tilde{A}_t,R_{t,\tilde{A}_t})) > I_t(A^\star_{t+1:n};(\tilde{A}_t,R_{t,\tilde{A}_t}))$. Then

$$\begin{aligned}
\text{term1}(t) &\overset{(a)}{\leqslant} \sigma\sqrt{2K}\cdot\sqrt{I_t(A^\star_t;(\tilde{A}_t,R_{t,\tilde{A}_t})) - I_t(A^\star_{t+1:n};(\tilde{A}_t,R_{t,\tilde{A}_t}))} \\
&\overset{(b)}{=} \sigma\sqrt{2K}\cdot\sqrt{I_t(A^\star_t;(\tilde{A}_t,R_{t,\tilde{A}_t})|A^\star_{t+1:n}) - I_t(A^\star_{t+1:n};(\tilde{A}_t,R_{t,\tilde{A}_t})|A^\star_t)} \\
&\leqslant \sigma\sqrt{2K}\cdot\sqrt{I_t(A^\star_t;(\tilde{A}_t,R_{t,\tilde{A}_t})|A^\star_{t+1:n})}
\end{aligned}$$

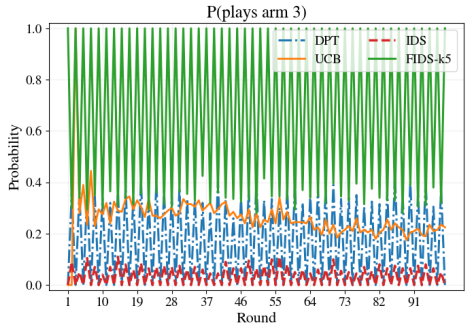
(a) holds because of Lemma C.7; (b) holds because of Lemma C.4.

We conclude the proof by combining the two cases. □

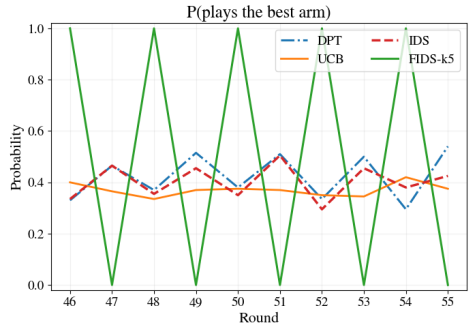
Lemma D.4.

$$\begin{aligned}
\sigma\sqrt{2K}\cdot\mathbb{E}\left[\sum_{t=1}^{n}\sqrt{g_t\left(\pi_t^{TS}\right)} - \sqrt{\cdot\tilde{g}_t\left(\pi_t^{TS}\right)}\right] &\overset{(a)}{\leqslant} \sigma\sqrt{2K}\cdot\mathbb{E}\left[\sum_{t=1}^{n}\sqrt{I_t(A^\star_t;(\tilde{A}_t,R_{t,\tilde{A}_t})|A^\star_{t+1:n})}\right] \\
&\leqslant \sigma\sqrt{2K}\cdot\mathbb{E}\left[\sum_{t=1}^{n}\sqrt{H_t(A^\star_t|A^\star_{t+1:n})}\right] \\
&\leqslant \sigma\sqrt{2K}\cdot\sum_{t=1}^{n}\sqrt{\mathbb{E}\left[H_t(A^\star_t|[A^\star_{t+1},\ldots,A^\star_n]]\right)} \\
&= \sigma\sqrt{2K}\cdot\sum_{t=1}^{n}\sqrt{H(A^\star_t|([A^\star_{t+1},\ldots,A^\star_n],\mathcal{F}_{t-1}))} \\
&\overset{(b)}{\leqslant} \sigma\sqrt{2K}\cdot\sum_{t=1}^{n}\sqrt{H(A^\star_t|[A^\star_{t+1},\ldots,A^\star_n])} \\
&\leqslant \sigma\sqrt{2K}\cdot\sqrt{n\cdot\sum_{t=1}^{n}H(A^\star_t|[A^\star_{t+1},\ldots,A^\star_n])} \\
&\overset{(c)}{=} \sigma\sqrt{2K\cdot n\cdot H([A^\star_1,\ldots,A^\star_n])} \qquad (17)
\end{aligned}$$

(a) uses Lemma D.3; (b) uses Corollary C.5; (c) uses Lemma C.2.



(a) The probability of choosing arm 3



(b) The probability of choosing the true best arm

E Numerical Results

Here we report additional numerical results for the one-step predictive environment in Sec. 5.1.

F Other Results

F.1 Proof of Lemma 3.1

Proof. Denote $E_{t,a} := \mathbb{E}_t\left[R_{t,a}\right]$, $g_{t,a} := I_t\left(\left[A^\star_{t+1}, \ldots, A^\star_n\right]; R_{t,a}\right)$, and $E_t := \left[E_{t,1}, \ldots, E_{t,K}\right]^T$, $g_t := \left[g_{t,1}, \ldots, g_{t,K}\right]^T$. Our objective function can then be written as $\rho(\pi_t) := \pi_t \cdot E_t + \sqrt{\Gamma \cdot \pi_t \cdot g_t}$, where $\Gamma = 2\sigma^2 K$ Consider some fixed maximizer $\pi^\star$ of Equation (5). The partial derivative of ρ with respect to $\pi_{t,a}$ at $\pi^\star$ is $\frac{\partial \rho}{\partial \pi_{t,a}}\left(\pi^\star\right) := E_{t,a} + \frac{\Gamma g_{t,a}}{2\sqrt{\Gamma g_t \cdot \pi^\star}}$.

Denote $d^\star := \max_{a \in \mathcal{A}} \frac{\partial \rho}{\partial \pi_{t,a}}\left(\pi^\star\right)$. The important observation here is that for any a such that $\pi^\star_a > 0$, $\frac{\partial \rho}{\partial \pi_{t,a}}\left(\pi^\star\right) = d^\star$, because otherwise one must be able to construct a new valid policy by transferring some probability from a to other arms with larger partial derivative and that will only increase the objective value. Therefore, for all $a \in B_t := \{a \in \mathcal{A} : \pi^\star_a > 0\}$,

$$E_{t,a} + \frac{\Gamma g_{t,a}}{2\sqrt{\Gamma g_t \cdot \pi^\star}} = d^\star \tag{18}$$

Reorder the set B_t such that $E_{t,a_1} \geqslant \cdots \geqslant E_{t,a_{|B_t|}}$. Note that there always exists a $\beta \in [0,1]$ such that $\sum_{a \in B_t} \pi^\star_a E_{t,a} = \beta E_{t,a_1} + (1-\beta) E_{t,a_{|B_t|}}$, which, by Equation (18), also implies that $\sum_{a \in B_t} \pi^\star_a g_{t,a} = \beta g_{t,a_1} + (1-\beta) g_{t,a_{|B_t|}}$. Therefore, we can construct a new policy that only assigns possibilities on the two arms a_1 and $a_{|B_t|}$, and this policy has the same objective value as $\pi^\star$ and thus optimal. □

F.2 Analytic Solutions

Depending on which round we are currently in, the mutual information term $I_t\left(\left[A^\star_{t+1}, \ldots, A^\star_n\right]; R_{t,a}\right)$ is equal to $I_t\left(\left[Z_1, Z_2\right]; R_{t,a}\right)$, if we are in the first half, or $I_t\left(Z_2; R_{t,a}\right)$ if we are in the second half. By independency, $I_t\left(\left[Z_1, Z_2\right]; R_{t,a}\right) = I_t\left(Z_1; R_{t,a}\right) + I_t\left(Z_2; R_{t,a}\right)$.

In the first half, by independency, when $a \neq 6$ $I_t\left(\left[Z_1, Z_2\right]; R_{t,a}\right) = I_t\left(Z_1; R_{t,a}\right)$, while $a = 6$ $I_t\left(\left[Z_1, Z_2\right]; R_{t,a}\right) = I_t\left(Z_2; R_{t,6}\right)$. The main problem is to calculate the posterior distribution. Essentially, we need to maintain $q_t(Z_1)$ and $w_t(Z_2)$.

if $a_t = 6$, $q_{t+1}(Z_1) = q_t(Z_1)$, and the posterior update rule for $w_{t+1}(Z_2)$ is

$$w_{t+1}(Z_2) \propto w_t(Z_2) \mathcal{N}\left(r_t; \frac{Z_2}{7}, 0.1^2\right)$$

While for $a_t \neq 6$, $w_{t+1}(Z_2) = w_t(Z_2)$, and the posterior update rule for $q_{t+1}(Z_1)$ is

$$q_{t+1}\left(Z_1\right) \propto q_t(Z_1) \mathcal{N}\left(r_t; \mu_z, 0.1^2\right)$$

This equation can be calculated analytically because z is finite. The posterior reward distribution can also be calculated by

$$\mathbb{P}_t\left(R_{t,a}=r\right)=\sum_z \mathbb{P}_t\left(R_{t,a}=r|Z_1=z\right)\mathbb{P}_t\left(Z_1=z\right)$$

In the second half, the posterior update rule for $w_{t+1}(Z_2)$ is

$$w_{t+1}(Z_2)\propto w_t(Z_2)\mathcal{N}\left(r_t;\mu_z,0.5^2\right)$$

F.3 Other Lemmas

Proposition F.1. *In Example 3.1, FIDS is the optimal policy and collects* $0.35\cdot n$ *cumulative reward, while TS and IDS collect* $0.325\cdot n$ *and* $0.3294\cdot n$ *cumulative reward respectively.*

Proof. Notice that $\mathbb{E}_t\left[R_{t,1}\right]=0.35, \mathbb{E}_t\left[R_{t,2}\right]=0.3$ for every t. Thus, the optimal policy is to choose arm 1 w.p. 1 at every t. FIDS will actually do that by noticing that $I_t\left(\left[A^\star_{t+1},\ldots,A^\star_n\right];R_{t,a}\right)=0$ for every t and a. And this will give cumulative reward $0.35\cdot n$.

TS will play each arm with equal probability at each round, and the cumulative reward is $0.5\cdot(0.35+0.3)\cdot n=0.325\cdot n$.

We calculate the mutual information terms $I_t\left(A^\star_t;(R_{t,1})\right)$ and $I_t\left(A^\star_t;(R_{t,2})\right)$ explicitly as shown below

$I_t\left(A^\star_t;(R_{t,1})\right)=\mathbb{P}(A^\star_t=1)\mathrm{KL}(\mathbb{P}(R_{t,1}\in\cdot|A^\star_t=1)||\mathbb{P}(R_{t,1}\in\cdot))+\mathbb{P}(A^\star_t=2)\mathrm{KL}(\mathbb{P}(R_{t,1}\in\cdot|A^\star_t=2)||\mathbb{P}(R_{t,1}\in\cdot))\approx 0.0055$

$I_t\left(A^\star_t;(R_{t,2})\right)=\mathbb{P}(A^\star_t=1)\mathrm{KL}(\mathbb{P}(R_{t,2}\in\cdot|A^\star_t=1)||\mathbb{P}(R_{t,2}\in\cdot))+\mathbb{P}(A^\star_t=2)\mathrm{KL}(\mathbb{P}(R_{t,2}\in\cdot|A^\star_t=2)||\mathbb{P}(R_{t,2}\in\cdot))\approx 0.0242$

And then we calculate the IDS policy by solving Equation (3) - $\pi^{\text{IDS}}_t(1)\approx 0.588$ and $\pi^{\text{IDS}}_t(2)\approx 0.412$. The cumulative reward will be $(0.588\cdot 0.35+0.412\cdot 0.3)\cdot n=0.3294\cdot n$. □

Lemma F.2. *Fix constants* $\alpha,\beta,\sigma\in\mathbb{R}$*, for the environment with* $\mu_{t,a}\in[\alpha,\beta]$ *and* $R_{t,a}\sim\mathcal{N}(\mu_{t,a},\sigma^2_{t,a})$ *where* $\sigma_{t,a}$ *is a deterministic non-negative constant no larger than* σ*. Then, conditioned on* $\mathcal{F}_t$*,*

$$R_{t,a}-\mathbb{E}\left[R_{t,a}\mid\mathcal{F}_t\right]$$

is $\sqrt{((b-a)/2)^2+\sigma^2}$ *sub-Gaussian.*

Proof. Note that $R_{t,a}-\mathbb{E}\left[R_{t,a}\mid\mathcal{F}_t\right]=\mu_{t,a}+\epsilon-\mathbb{E}\left[\mu_{t,a}\mid\mathcal{F}_t\right]$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\left(R_{t,a}-\mathbb{E}\left[R_{t,a}\mid\mathcal{F}_t\right]\right)\right)\mid\mathcal{F}_t\right]&=\mathbb{E}\left[\exp\left(\lambda\left(\mu_{t,a}+\epsilon-\mathbb{E}\left[\mu_{t,a}\mid\mathcal{F}_t\right]\right)\right)\mid\mathcal{F}_t\right]\\
&=\mathbb{E}\left[\exp\left(\lambda\left(\mu_{t,a}-\mathbb{E}\left[\mu_{t,a}\mid\mathcal{F}_t\right]\right)\right)\cdot\exp\left(\lambda\epsilon\right)\mid\mathcal{F}_t\right]\\
&=\mathbb{E}\left[\exp\left(\lambda\left(\mu_{t,a}-\mathbb{E}\left[\mu_{t,a}\mid\mathcal{F}_t\right]\right)\right)\mid\mathcal{F}_t\right]\cdot\mathbb{E}\left[\exp(\lambda\epsilon)\right]\\
&\leqslant\exp\left(\lambda^2(\beta-\alpha)^2/8\right)\cdot\exp\left(\lambda^2\sigma^2/2\right)\\
&=\exp\left(\frac{\lambda^2\left((\beta-\alpha)^2/4+\sigma^2\right)}{2}\right)
\end{aligned}$$

□