# Bandit Problems With Side Observations

Chih-Chun Wang, *Student Member, IEEE*, Sanjeev R. Kulkarni, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—An extension of the traditional two-armed bandit problem is considered, in which the decision maker has access to some side information before deciding which arm to pull. At each time $t$, before making a selection, the decision maker is able to observe a random variable $X_t$ that provides some information on the rewards to be obtained. The focus is on finding uniformly good rules (that minimize the growth rate of the inferior sampling time) and on quantifying how much the additional information helps. Various settings are considered and for each setting, lower bounds on the achievable inferior sampling time are developed and asymptotically optimal adaptive schemes achieving these lower bounds are constructed.**

***Index Terms*—Adaptive, asymptotic, allocation rule, inferior sampling time, efficient, side information, two-armed bandit.**

## I. Introduction

SINCE THE publication of [1], bandit problems have attracted much attention in various areas of statistics, control, learning, and economics (e.g., see [2]–[10]). In the classical two-armed bandit problem, at each time a player selects one of two arms and receives a reward drawn from a distribution associated with the arm selected. The essence of the bandit problem is that the reward distributions are unknown, and so there is a fundamental tradeoff between gathering information about the unknown reward distributions and choosing the arm we currently think is the best. A rich set of problems arises in trying to find an optimal/reasonable balance between these conflicting objectives (also referred to as learning versus control, or exploration versus exploitation).

We let $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$ denote the sequences of rewards from arms 1 and 2 in a two-armed bandit machine. In the traditional parametric setting, the underlying configurations/distributions of the arms are expressed by a pair of parameters $C_0 = (\theta_1, \theta_2)$ such that $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$ are independent and identically distributed (i.i.d.) with distribution $(F_{\theta_1}, F_{\theta_2})$, where $\{F_\theta\}$ is a known family of distributions parametrized by $\theta$. The goal is to maximize the sum of the expected rewards. Results on achievable performance have been obtained for a number of variations and extensions of the basic problem defined in [9] (e.g., see [11]–[17]).

In this paper, we consider an extension of the classical two-armed bandit where we have access to side information before making our decision about which arm to pull. Suppose at time $t$, in addition to the history of previous decisions, outcomes, and observations, we have access to a side observation $X_t$ to help us make our current decision. The extent to which this side observation can help depends on the relationship of $X_t$ to the reward distributions of $Y_t^1$ and $Y_t^2$.

Previous work on bandit problems with side observations includes [18]–[22]. Woodroofe [21] considered a one-armed bandit in a Bayesian setting, and constructed a simple criterion for asymptotically optimal rules. Sarkar [20] extended the side information model of [21] to the exponential family. In [19], Kulkarni considered classes of reward distributions and their effects on performance using results from learning theory. Most of the previous work with side observations is on one-armed bandit problems, which can be viewed as a special case of the two-armed setting by letting arm 2 always return zero.

In contrast with this previous work, we consider various general settings of side information for a two-armed bandit problem. Our focus is on providing both lower bounds and bound-achieving algorithms for the various settings. The results and proofs are very much along the lines of [8] and subsequent works as in [11]–[15].

We now describe the settings considered in this paper.

1) **Direct Information:** In this case, $X_t$ provides information directly about the underlying configuration $C_0 = (\theta_1, \theta_2)$, which allows a type of separation between the learning and control. This has a dramatic effect on the achievable inferior sampling time. Specifically, estimating $(\theta_1, \theta_2)$ by observing $\{X_\tau\}$, and using the estimate $(\hat{\theta}_1, \hat{\theta}_2)$ to make the decision, results in bounded expected inferior sampling time.

If the distribution of $\{X_\tau\}$ is not a function of $C_0$, we are not able to learn $C_0$ through $\{X_\tau\}$. However, different values of the side observation $X_t$ will result in different conditional distributions of the rewards $Y_t^i$. By exploiting this new structure (observing $X_t$ in advance), we can hope to do better than the case without any side observation.

An interpretation of the aforementioned scenario (constant distribution on $\{X_\tau\}$) is that a two-armed bandit with the side observations drawn from a finite set $\{x_1, x_2, \ldots, x_n\}$ can be viewed as a set of $n$ different two-armed sub-bandit machines indexed from $x_1$ to $x_n$. The player does not know the order of sub-machines he is going to play, which is determined by rolling a die with $n$ faces. However, by observing $X_t$, the player knows which machine (out of the $n$ different ones) he is facing now before selecting which arm to play. The connection between these sub-machines is that they share the same common configuration pair $(\theta_1, \theta_2)$, so that the rewards observed from one machine provide information on the common $(\theta_1, \theta_2)$, which can then be applied to *all* of the others (different values of $X_t$).

Manuscript received November 15, 2002; revised November 1, 2004. Recommended by Associate Editor D. Li. This work was supported in part by the National Science Foundation under Grants ANI-0338807 and ECS-9873451, the Army Research Office under Contract DAAD19-00-1-0466, the Office of Naval Research under Grant N00014-03-1-0102, and the New Jersey Center for Pervasive Information Technologies.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: chihw@princeton.edu; kulkarni@princeton.edu; poor@princeton.edu).

Digital Object Identifier 10.1109/TAC.2005.844079

This is the key aspect that makes this setup distinct from simply having many independent bandit problems with random access opportunity.

We consider the following three cases of different relationships among the most rewarding arm, $C_0$, and $X_t$.

2) **For all possible $C_0$, the best arm is a function of $X_t$**: That is, $\forall(\theta_1,\theta_2),\exists x_1,x_2$ such that at time $t$, arm 1 yields higher expected reward conditioned on $X_t = x_1$ while arm 2 is preferred when $X_t = x_2$. Surprisingly, we exhibit an algorithm that achieves *bounded* expected inferior sampling time in this case. Woodroofe's result [21] can then be viewed as a special case of this scenario.

3) **For all possible $C_0$, the best arm is not a function of $X_t$**: In this case, for *all* configurations $(\theta_1,\theta_2)$, one of the arms is always preferred regardless of the value of $X_t$. Since the conditional reward distributions are functions of $X_t$, the intuition is that we can postpone our learning until it is most advantageous to us. We show that, asymptotically, our performance will be governed by the most "informative" bandit (among the different values taken on by $X_t$).

4) **Mixed Case:** This is a general case that combines the previous two, and contains the main contribution of this paper. For some possible configurations, one arm may always be preferred (for all $X_t$), while for other possible configurations, the preferred arm is a function of $X_t$. We exhibit an algorithm that achieves the best possible in either case. That is, if the best arm is a function of $X_t$, it achieves bounded expected inferior sampling time as in case 2), while if the underlying configuration is such that one arm is always preferred, then we get the results of case 3).

This paper is organized as follows. In Section II, we introduce the general formulation. In Section III, we provide background on the asymptotic analysis of traditional bandit problems (without side observations). In Sections IV through VII, we consider the previous four cases respectively. The results are included in each section, while details of the proofs are provided in the Appendix.

## II. General Formulation

Consider the two-armed bandit problem defined as follows. Suppose we have two sequences of (real-valued) random variables (r.v.'s), $\{Y_\tau^i\}_{i=1,2}$, and an i.i.d. side observation sequence $\{X_\tau\}$, taking values in $\mathbf{X}\subset\mathbb{R}$. $\{Y_\tau^i\}$ denotes the reward sequence of arm $i$ while $X_t$ is the side information observed at time $t$ before making the decision. The formal parametric setting is as follows. For each configuration pair $C_0=(\theta_1,\theta_2)$ and each $i$, the sequence of vectors $(X_t,Y_t^i)$ is i.i.d. with joint distribution $G_{C_0}(dx)F_{\theta_i}(dy\,|\,x)$, where the families $\{G_C\}_{C\in\mathbf{\Theta}^2}$ and $\{F_\theta(\cdot\,|\,\cdot)\}_{\theta\in\mathbf{\Theta}}$ are known to the player, but the true value of the corresponding index $C_0$ must be learned through experiments. For notational simplicity, we further assume that the parameter set $\mathbf{\Theta}$ is a set of real numbers.

Note that the concept of the i.i.d. bandit is now extended to the assumption that the vector sequence $\{(X_\tau,Y_\tau^i)\}_t$ is i.i.d. The unconditioned marginal sequence $\{Y_\tau^i\}$ remains i.i.d. However, rather than the unconditional marginals, the player is now facing the conditional distribution of $Y_t^i$, which is a function of the observed side information $X_t$ (and is not identically distributed given different $X_t$).

The goal is to find an adaptive allocation rule $\{\phi_\tau\}$ to maximize the growth rate of the expected reward

$$\mathsf{E}_{C_0}\{W_\phi(t)\} := \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t}\left(1_{\{\phi_\tau=1\}}Y_\tau^1 + 1_{\{\phi_\tau=2\}}Y_\tau^2\right)\right\}$$

or, equivalently, to minimize the growth rate of the expected inferior sampling time,[1] namely $\mathsf{E}_{C_0}\{T_{\inf}(t)\}$. To be more explicit, at any time $t$, $\phi_t$ takes a value in $\{1,2\}$ and depends only on the past rewards $(\tau<t)$ and the current side observation $X_t$.

We define a uniformly good rule as follows.

*Definition 1 (Uniformly Good Rules):* An allocation rule $\{\phi_\tau\}$ is uniformly good if for all $C\in\mathbf{\Theta}^2, \mathsf{E}_C\{T_{\inf}(t)\} = o(t^\alpha), \forall\alpha>0$.

In what follows, we consider only uniformly good rules and regard other rules as uninteresting. Necessary notation and several quantities of interest are defined in Table I. We assume that all the given expectations exist and are finite.

## III. Traditional Bandits

Under the general formulation provided in Section II, the traditional non-Bayesian, parametric, infinite horizon, two-armed bandit is simply a degenerate case, i.e., the traditional bandit problem is equivalent to having only one element in $\mathbf{X}$ (say $\mathbf{X}=\{x_0\}$). This formulation of traditional bandit problems is identical to the two-armed case of [14], [8], and [9]. For simplicity, the argument $x_0$ can be omitted in this traditional setting, i.e., $M_C := M_C(x_0), \mu_\theta := \mu_\theta(x_0)$, $I(\theta_1,\theta_2) := I(\theta_1,\theta_2\,|\,x_0)$, etc.

The main contribution of [14], [8], and [9] is the asymptotic analysis stated via the following two theorems.

*Theorem 1 ($\log t$ Lower Bound):* For any uniformly good rule, $\{\phi_\tau\}, T_{\inf}(t)$ satisfies

$$\lim_{t\to\infty}\mathsf{P}_{C_0}\left(T_{\inf}(t)\ge\frac{(1-\epsilon)\log t}{K_{C_0}}\right)=1 \qquad \forall\epsilon>0$$

and

$$\liminf_{t\to\infty}\frac{\mathsf{E}_{C_0}\{T_{\inf}(t)\}}{\log t}\ge\frac{1}{K_{C_0}}$$

[1] In the literature of bandit problems, the term "regret" is more typically used rather than the inferior sampling time. For traditional two-armed bandits, the regret is defined as

$$\text{regret} := t\cdot\max\{\mu_{\theta_1},\mu_{\theta_2}\} - \mathsf{E}_{C_0}\{W_\phi(t)\}$$

the difference between the best possible reward and that of the strategy of interest $\{\phi_\tau\}$. The relationship between the regret and $T_{\inf}(t)$ is as follows:

$$\text{regret} = \left|\mu_{\theta_1}-\mu_{\theta_2}\right|\cdot\mathsf{E}_{C_0}\{T_{\inf}(t)\}.$$

For greater simplicity in the discussion of bandit problems with side observations, we consider $T_{\inf}(t)$ rather than the regret.

TABLE I
GLOSSARY

| Not'n | Description |
|---|---|
| $G_C(dx)$ | The marginal distribution of the i.i.d. $\{X_\tau\}$ under configuration $C$. |
| $F_{\theta_i}(dy\|x)$ | The conditional distribution of the reward of arm $i$, $Y_t^i$, under parameter $\theta_i$. |
| $\mu_\theta(x)$ | The conditional expectation of the reward, $\mu_\theta(x) = \mathsf{E}_\theta\{Y\|x\} = \int yF_\theta(dy\|x)$. |
| $1(C_0)$, $2(C_0)$ | The first and the second coordinates of the configuration pair $C_0$, i.e. $1(C_0) = \theta_1$, $2(C_0) = \theta_2$. For example: $F_{1(C_0)}(dy\|x) = F_{\theta_1}(dy\|x)$ and $\mu_{2(C_0)}(x) = \mu_{\theta_2}(x)$. |
| $M_C(x)$ | The index of the preferred arm, i.e. $\arg\max_{i=1,2}\{\mu_{i(C)}(x)\}$. |
| $\phi_t$ | The decision rule taking values in $\{1,2\}$ and depending only on the past outcomes and the current side information $X_t$. |
| $T_i(t)$ | The total number of samples taken on arm $i$ up to time $t$, $T_i(t) = \sum_{\tau=1}^{t} 1_{\{\phi_\tau = i\}}$. |
| $T_{inf}(t)$ | The total number of samples taken on the inferior arm up to time $t$: $T_{inf}(t) = \sum_{\tau=1}^{t} 1_{\{\phi_\tau \neq M_{C_0}(X_\tau)\}}$. |
| $I(P,Q)$ | The Kullback-Leibler (K-L) information number between distributions $P$ and $Q$: $I(P,Q) = \mathsf{E}_P\left\{\log\left(\frac{dP}{dQ}\right)\right\}$. |
| $I(\theta_1,\theta_2\|x)$ | The conditional K-L information number: $I(\theta_1,\theta_2\|x) = I(F_{\theta_1}(\cdot\|x), F_{\theta_2}(\cdot\|x))$. |

where $K_{C_0}$ is a constant depending on $C_0$. If $M_{C_0} = 2$, then $T_{\inf}(t) = T_1(t)$ and $K_{C_0}$ is defined[2] as follows:

$$K_{C_0} = \inf\{I(\theta_1,\theta): \forall\theta, \mu_\theta > \mu_{\theta_2}\}. \tag{1}$$

The expression for $K_{C_0}$ for the case in which $M_{C_0} = 1$ can be obtained by symmetry.

*Theorem 2 (Asymptotic Tightness):* Under certain regularity conditions,[3] the aforementioned lower bound is asymptotically tight. Formally stated, given the distribution family $\{F_\theta\}$, there exists a decision rule $\{\phi_\tau\}$ such that for all $C_0 = (\theta_1,\theta_2) \in \Theta^2$

$$\limsup_{t\to\infty} \frac{\mathsf{E}_{C_0}\{T_{\inf}(t)\}}{\log t} \leq \frac{1}{K_{C_0}}$$

where $K_{C_0}$ is the same as in Theorem 1.

The intuition behind the $\log t$ lower bound is as follows. Suppose $M_{C_0} = 2$ and consider another configuration $C' = (\theta,\theta_2)$ such that $M_{C'} = 1$. It can be shown that if under configuration $C_0 = (\theta_1,\theta_2)$, $\mathsf{E}_{C_0}\{T_{\inf}(t)\}$ is less than the $\log t$ lower bound, $\mathsf{E}_{C'}\{T_{\inf}(t)\}$ must be greater than $o(t^\alpha)$ for some $\alpha > 0$, which contradicts the assumption that $\{\phi_\tau\}$ is uniformly good.

## IV. DIRECT INFORMATION

### A. Formulation

In this setting, the side observation $X_t$ directly reveals information about the underlying configuration pair $C_0 = (\theta_1,\theta_2)$ in the following way.

Dependence: $G_{C_1} = G_{C_2}$ iff $C_1 = C_2$.

As a result, observing the empirical distribution of $X_t$ gives us useful information about the underlying parameter pair $C_0$. Thus, this is a type of identifiability condition.

[2]Throughout this paper, we will adopt the conventions that the infimum of the null set is $\infty$, and $(1/\infty) = 0$.

[3]If the parameter set is finite, Theorem 2 always holds. If $\Theta$ is the set of reals, the required regularity conditions are on the unboundedness and the continuity of $\mu_\theta$ w.r.t. $\theta$ and on the continuity of $I(\theta_1,\theta)$ w.r.t. $\mu_\theta$.

Examples:

- $\Theta = (0, 0.5)$ and $\mathbf{X} = \{x_1, x_2, x_3\}$

$$\mathsf{P}_{(\theta_1,\theta_2)}(X_t = x_k) = \begin{cases} \theta_k, & \text{if } k = 1,2 \\ 1-\theta_1-\theta_2, & \text{otherwise} \end{cases}.$$

- $\Theta = (0,\infty)$ and $\mathbf{X} = [0,1]$. $X_t$ is beta distributed with parameters $(\theta_1,\theta_2)$.

### B. Scheme With Bounded $\mathsf{E}_{C_0}\{T_{\inf}(t)\}$

Consider the following condition.

*Condition 1:* For any fixed $C_0$

$$\inf\{\rho(G_{C_0}, G_{C_e}) : C_e \in \Theta^2, \exists x, M_{C_e}(x) \neq M_{C_0}(x)\} > 0$$

where $\rho$ denotes the Prohorov metric[4] on the space of distributions. Two examples satisfying Condition 1 are as follows.

- *Example 1:* $\mathbf{X}$ is finite, and $\forall x \in \mathbf{X}, \mu_\theta(x)$ is continuous with respect to (w.r.t.) $\theta$.
- *Example 2:* $F_\theta(\cdot\,|\,x) \sim \mathcal{N}(\theta x, 1)$ is a Gaussian distribution with mean $\theta x$ and variance 1.

Under this condition, we obtain the following result.

*Theorem 3 (Bounded $\mathsf{E}_{C_0}\{T_{\inf}(t)\}$):* If Condition 1 is satisfied, then there exists an allocation rule $\{\phi_\tau\}$, such that $\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\inf}(t)\} < \infty$ and $\lim_{t\to\infty} T_{\inf}(t) < \infty$ a.s.

- Note: The information directly revealed by $X_t$ helps the sequential control scheme surpass the $\log t$ lower bound stated in Theorem 1. This significant improvement (bounded expected inferior sampling time) is due to the fact that the dilemma between learning and control no longer exists in the direct information case.

We provide a scheme achieving bounded $\mathsf{E}_{C_0}\{T_{\inf}(t)\}$ as in Algorithm 1

**Algorithm 1** $\phi_{t+1}$, the decision at time $t+1$ (after observing $X_{t+1}$)

1: Construct

$$\mathbf{C}_{t+1} := \{C \in \Theta^2 : \rho(G_C, L_X(t+1)) \leq \inf_{C\in\Theta^2} \rho(G_C, L_X(t+1)) + \frac{1}{t+1}\},$$

where $L_X(t+1)$ is the empirical measure of the side observations $\{X_\tau\}$ until time $t+1$, and $\rho$ is the Prohorov metric as before.

2: Arbitrarily pick $\hat{C}_{t+1} \in \mathbf{C}_{t+1}$, and set $\phi_{t+1} = M_{\hat{C}_{t+1}}(X_{t+1})$.

of which a detailed analysis is given in Appendix II.

## V. BEST ARM AS A FUNCTION OF $X_t$

For all of the following sections (Sections V–VII), we consider only the case in which observing $X_t$ will not reveal any information about $C_0$, but only reveals information about the upcoming reward $Y_t^i$, that is

- $G_{C_0}$ does not depend on the value of $C_0$; we use $G := G_{C_0}$ as shorthand notation.

Three further refinements regarding the relationship between $M_C(x)$ and $x$ will be discussed separately (each in one section).

[4]A definition of the Prohorov metric is stated in Appendix I.

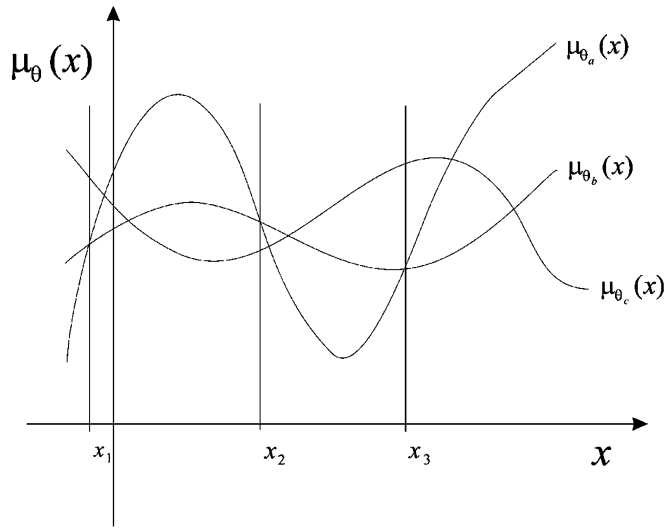



Fig. 1. Best arm at time $t$ *always* depends on the side observation $X_t$. That is, for any possible pair $(\theta_1, \theta_2)$ the two curves, $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$, (w.r.t. $x$) always intersect each other.

### A. Formulation

In this section, we assume that for all possible $C$, the side observation $X_t$ is *always* able to change the preference order as shown in Fig. 1. That is

- for all $C \in \mathbf{\Theta}^2$, there exist $x_1$ and $x_2$ such that $M_C(x_1) = 1$ and $M_C(x_2) = 2$.

The needed regularity conditions are as follows.

1) $\mathbf{X}$ is a finite set and $\mathsf{P}_G(X_t = x) > 0$ for all $x \in \mathbf{X}$.
2) $\forall \theta_1, \theta_2, x, I(\theta_1, \theta_2 \,|\, x)$ is strictly positive and finite.
3) $\forall x, \mu_\theta(x)$ is continuous w.r.t. $\theta$.

The first condition embodies the idea of treating $X_t$ as the index of several different bandit machines, which also simplifies our proof. The second condition is to ensure that all these different bandit problems are nontrivial, with *nonidentical* pairs of arms.

Example:

- $\mathbf{\Theta} = (0, \infty), \mathbf{X} = \{-1, 1\}$, and the conditional reward distribution $F_\theta(\cdot \,|\, x) \sim \mathcal{N}(\theta x, 1)$.

### B. Scheme With Bounded $\mathsf{E}_{C_0}\{T_{\text{inf}}(t)\}$

*Theorem 4 (Bounded $\mathsf{E}_{C_0}\{T_{\text{inf}}(t)\}$):* If the aforementioned conditions are satisfied, there exists an allocation rule $\{\phi_\tau\}$ such that

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\text{inf}}(t)\} < \infty.$$

Such a rule is obviously uniformly good.

- Note: Although the side observation $X_t$ does not reveal any information about $C_0$ in this setting, the alternation of the best arm as the i.i.d. $X_t$ takes on different values $x$ makes it possible to always perform the control part $\phi_t = M_{\hat{C}_{t-1}}(X_t)$, and simultaneously sample both arms often enough. Since the information about both arms will be implicitly revealed [through the alternation of $M_{C_0}(X_t)$], the dilemma of learning and control no longer exists, and a significant improvement $(\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\text{inf}}(t)\} < \infty)$ is obtained over the $\log t$ lower bound in Theorem 1.

We construct an allocation rule with bounded $\mathsf{E}_{C_0}\{T_{\text{inf}}(t)\}$ given as Algorithm 2.

**Algorithm 2** $\phi_{t+1}$, the decision at time $t+1$

**Variables:** Denote $T_i^x(t)$ as the total number of time instants until time $t$ when arm $i$ has been pulled and $X_\tau = x$, i.e.

$$T_i^x(t) := \sum_{\tau=1}^{t} 1_{\{X_\tau = x, \phi_\tau = i\}},$$

and define $x_i^\star := \arg\max_x \{T_i^x(t)\}$ and $T_i^{x^\star}(t) := \max_x \{T_i^x(t)\}$. Construct

$$\begin{aligned} \mathbf{C}_t \quad := \quad & \{C = (\theta_1, \theta_2) \in \mathbf{\Theta}^2 : \\ & \sigma(C, t) \le \inf\{\sigma(C, t) : C \in \mathbf{\Theta}^2\} + \frac{1}{t}\}, \end{aligned}$$

$$\begin{aligned} \text{with} \qquad \sigma(C, t) \quad := \quad & \rho(F_{1(C)}(\cdot|x_1^\star), L_1^{x^\star}(t)), \\ & + \rho(F_{2(C)}(\cdot|x_2^\star), L_2^{x^\star}(t)), \end{aligned}$$

where $L_i^x(t)$ is the empirical measure of rewards sampled from arm $i$ at those time instants $\tau \le t$ when $X_\tau = x$. (As before $\rho$ is the Prohorov metric.) Arbitrarily choose $\hat{C}_t \in \mathbf{C}_t$.

**Algorithm:**

1: **if** $t+1 \le 6$ **then**
2: $\quad \phi_{t+1} = (t \bmod 2) + 1$.
3: **else if** $\exists i$ such that $T_i(t) < \sqrt{t+1}$ **then**
4: $\quad \phi_{t+1} = i$.
5: **else**
6: $\quad \phi_{t+1} = M_{\hat{C}_t}(X_{t+1})$.
7: **end if**

(Note that Line 1 guarantees that there is only one $i$ such that $T_i(t) < \sqrt{t+1}$.)

The intuition as to why the proposed scheme has bounded $\mathsf{E}_{C_0}\{T_{\text{inf}}(t)\}$ is as follows. The forced sampling $T_i(t) < \sqrt{t+1}$ ensures there are enough samples on both arms, which implies good enough estimates of $C_0$. Based on the good enough estimates, the myopic action of sampling the seemingly better arm $\phi_{t+1} = M_{\hat{C}_t}(X_{t+1})$ will result in very few inferior samplings. Unlike the traditional two-armed bandits, in this scenario, the best arm $M_{C_0}(x)$ varies from one outcome of $X_t$ to the other. Therefore, the myopic action and the even appearances of the i.i.d. $\{X_\tau\}$ will eventually make both $T_1(t)$ and $T_2(t)$ grow linearly with the elapsed time $t$, and the forced sampling should occur only rarely. This situation differs significantly from the traditional bandits, where the forced sampling will inevitably make the $T_{\text{inf}}(t)$ of the order of $\sqrt{t}$, which is an undesired result.

A detailed proof of the boundedness of $\mathsf{E}_{C_0}\{T_{\text{inf}}(t)\}$ for this scheme is provided in Appendix III.

## VI. Best Arm Is Not a Function of $X_t$

### A. Formulation

Besides the assumption of constant $G$, in this section, we consider the case in which for all $C \in \mathbf{\Theta}^2$, $M_C(x)$ is not a function of $x$, and we thus can use $M_C := M_C(x)$ as shorthand notation. Fig. 2 illustrates this situation.

The needed regularity conditions are similar to those in Section V.

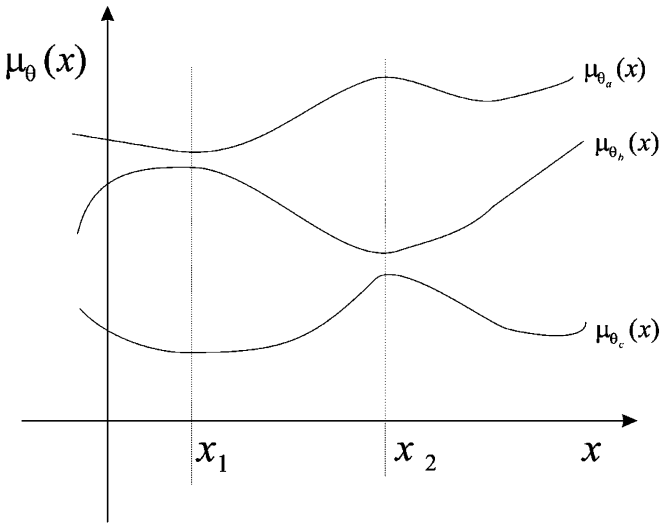



Fig. 2. Best arm at time $t$ *never* depends on the side observation $X_t$. That is, for any possible pair $(\theta_1, \theta_2)$, the two curves $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ do not intersect each other. However, in this case, we can postpone our sampling to the most informative time instants.

1) $\mathbf{X}$ is a finite set and $\mathsf{P}_G(X_t = x) > 0$ for all $x \in \mathbf{X}$.
2) $\forall \theta_1, \theta_2, x, I(\theta_1, \theta_2 \,|\, x)$ is strictly positive and finite.

In this case, one arm is always better than the other no matter what value of $X_t$ occurs. The conflict between learning and control still exists. As expected, the growth rate of the expected inferior sampling time is again lower bounded by $\log t$, but with the additional help of $X_t$ we can see improvements over the traditional bandit problems.

To greatly simplify the notation, we also assume that

4) for all $x$, the conditional expected reward $\mu_\theta(x)$ is strictly increasing w.r.t. $\theta$.

This condition gives us the notational convenience that the order of $(\mu_{\theta_1}(x), \mu_{\theta_2}(x))$ is simply the same as the order of $(\theta_1, \theta_2)$.

Example:

- $\mathbf{\Theta} = (1, \infty), \mathbf{X} = \{1, 2, 3\}$, and the conditional reward distribution $F_\theta(\cdot \,|\, x) \sim \mathcal{N}(\theta x, 1)$.

### B. Lower Bound

*Theorem 5 ($\log t$ Lower Bound):* Under the previou assumptions, for any uniformly good rule $\{\phi_\tau\}$, $T_{\inf}(t)$ satisfies

$$\lim_{t\to\infty} \mathsf{P}_{C_0}\left(T_{\inf}(t) \geq \frac{(1-\epsilon)\log t}{K_{C_0}}\right) = 1 \qquad \forall \epsilon > 0$$

and

$$\liminf_{t\to\infty} \frac{\mathsf{E}_{C_0}\{T_{\inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}} \tag{2}$$

where $K_{C_0}$ is a constant depending on $C_0$. If $M_{C_0} = 2$, then $T_{\inf}(t) = T_1(t)$. The constant $K_{C_0}$ can be expressed as follows:

$$K_{C_0} = \inf_{\theta:\theta>\theta_2} \sup_{x\in\mathbf{X}} \{I(\theta_1, \theta \,|\, x)\}. \tag{3}$$

The expression for $K_{C_0}$ for the case in which $M_{C_0} = 1$ can be obtained by symmetry.

Note 1: If the decision maker is not able to access the side observation $X_t$, the player will then face the *unconditional* reward distribution $\int_x F_{\theta_i}(dy \,|\, x) G(dx)$ rather than $F_{\theta_i}(dy \,|\, x)$. Let $I(\theta_1, \theta_2)$ denote the Kullback–Leibler information between the *unconditional* reward distributions. By the convexity of the Kullback–Leibler information, we have

$$\sup_x I(\theta_1, \theta \,|\, x) \geq \int_x I(\theta_1, \theta \,|\, x) G(dx) = I(\theta_1, \theta).$$

This shows that the new constant in front of $\log t$, in (3), is no larger than the corresponding constant in (1), and the additional side information $X_t$ generally improves the decision made in the bandit problem. As we would expect, Theorem 5 collapses to Theorem 1 when $|\mathbf{X}| = 1$.

Note 2: This situation is like having several related bandit machines, whose reward distributions are all determined by the common configuration pair $(\theta_1, \theta_2)$. The information obtained from one machine is also applicable to the other machines. If arm 2 is always better than arm 1, we wish to sample arm 2 most of the time (the control part), and force sample arm 1 once in a while (the learning part). With the help of the side information $X_t$, we can postpone our forced sampling (learning) to the most informative machine $X_t = x$. As a result, the constant in the $\log t$ lower bound in Theorem 1 has been further reduced to this new $(1/(K_{C_0}))$.

A detailed proof of Theorem 5 is provided in Appendix IV.

### C. Scheme Achieving the Lower Bound

Consider the additional conditions as follows.

1) $\mathbf{\Theta}$ is finite.
2) A saddle point for $K_{C_0}$ exists; that is, for all $\theta_1 < \theta_2$

$$\inf_{\theta:\theta>\theta_2} \sup_x I(\theta_1, \theta \,|\, x) = \sup_x \inf_{\theta:\theta>\theta_2} I(\theta_1, \theta \,|\, x).$$

With these conditions, we construct a $\log t$-lower-bound-achieving scheme $\{\phi_\tau\}$, which is inspired by [12]. The following notation and quantities are necessary in the expression of $\{\phi_\tau\}$.

- Denote $\hat{C}_t := (\theta^\alpha, \theta^\beta)$. Instead of the traditional $(\hat{\theta}_1, \hat{\theta}_2)$ representation, we use $(\theta^\alpha, \theta^\beta)$. Based on this representation, we are able to derive the following useful notation:

$$\begin{aligned} \theta^{\alpha\wedge\beta} &:= \min(\theta^\alpha, \theta^\beta) \\ \alpha\wedge\beta &:= \arg\min(\theta^\alpha, \theta^\beta) \\ \theta^{\alpha\vee\beta} &:= \max(\theta^\alpha, \theta^\beta) \\ \alpha\vee\beta &:= \arg\max(\theta^\alpha, \theta^\beta). \end{aligned}$$

  For instance, if $\theta^\alpha < \theta^\beta$, $\mu_{\theta^\alpha}(x) = \mu_{\theta^{\alpha\wedge\beta}}(x)$; arm $\alpha\wedge\beta$ represents arm 1; $Y_t^{\alpha\vee\beta}$ is the reward of arm 2; and $T_{\inf}(t) = T_{\alpha\wedge\beta}(t) = T_1(t)$.
- Choose an $\epsilon$ such that

$$0 < \epsilon < (1/2)\min\{\rho(F_\theta(\cdot \,|\, x), F_\vartheta(\cdot \,|\, x)) : \forall x \in \mathbf{X}, \theta \neq \vartheta \in \mathbf{\Theta}\}$$

  where $\rho$ is the Prohorov metric. The whole system is *well-sampled* if there exists a unique estimate $\hat{C}_t = (\theta^\alpha, \theta^\beta)$, such that the empirical measure $L_i^x(t)$ falls

into the $\epsilon$-neighborhood of $F_{i(\hat{C}_t)}(\cdot\,|\,x)$, for all $x \in \mathbf{X}$ and $i \in \{1,2\}$. That is

$$\exists \hat{C}_t \quad \text{s.t. } \rho\left(L_i^x(t), F_{i(\hat{C}_t)}(\cdot\,|\,x)\right) < \epsilon \qquad \forall x \in \mathbf{X},\, i \in \{1,2\}.$$

- For any estimate $\hat{C}_t = (\theta^\alpha, \theta^\beta)$, define the most informative bandit according to $\hat{C}_t$ as

$$x^*(\hat{C}_t) := \arg\max_x \inf_{\theta:\theta>\theta^{\alpha\vee\beta}} I(\theta^{\alpha\wedge\beta}, \theta, x)$$

and $\Lambda_t(\hat{C}_t, \theta)$ to be the conditional likelihood ratio between the seemingly inferior arm $\theta^{\alpha\wedge\beta}$ and the competing parameter $\theta$

$$\Lambda_t(\hat{C}_t, \theta) := \prod_{m=1}^{T^{x^*}_{\alpha\wedge\beta}(t)} \frac{F_{\theta^{\alpha\wedge\beta}}\left(dY^{\alpha\wedge\beta}_{\tau_{x^*}(m)} \,\middle|\, x^*(\hat{C}_t)\right)}{F_\theta\left(dY^{\alpha\wedge\beta}_{\tau_{x^*}(m)} \,\middle|\, x^*(\hat{C}_t)\right)}$$

where $\tau_{x^*}(m)$ denotes the time instant of the $m$th pull of arm $\alpha\wedge\beta$ when the side observation $X_\tau = x^*(\hat{C}_\tau)$.
- Set a total number of $|\mathbf{X}| + |\mathbf{\Theta}|^2 + |\mathbf{\Theta}|^3$ counters, including $|\mathbf{X}|$ counters, named "ctr$(x)$;" $|\mathbf{\Theta}|^2$ counters, named "ctr$(\hat{C})$" for all possible $\hat{C} \in \mathbf{\Theta}^2$; and $|\mathbf{\Theta}|^3$ counters, named "ctr$(\hat{C}, \theta)$" for all possible $\hat{C}$ and $\theta$. Initially, all counters are set to zero.

*Theorem 6 (Asymptotic Tightness):* With the previous conditions, the scheme described in Algorithm 3

**Algorithm 3** $\phi_{t+1}$, the decision at time $t+1$

**if** there exists $i \in \{1,2\}$ and $x \in \mathbf{X}$ such that $T_i^x(t) = 0$, **then** {Cond0}
  $\phi_{t+1} \leftarrow (t+1) \bmod 2$.
**else if** the whole system is not *well-sampled* or $\theta^\alpha = \theta^\beta$, **then** {Cond1}
  ctr$(X_{t+1}) \leftarrow$ ctr$(X_{t+1}) + 1$ and $\phi_{t+1} \leftarrow$ ctr$(X_{t+1}) \bmod 2$.
**else if** $\theta^{\alpha\vee\beta} = \bar{\theta} := \max \mathbf{\Theta}$, **then** {.....Cond2}
  $\phi_{t+1} \leftarrow 1$ if it is $\theta^\alpha = \bar{\theta}$. Otherwise, $\phi_{t+1} \leftarrow 2$.
**else** {.....Cond3}
  ctr$(\hat{C}_t) \leftarrow$ ctr$(\hat{C}_t) + 1$.
  **if** ctr$(\hat{C}_t)$ is odd, **then** {.....Cond3a}
    $\phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$.
  **else** {.....Cond3b}
    $\theta^* \leftarrow \arg\min\{\Lambda_t(\hat{C}_t, \theta) : \theta > \theta^{\alpha\vee\beta}\}$.
    **if** $X_{t+1} = x^*(\hat{C}_t)$ **then** {.....Cond3b1}
      **if** $\Lambda_t(\hat{C}_t, \theta^*) \le t(\log t)^2$, **then** {.....Cond3b1a}
        ctr$(\hat{C}_t, \theta^*) \leftarrow$ ctr$(\hat{C}_t, \theta^*) + 1$.
        **if** $\exists k \in \mathbb{N}$ s.t. ctr$(\hat{C}_t, \theta^*) = k^2$, **then** {.....Cond3b1a1}
          $\phi_{t+1} \leftarrow k \bmod 2$.
        **else** {.....Cond3b1a2}
          $\phi_{t+1} \leftarrow 3 - M_{\hat{C}_t}(X_{t+1})$.
        **end if**
      **else** {.....Cond3b1b}
        $\phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$.
      **end if**
    **else** {.....Cond3b2}
      $\phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$.
    **end if**
  **end if**
**end if**

achieves the $\log t$ lower bound (2), so that this $\{\phi_\tau\}$ is uniformly good and asymptotically optimal.

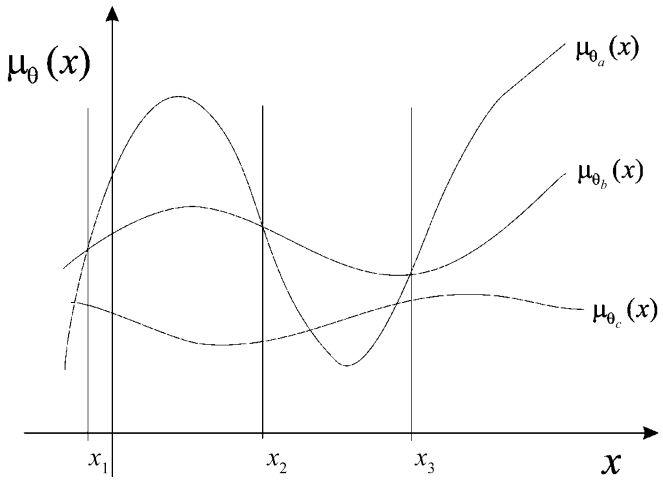



Fig. 3. If $(\theta_1, \theta_2) = (\theta_a, \theta_b)$, the best arm depends on $x$, i.e., $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ intersect each other as in Section V. If $(\theta_1, \theta_2) = (\theta_b, \theta_c)$, the best arm does not depend on $x$, i.e., $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ do not intersect each other as in Section VI.

A complete analysis is provided in Appendix V.

## VII. Mixed Case

The main difference between Sections V and VI is that in one case, for all possible $C_0, X_t$ *always* changes the preference order, while in the other, for all possible $C_0, X_t$ *never* changes the order. A more general case is a mixture of these two. In this section, we consider this mixed case, which is the main result of this paper.

### A. Formulation

Besides the assumption of constant $G$, in this section, we consider the case in which for some $C \in \mathbf{\Theta}^2, M_C(x)$ is not a function of $x$. For the remaining $C$, there exist $x_1$ and $x_2$ s.t. $M_C(x_1) = 1$ and $M_C(x_2) = 2$. For future reference, when the configuration pair $C_0$ satisfies the latter case, we say the configuration pair $C_0$ is *implicitly revealing*. Fig. 3 illustrates this situation.

However, without knowledge of the authentic underlying configuration $C_0$, we do not know whether $C_0$ is implicitly revealing or not. In view of the results of Sections V and VI, we would like to find a single scheme that is able to achieve bounded $\mathsf{E}_{C_0}\{T_{\inf}(t)\}$ when being applied to an implicitly revealing $C_0$, and on the other hand to achieve the $\log t$ lower bound when being applied to those $C_0$ which are not implicitly revealing.

The needed regularity conditions are the same as those in Sections V and VI.

1) $\mathbf{X}$ is a finite set and $\mathsf{P}_G(X_t = x) > 0$ for all $x \in \mathbf{X}$.
2) $\forall \theta_1, \theta_2, x, I(\theta_1, \theta_2\,|\,x)$ is strictly positive and finite.

To simplify the notation and the following proof, we define a partial ordering as $\theta \prec \vartheta$ iff $\forall x, \mu_\theta(x) \le \mu_\vartheta(x)$, and $\theta \succ \vartheta$ is defined similarly. Note that for a configuration $C_0 = (\theta_1, \theta_2)$, it can be the case that neither $\theta_1 \prec \theta_2$ nor $\theta_1 \succ \theta_2$.

Example:

- $\Theta = (0,\infty), \mathbf{X} = \{-1,1\}$ and the conditional reward distribution $F_\theta(\cdot\,|\,x) \sim \mathcal{N}(\theta^2 - \theta x, 1)$. Then, $C_0 = (\theta_1,\theta_2) = (0.1, 0.2)$ is implicitly revealing, but $C_0 = (0,10)$ is not.

### B. Lower Bound

*Theorem 7 ($\log t$ Lower Bound):* Under the previous assumptions, for any uniformly good rule $\{\phi_\tau\}$, if $C_0$ is not implicitly revealing, $T_{\inf}(t)$ satisfies

$$\lim_{t\to\infty} \mathsf{P}_{C_0}\left(T_{\inf}(t) \geq \frac{(1-\epsilon)\log t}{K_{C_0}}\right) = 1 \qquad \forall \epsilon > 0$$

and

$$\liminf_{t\to\infty} \frac{E_{C_0}\{T_{\inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}} \tag{4}$$

where $K_{C_0}$ is a constant depending on $C_0$. If $M_{C_0} = 2$, $T_{\inf}(t) = T_1(t)$, and the constant $K_{C_0}$ can be expressed as follows:

$$K_{C_0} = \inf_{\{\theta:\exists x_0, \text{s.t. } \mu_\theta(x_0) > \mu_{\theta_2}(x_0)\}} \sup_x \{I(\theta_1, \theta\,|\,x)\}.$$

The expression for $K_{C_0}$ for the case in which $M_{C_0} = 1$ can be obtained by symmetry.

The only difference between the lower bounds (2) and (4) is that, in (4), $K_{C_0}$ has been changed from taking the infimum over $\{\theta > \theta_2\} = \{\forall x, \mu_\theta(x) > \mu_{\theta_2}(x)\}$ to a larger set, $\{\theta : \exists x, \mu_\theta(x) > \mu_{\theta_2}(x)\}$. The reason for this is that under this case, consider a $\theta$ for which there exists $x$ such that $\mu_\theta(x) > \mu_{\theta_2}(x)$. If the authentic configuration is $C' = (\theta, \theta_2)$ rather than $(\theta_1, \theta_2)$, a linear order of incorrect sampling will be introduced, which violates the uniformly good-rule assumption. As a result, a broader class of competing distributions $C' = (\theta, \theta_2)$ must be considered, i.e., we must consider a different set of configurations, over which the infimum is taken.

A detailed proof is contained in Appendix VI.

### C. Scheme Achieving the Lower Bound

Consider the same two additional conditions as those in Section VI.

1) $\Theta$ is finite.
2) A saddle point for $K_{C_0}$ exists; that is, for all $\theta_1$

$$\inf_{\{\theta:\exists x_0, \mu_\theta(x_0) > \mu_{\theta_2}(x_0)\}} \sup_x I(\theta_1, \theta\,|\,x) = \sup_x \inf_{\{\theta:\exists x_0, \mu_\theta(x_0) > \mu_{\theta_2}(x_0)\}} I(\theta_1, \theta\,|\,x).$$

A proposed scheme is described in Algorithm 4

**Algorithm 4** $\phi_{t+1}$, the decision at time $t+1$

```
1: if there exists i ∈ {1,2} and x ∈ X such that T_i^x(t) = 0, then { Cond0}
2:    φ_{t+1} ← (t+1) mod 2.
3: else if the whole system is not well-sampled or θ^α = θ^β, then {Cond1}
4:    ctr(X_{t+1}) ← ctr(X_{t+1}) + 1 and φ_{t+1} ← ctr(X_{t+1}) mod 2.
5: else if there exists i ∈ {1,2}, such that ∀θ, x, μ_θ(x) ≤ μ_{i(Ĉ_t)}(x), then
   { ..... Cond2}
6:    φ_{t+1} ← i, where i is the satisfying index.
7: else if Ĉ_t is implicitly revealing, then {..... Cond2.5}
8:    φ_{t+1} ← M_{Ĉ_t}(X_{t+1}).
9: else {..... Cond3}
10:   ctr(Ĉ_t) ← ctr(Ĉ_t) + 1.
11:   if ctr(Ĉ_t) is odd, then {..... Cond3a}
12:      φ_{t+1} ← M_{Ĉ_t}(X_{t+1}).
13:   else {..... Cond3b}
14:      θ* ← arg min{Λ_t(Ĉ_t, θ) : ∀θ, ∃x_0, s.t. μ_θ(x_0) > μ_{θ^{α∨β}}(x_0)}.
15:      if X_{t+1} = x*(Ĉ_t) then {..... Cond3b1}
16:         if Λ_t(Ĉ_t, θ*) ≤ t(log t)^2, then {..... Cond3b1a}
17:            ctr(Ĉ_t, θ*) ← ctr(Ĉ_t, θ*) + 1.
18:            if ∃k ∈ ℕ s.t. ctr(Ĉ_t, θ*) = k^2, then {..... Cond3b1a1}
19:               φ_{t+1} ← k mod 2.
20:            else {..... Cond3b1a2}
21:               φ_{t+1} ← 3 − M_{Ĉ_t}(X_{t+1}).
22:            end if
23:         else {..... Cond3b1b}
24:            φ_{t+1} ← M_{Ĉ_t}(X_{t+1}).
25:         end if
26:      else {..... Cond3b2}
27:         φ_{t+1} ← M_{Ĉ_t}(X_{t+1}).
28:      end if
29:   end if
30: end if
```

which is similar to the scheme in Section VI-C. The only differences are the insertion of Cond2.5, Lines 7 and 8; the modification of Cond2, Lines 5 and 6; and the modification of Cond3b, Line 14.

*Notes:*

1) When the estimate $\hat{C}_t = (\theta^\alpha, \theta^\beta)$ is not implicitly revealing, an ordering between $\theta^\alpha$ and $\theta^\beta$ exists. As a result, all notation regarding $\alpha\vee\beta$, $\theta^{\alpha\vee\beta}$, etc., remains valid.
2) The definition of $\Lambda_t(\hat{C}_t, \theta)$ is slightly different. For any estimate $\hat{C}_t = (\theta^\alpha, \theta^\beta)$ that is not implicitly revealing, we can define the most informative bandit according to $\hat{C}_t$ as

$$x^*(\hat{C}_t) := \arg\max_x \inf_{\{\theta:\exists x_0, \mu_\theta(x_0) > \mu_{\theta^{\alpha\vee\beta}}(x_0)\}} I(\theta^{\alpha\wedge\beta}, \theta\,|\,x) \tag{5}$$

and $\Lambda_t(\hat{C}_t, \theta)$ to be the conditional likelihood ratio between the seemingly inferior arm $\theta^{\alpha\wedge\beta}$ and the competing parameter $\theta$. That is

$$\Lambda_t(\hat{C}_t, \theta) := \prod_{m=1}^{T^{x^*}_{\alpha\wedge\beta}(t)} \frac{F_{\theta^{\alpha\wedge\beta}}\left(dY^{\alpha\wedge\beta}_{\tau_{x^*}(m)} \,\middle|\, x^*(\hat{C}_t)\right)}{F_\theta\left(dY^{\alpha\wedge\beta}_{\tau_{x^*}(m)} \,\middle|\, x^*(\hat{C}_t)\right)}$$

where $\tau_{x^*}(m)$ denotes the time instant of the $m$th pull of arm $\alpha\wedge\beta$ when the side observation $X_\tau = x^*(\hat{C}_\tau)$. [The difference between this new $\Lambda_t(\hat{C}_t, \theta)$ and the previous one in Algorithm 3 is that we have a new $x^*(\hat{C}_t)$ defined in (5).]

TABLE II
SUMMARY OF THE BENEFIT OF THE SIDE OBSERVATIONS AND THE REQUIRED REGULARITY CONDITIONS.

| Characterization | Regularity Conditions | Results |
|---|---|---|
| $G_{C_1} \neq G_{C_2}$ iff $C_1 \neq C_2$. | As $\hat{C}_t \to C_0$, $\forall x$, $M_{\hat{C}_t}(x) = M_{C_0}(x)$. | $\exists\{\phi_\tau\}$ s.t. $\forall C_0$, $\lim_t \mathsf{E}_{C_0}\{T_{inf}(t)\} < \infty$. |
| (i) Constant $G_C$, i.e., $G_C := G$,<br>(ii) $\forall C, \exists x_1, x_2$, s.t. $M_C(x_1) = 1$, $M_C(x_2) = 2$ (implicitly revealing). | (i) $\mathbf{X}$ is finite.<br>(ii) $\forall \theta_1 \neq \theta_2, x$, $0 < I(\theta_1, \theta_2 \vert x) < \infty$.<br>(iii) $\forall x$, $\mu_\theta(x)$ is continuous w.r.t. $\theta$. | $\exists\{\phi_\tau\}$ such that $\forall C_0$, $\lim_t \mathsf{E}_{C_0}\{T_{inf}(t)\} < \infty$. |
| (i) Constant $G_C$, i.e., $G_C := G$,<br>(ii) $\forall C$, $M_C(x)$ depends only on $C$, not on $x$. | (i) $\mathbf{X}$ is finite.<br>(ii) $\forall \theta_1 \neq \theta_2, x$, $0 < I(\theta_1, \theta_2 \vert x) < \infty$,<br>(iii) $\forall x$, $\mu_\theta(x)$ is strictly increasing w.r.t. $\theta$. | For any uniformly good $\{\phi_\tau\}$, we have $\lim_t \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}}$,<br>$K_{C_0} := \inf_\theta \sup_x I(\theta_1, \theta \vert x)$.<br>For finite $\mathbf{\Theta}$, $\exists\{\phi_\tau\}$, s.t. $\lim_t \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{\log t} \leq \frac{1}{K_{C_0}}$. |
| (i) Constant $G_C$, i.e., $G_C := G$,<br>(ii) The underlying $C_0$ may be implicitly revealing or not. | (i) $\mathbf{X}$ is finite.<br>(ii) $\forall \theta_1 \neq \theta_2, x$, $0 < I(\theta_1, \theta_2 \vert x) < \infty$. | For any uniformly good $\{\phi_\tau\}$, if $C_0$ is not implicitly revealing, we have $\lim_t \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}}$,<br>$K_{C_0} := \inf_\theta \sup_x I(\theta_1, \theta \vert x)$.<br>For finite $\mathbf{\Theta}$, $\exists\{\phi_t\}$ s.t.<br>(1) if $C_0$ is implicitly revealing, $\mathsf{E}_{C_0}\{T_{inf}(t)\} < \infty$,<br>(2) if $C_0$ is not *i.r.*, $\lim_t \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{\log t} \leq \frac{1}{K_{C_0}}$. |

*Theorem 8 (Asymptotic Tightness):* With the aforementioned conditions, the scheme described in Algorithm 4 has bounded $\lim_t \mathsf{E}_{C_0}\{T_{\inf}(t)\}$, or achieves the $\log t$ lower bound (4), depending on whether the underlying configuration pair $C_0$ is implicitly revealing or not.

A detailed analysis is given in Appendix VI.

## VIII. CONCLUSION

We have shown that observing additional side information can significantly improve sequential decisions in bandit problems. If the side observation itself directly provides information about the underlying configuration, then it resolves the dilemma of forced sampling and optimal control. The expected inferior sampling time will be bounded, as has been shown in Section IV. If the side observation does not provide information on the underlying configuration $(\theta_1, \theta_2)$, but *always* affects the preference order (implicitly revealing), then the myopic approach of sampling the seemingly-best arm will automatically sample both arms enough. The expected inferior sampling time is bounded, as shown in Section V. If the side observation *does not* affect the preference order at all, the dilemma still exists. However, by postponing our forced sampling to the most informative time instants, we can reduce the constant in the $\log t$ lower bound, as shown in Section VI. In Section VII, we have combined the settings of Sections V and VI, and have obtained a general result. When the underlying configuration $C_0$ is implicitly revealing (such that $X_t$ will change the preference order), we have obtained bounded expected inferior sampling time as in Section V. Even if $C_0$ is not implicitly revealing (in that $X_t$ does not change the preference order), the new $\log t$ lower bound can be achieved as in Section VI. Our results are summarized in Table II.

## APPENDIX I
## SANOV'S THEOREM AND THE PROHOROV METRIC

For two distributions $P$ and $Q$ on the reals, the Prohorov metric is defined as follows.

*Definition 2 (The Prohorov Metric):* For any closed set $A \subset \mathbb{R}$ and $\epsilon > 0$, define $A^\epsilon$, the $\epsilon$-flattening of $A$, as

$$A^\epsilon := \{x \in \mathbb{R}\colon \inf_{y \in A} |x - y| < \epsilon\}.$$

The Prohorov metric $\rho$ is then defined as follows.

$$\rho(P, Q) := \inf\{\epsilon > 0 : P(A) \leq Q(A^\epsilon) + \epsilon \text{ for all closed } A \subset \mathbb{R}.\}.$$

The Prohorov metric generates the topology corresponding to convergence in distribution. Throughout this paper, the open/closed sets on the space of distributions are thus defined accordingly.

*Theorem 9 (Sanov's Theorem):* Let $L_X(n)$ denote the empirical measure of the real-valued i.i.d. random variables $X_1, X_2, \ldots, X_n$. Suppose $X_i$ is of distribution $P$ and consider any open set $A$ and closed set $B$ from the topological space of distributions, generated by the Prohorov metric. We have

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathsf{P}_P(L_X(n) \in A) \geq -\inf_{Q \in A} I(Q, P)$$
$$\limsup_{n\to\infty} \frac{1}{n} \log \mathsf{P}_P(L_X(n) \in B) \leq -\inf_{Q \in B} I(Q, P).$$

Further discussion of the Prohorov metric and Sanov's theorem can be found in [23] and [24].

## APPENDIX II
## PROOF OF THEOREM 3

*Proof:* For any underlying configuration pair $C_0 = (\theta_1, \theta_2)$, define the error set $\mathbf{C}_e$ as follows:

$$\mathbf{C}_e := \bigcup_{x \in \mathbf{X}} \{C \in \mathbf{\Theta}^2\colon M_C(x) \neq M_{C_0}(x)\}. \tag{6}$$

Let $\bar{\mathbf{C}}_e$ denote the closure of $\mathbf{C}_e$. By Condition 1, $C_0 \notin \bar{\mathbf{C}}_e$. For any $t$, we can write

$$\begin{aligned}\mathsf{P}_{C_0}\left(\phi_t \neq M_{C_0}(X_t)\right) &= \mathsf{P}_{C_0}\left(M_{\hat{C}_t}(X_t) \neq M_{C_0}(X_t)\right)\\ &\leq \mathsf{P}_{C_0}\left(\exists x, M_{\hat{C}_t}(x) \neq M_{C_0}(x)\right)\\ &= \mathsf{P}_{C_0}(\hat{C}_t \in \mathbf{C}_e)\\ &\leq \mathsf{P}_{C_0}(\hat{C}_t \in \bar{\mathbf{C}}_e).\end{aligned}$$

Let $\epsilon = (1/3)\inf\{\rho(G_{C_0}, G_{C_e})\colon C_e \in \bar{\mathbf{C}}_e\}$, which is strictly positive by Condition 1, and consider sufficiently large $t \geq (1/\epsilon)$. If $\rho(G_{C_0}, L_X(t)) < \epsilon$, then by the definition of $\mathbf{C}_t, \rho(G_{\hat{C}_t}, L_X(t)) < \epsilon + \epsilon = 2\epsilon$. By the triangle inequality, $\rho(G_{C_0}, G_{\hat{C}_t}) < 3\epsilon$ and $\hat{C}_t \notin \bar{\mathbf{C}}_e$. As a result

$$\{\hat{C}_t \in \bar{\mathbf{C}}_e\} \subset \{\rho(L_X(t), G_{C_0}) \geq \epsilon\} \triangleq \mathbf{K}_t.$$

$\mathbf{K}_t$ is a closed set. By Sanov's theorem, the probability of $\mathbf{K}_t$ is exponentially upper bounded w.r.t. $t$, and so is $\mathsf{P}_{C_0}(\hat{C}_t \in \bar{\mathbf{C}}_e)$. As a result, we have

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\inf}(t)\} = \lim_{t\to\infty} \sum_{\tau=1}^{t} \mathsf{P}_{C_0}\left(\phi_\tau \neq M_{C_0}(X_\tau)\right) < \infty.$$

By the monotone convergence theorem, the expectation of $\lim_{t\to\infty} T_{\inf}(t)$ is finite, which implies that $\lim_{t\to\infty} T_{\inf}(t)$ is finite a.s. ∎

## Appendix III
## Proof of Theorem 4

Similarly, we define $\mathbf{C}_e$ as that in (6). We need the following lemma to complete the analysis.

*Lemma 1:* With the regularity conditions specified in Section V, $\exists a_1, a_2 > 0$ such that $\mathsf{P}_{C_0}(\hat{C}_t \in \mathbf{C}_e) \leq a_1 \exp(-a_2 \min\{T_1^{x^\star}(t), T_2^{x^\star}(t)\})$.

*Proof of Lemma 1:* By the continuity of $\mu_\theta(x)$ w.r.t. $\theta$ and the assumption of finite $\mathbf{X}$, it can be shown that $C_0 \in \bar{\mathbf{C}}_e^c$.[5] Therefore, there exists a neighborhood of $C_0, \mathbf{C}_\delta = (\theta_1 - \delta, \theta_1 + \delta) \times (\theta_2 - \delta, \theta_2 + \delta)$, such that $\mathbf{C}_\delta \subset \bar{\mathbf{C}}_e^c \Leftrightarrow \bar{\mathbf{C}}_e \subset \mathbf{C}_\delta^c$.

Define a strictly positive $\epsilon > 0$ as follows:

$$\epsilon := \frac{1}{4}\inf\Big\{\rho\left(F_{\hat{\theta}_i}(\cdot\,|\,x), F_{\theta_i}(\cdot\,|\,x)\right)\colon \forall x \in \mathbf{X}, i \in \{1,2\}, (\hat{\theta}_1, \hat{\theta}_2) \in \mathbf{C}_\delta^c\Big\}.$$

We would like to prove that for sufficiently large $t > (1/\epsilon)$

$$\left\{\hat{C}_t \in \mathbf{C}_\delta^c\right\} \subset \left\{\exists i, \rho\left(L_i^{x^\star}(t), F_{\theta_i}(\cdot\,|\,x^\star)\right) > \epsilon\right\}.$$

Suppose $\rho(L_i^{x^\star}(t), F_{\theta_i}(\cdot\,|\,x^\star)) \leq \epsilon$ for both $i = 1, 2$. By the definition of $\sigma(C_0, t)$, we have

$$\sigma(C_0, t) \leq 2\epsilon. \tag{7}$$

[5] $\bar{\mathbf{C}}_e^c$ denotes the complement of $\bar{\mathbf{C}}_e$.

However, for those $\hat{C}_t \in \mathbf{C}_\delta^c$, by the definition of $\epsilon$, for some $i \in \{1, 2\}$, we have

$$\begin{aligned}\sigma(\hat{C}_t, t) &\geq \rho\left(F_{\hat{\theta}_i}(\cdot\,|\,x^\star), L_i^{x^\star}(t)\right)\\ &\geq \rho\left(F_{\hat{\theta}_i}(\cdot\,|\,x^\star), F_{\theta_i}(\cdot\,|\,x^\star)\right) - \rho\left(F_{\theta_i}(\cdot\,|\,x^\star), L_i^{x^\star}(t)\right)\\ &\geq 3\epsilon\end{aligned} \tag{8}$$

which contradicts the definition of $\mathbf{C}_t$ since (7) and (8) imply $\sigma(\hat{C}_t, t) > (1/t) + \sigma(C_0, t)$. As a result, for sufficiently large $t$, we have

$$\begin{aligned}\{\hat{C}_t \in \mathbf{C}_e\} &\subset \left\{\hat{C}_t \in \mathbf{C}_\delta^c\right\}\\ &\subset \left\{\exists i, \rho\left(L_i^{x^\star}(t), F_{\theta_i}(\cdot\,|\,x^\star)\right) > \epsilon\right\}\\ &= \bigcup_{i=1,2} \left\{\rho\left(L_i^{x^\star}(t), F_{\theta_i}(\cdot\,|\,x^\star)\right) > \epsilon\right\}.\end{aligned} \tag{9}$$

By Sanov's theorem, the probability of each term in the union of the right-hand side of (9) is exponentially bounded w.r.t. $T_i^{x^\star}(t)$. As a result, the probability of this finite union is bounded by $a_1 \exp(-a_2 \min\{T_1^{x^\star}(t), T_2^{x^\star}(t)\})$ for some $a_1, a_2 > 0$. ∎

*Analysis of the Scheme:* We first use induction to show that $\forall t \geq 6, T_i(t) \geq \sqrt{t}$. This statement is true for $t = 6$. Suppose $T_t(t-1) \geq \sqrt{t-1}$. If $T_i(t-1) \geq \sqrt{t}$, by the monotonicity of $T_i(t)$ w.r.t. $t$, we have $T_i(t) \geq T_i(t-1) \geq \sqrt{t}$. If $T_i(t-1) < \sqrt{t}$, by the forced sampling mechanism, $T_i(t) = T_i(t-1) + 1 \geq \sqrt{t-1} + 1 \geq \sqrt{t}$.

We consider the event of the inferior sampling at time $(t+1)$

$$\begin{aligned}\{\phi_{t+1} \neq M_{C_0}(X_{t+1})\} &= \left\{\phi_{t+1} \neq M_{C_0}(X_{t+1}), \hat{C}_t \in \mathbf{C}_e\right\}\\ &\quad \cup \left\{\phi_{t+1} \neq M_{C_0}(X_{t+1}), \hat{C}_t \in \mathbf{C}_e^c\right\}\\ &\subset \{\hat{C}_t \in \mathbf{C}_e\}\\ &\quad \cup \left\{\phi_{t+1} \neq M_{C_0}(X_{t+1}), \hat{C}_t \in \mathbf{C}_e^c\right\}\\ &\triangleq A_{t+1} \cup B_{t+1}.\end{aligned} \tag{10}$$

Since $T_i(t) \geq \sqrt{t}$, we have $\min_i T_i(t) \geq \sqrt{t}$ and $\min_i T_i^{x^\star}(t) \geq (\sqrt{t}/|\mathbf{X}|)$. By Lemma 1, we have $\mathsf{P}_{C_0}(A_{t+1}) \leq a_1 e^{-a_2(\sqrt{t}/|\mathbf{X}|)}$ and, hence, $\sum_{t+1=7}^{\infty} \mathsf{P}_{C_0}(A_{t+1}) < \infty$.

For $\mathsf{P}_{C_0}(B_{t+1})$, we can write

$$\begin{aligned}B_{t+1} &= \left\{\phi_{t+1} \neq M_{\hat{C}_t}(X_{t+1}), \hat{C}_t \in \mathbf{C}_e^c\right\}\\ &\subset \left\{\min\{T_i(t)\}_i < \sqrt{t+1}, \hat{C}_t \in \mathbf{C}_e^c\right\}\\ &= \left\{\min\{T_i(t)\}_i = \sqrt{t} \in \mathbb{N}, \hat{C}_t \in \mathbf{C}_e^c\right\}\\ &\subset \{\exists i, \phi_a \neq i, \forall a \in (\tau_0, t], T_i(\tau_0) = \sqrt{t}\}\\ &\triangleq B_{t+1}^1 \cup B_{t+1}^2\end{aligned} \tag{11}$$

where $\tau_0 = (\sqrt{t} - 1)^2 + 1$ and $B_{t+1}^1, B_{t+1}^2$ correspond to $i = 1, 2$, respectively. The first equality follows from the fact that since $\hat{C}_t \in \mathbf{C}_e^c, M_{C_0}(X_{t+1}) = M_{\hat{C}_t}(X_{t+1})$. The first subset sign follows from the fact that $\phi_{t+1} \neq M_{\hat{C}_t}(X_{t+1})$ implies the decision rule $\phi_{t+1}$ is in the stage of forced sampling.

The second equality follows by combining both the inequalities: $\min\{T_i(t)\}_i \geq \sqrt{t}$ and $\min\{T_i(t)\}_i < \sqrt{t+1}$ and the fact that both $t$ and $T_i(t)$ are integers.

The reasoning behind the second subset inequality is as follows. By again using the fact that $T_i(t) \geq \sqrt{t}$ and substituting $\tau_0$ for $t$, we have $\sqrt{t}-1 < T_i(\tau_0)$ and thus have $T_i(\tau_0) = \sqrt{t} = T_i(t)$, which guarantees that arm $i$ has not been sampled from time $\tau_0 + 1$ to $t$.

By the symmetry between $B^1_{t+1}$ and $B^2_{t+1}$, we can consider only $B^1_{t+1}$, for example. We have

$$\begin{aligned}
&\mathsf{P}_{C_0}\left(B^1_{t+1}\right) \\
&\quad \leq \mathsf{P}_{C_0}\left(M_{\hat{C}_{a-1}}(X_a) = 2, \forall a \in (\tau_0, t]\right) \\
&\quad = \prod_{a\in(\tau_0,t]} \mathsf{P}_{C_0}\left(M_{\hat{C}_{a-1}}(X_a) = 2 \,\middle|\, M_{\hat{C}_{b-1}}(X_b) = 2,\right. \\
&\qquad\qquad \left.\forall b \in (\tau_0, a)\right) \\
&\quad \leq \left(1 - \min_x\{\mathsf{P}_G(X_t = x)\}\right)^{t-\tau_0}. \qquad (12)
\end{aligned}$$

The first inequality follows from the definition of $\{\phi_\tau\}$ which implies that if $T_1(\tau_0) = T_1(t) \geq \sqrt{t}$, the forced sampling mechanism is not active during the time interval $(\tau_0, t]$. So $\phi_a = 2$ implies $M_{\hat{C}_{a-1}}(X_a) = 2, \forall a \in (\tau_0, t]$. The second inequality follows from the assumption of i.i.d. $\{X_\tau\}$, which implies that $X_a$ is independent of $\hat{C}_b$ and $X_b$ for all $b < a$. Since at least one $x$ will make $M_{\hat{C}_{a-1}}(X_a) = 1$, each term in the product is then upper bounded by $1 - \min_x\{\mathsf{P}_G(X_t = x)\}$. It is worth noting that by the regularity assumption on $G$, $1-\min_x\{\mathsf{P}_G(X_t = x)\}$ is strictly less than 1.

Then, from (11), (12), and the union bound, we obtain $\mathsf{P}_{C_0}(B_{t+1}) \leq \mathsf{P}_{C_0}(B^1_{t+1}) + \mathsf{P}_{C_0}(B^2_{t+1}) \leq 2a^{t-((\sqrt{t}-1)^2+1)}$ for some $a < 1$. Hence, $\sum_{t+1=7}^{\infty} \mathsf{P}_{C_0}(B_{t+1}) < \infty$. From (10), we conclude that

$$\begin{aligned}
&\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\inf}(t)\} \\
&\qquad \leq 6 + \sum_{\tau+1=7}^{\infty} \left(\mathsf{P}_{C_0}(A_{\tau+1}) + \mathsf{P}_{C_0}(B_{\tau+1})\right) < \infty
\end{aligned}$$

which completes the proof. ∎

## Appendix IV
## Proof of Theorem 5

*Proof:* The proof is inspired by [14]. Without loss of generality, we assume $M_{C_0} = 2$, which immediately implies $T_{\inf}(t) = T_1(t)$. Fix a $\theta$ with $\mu_\theta > \mu_{\theta_2}$, and define $C' = (\theta, \theta_2)$. Let $\lambda(n)$ denote the log likelihood ratio between $\theta_1$ and $\theta$ based on the first $n$ observed rewards of arm 1. That is

$$\lambda(n) := \sum_{m=1}^{n} \log\left(\frac{F_{\theta_1}\left(dY^1_{\tau(m)} \,\middle|\, X_{\tau(m)}\right)}{F_{\theta}\left(dY^1_{\tau(m)} \,\middle|\, X_{\tau(m)}\right)}\right)$$

where $\tau(m)$ is a random variable corresponding to the time index of the $m$th pull of arm 1.

By conditioning on the sequence $\{X_{\tau(m)}\}$, $\lambda(n)$ is a sum of independent r.v.'s. Let $K_{C'} := \sup_{x\in\mathbf{X}}\{I(\theta_1, \theta \,|\, x)\}$, and suppose there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \frac{\lambda(n)}{n} > K_{C'} + \delta$$

with positive probability. Then with positive probability, there exists an $x_0$ such that the average of the subsequence for which $X_{\tau(m)} = x_0$, will be larger than $K_{C'} + \delta$. This, however, contradicts the strong law of large numbers since the subsequence is i.i.d. and with marginal expectation $I(\theta_1, \theta \,|\, x_0)$. Thus, we obtain

$$\limsup_{n\to\infty} \frac{\lambda(n)}{n} \leq K_{C'} \quad \mathsf{P}_{C_0} - \text{a.s.} \qquad (13)$$

Inequality (13) is equivalent to the statement that with probability one, there are finitely many $n$ such that $\lambda(n) > n(K_{C'} + \delta)$ for some $\delta > 0$. Since $K_{C'} > 0$, this in turn implies there are at most finitely manly $n$ such that $\max_{m\leq n} \lambda(m) > n(K_{C'} + \delta)$. As a result, we have

$$\limsup_{n\to\infty} \frac{\max_{m\leq n} \lambda(m)}{n} \leq K_{C'} \quad P_{C_0} - \text{a.s.}$$

and

$$\lim_{n\to\infty} \mathsf{P}_{C_0}\left(\exists m \leq n, \lambda(m) \geq (1+\delta) n K_{C'}\right) = 0. \qquad (14)$$

Henceforth, we proceed using contradiction. Suppose

$$\limsup_{t\to\infty} \mathsf{P}_{C_0}\left(T_1(t) < \frac{\log t}{(1+2\delta)K_{C'}}\right) > 0.$$

Using $A_1$ and $A_2$ as shorthand to denote events $A_1 := \{T_1(t) < ((\log t)/((1+2\delta)K_{C'}))\}$ and $A_2 := \{\lambda(T_1(t)) \leq (((1+\delta)\log t)/(1+2\delta))\}$, and by (14), we have

$$\limsup_{t\to\infty} \mathsf{P}_{C_0}(A_1 \cap A_2) > 0. \qquad (15)$$

The quantity $\mathsf{E}_{C'}\{T_{\inf}(t)\}$ can be rewritten as follows:

$$\begin{aligned}
&\mathsf{E}_{C'}\{T_{\inf}(t)\} \\
&\quad \overset{(a)}{=} \mathsf{E}_{C'}\{T_2(t)\} \\
&\quad \overset{(b)}{=} \mathsf{E}_{C'}\{t - T_1(t)\} \\
&\quad \overset{(c)}{\geq} \left(t - \frac{\log t}{(1+2\delta)K_{C'}}\right) \mathsf{P}_{C'}(A_1) \\
&\quad \overset{(d)}{\geq} \left(t - \frac{\log t}{(1+2\delta)K_{C'}}\right) \mathsf{P}_{C'}(A_1 \cap A_2) \\
&\quad \overset{(e)}{\geq} \left(t - \frac{\log t}{(1+2\delta)K_{C'}}\right) e^{-\frac{(1+\delta)\log t}{1+2\delta}} \mathsf{P}_{C_0}(A_1 \cap A_2) \\
&\quad \overset{(f)}{=} \mathcal{O}\left(t^{\frac{\delta}{1+2\delta}}\right). \qquad (16)
\end{aligned}$$

The equality marked $(a)$ follows from $M_{C'} = 1$ and $(b)$ follows from the fact that $T_1(t) + T_2(t) = t$. $(c)$ and $(d)$ follow from elementary probability inequalities. $(e)$ follows from the change-of-measure formula and the definition of $A_2$ in which $\lambda(T_1(t)) \leq (((1+\delta)\log t)/(1+2\delta))$. $(f)$ follows from simple arithmetic and (15).

Inequality (16) contradicts the assumption that $\{\phi_\tau\}$ is uniformly good for both $C_0 = (\theta_1, \theta_2)$ and $C' = (\theta, \theta_2)$ and, thus, we have

$$\lim_{t\to\infty} \mathsf{P}_{C_0}\left(T_1(t) \geq \frac{(1-\epsilon)\log t}{K_{C'}}\right) = 1 \qquad \forall \epsilon > 0.$$

By choosing the $\theta$ in $C' = (\theta, \theta_2)$ with the minimizing configuration $\inf_{\theta>\theta_2} \sup_x I(\theta_1, \theta \,|\, x)$, we complete the proof of the first statement of Theorem 5. The second statement in Theorem 5 can be obtained by simply applying Markov's inequality and the first statement. ■

## Appendix V
## Proof of Theorem 6

We prove Theorem 6 by decomposing the inferior sampling time instants into disjoint subsequences, each of which will be discussed in separate lemmas respectively. For simplicity, throughout this proof, we use $1\{\mathsf{Cond1}(t)\}$ as shorthand for $1\{\mathsf{Cond1}$ is satisfied at time $t\}$,[6] and use $\delta-\mathrm{nbd}(G)$ to denote the $\delta$-neighborhood of the distribution $G(x)$ on the $L^\infty$ space of distributions.

Suppose $M_{C_0} = 2$. To prove that for the $\{\phi_\tau\}$ in Algorithm 3, $\limsup_{t\to\infty}((\mathsf{E}_{C_0} T_1(t))/(\log t)) \leq (1/(\inf_{\theta>\theta_2} \max_x I(\theta_1, \theta \,|\, x)))$, we first note the following:

$$\begin{aligned}
T_1(t) &= \sum_{\tau=1}^{t} 1\{\phi_\tau = 1\} \\
&= \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \mathsf{Cond0}(\tau)\} \\
&\quad + \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \mathsf{Cond1}(\tau)\} \\
&\quad + \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \mathsf{Cond2}(\tau)\} \\
&\quad + \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta) \\
&\qquad \theta^\alpha < \theta^\beta \neq \theta_2, \mathsf{Cond3}(\tau)\} \\
&\quad + \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta) \\
&\qquad \theta_1 \neq \theta^\alpha > \theta^\beta, \mathsf{Cond3}(\tau)\} \\
&\quad + \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta) \\
&\qquad \theta_1 \neq \theta^\alpha < \theta^\beta = \theta_2, \mathsf{Cond3}(\tau)\} \\
&\quad + \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta) \\
&\qquad \theta_1 = \theta^\alpha > \theta^\beta, \mathsf{Cond3}(\tau)\} \\
&\quad + \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta) = C_0 = (\theta_1, \theta_2) \\
&\qquad \mathsf{Cond3}(\tau)\}. \qquad (17)
\end{aligned}$$

[6]"At time $t$" means after observing $X_t$ but before the final decision $\phi_t$ is made. It is basically the moment when we are performing the $\phi_t$-deciding algorithm.

These eight terms of the right-hand side of (17) will be treated separately in Lemmas 2–8.

*Lemma 2:* Suppose $M_{C_0} = 2$, i.e., $\theta_1 < \theta_2$.[7] Then

$$\forall C_0 \in \Theta^2 \quad \lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \mathsf{Cond0}(\tau)\}\right\} \leq \lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\mathsf{Cond0}(\tau)\}\right\} < \infty.$$

*Proof:* Let $T0 := \sum_{\tau=1}^{\infty} 1\{\mathsf{Cond0}(\tau)\}$. By the monotone convergence theorem, it is equivalent to prove that $\mathsf{E}_{C_0}\{T0\} < \infty$ for all $C_0$. By the definition of $\mathsf{Cond0}$, we have

$$\begin{aligned}
&\mathsf{P}_{C_0}(T0 = t) \\
&\quad \leq \sum_{x\in\mathbf{X}} \mathsf{P}_G(X_t = x, \forall \tau < t \text{ and } \tau \equiv t \bmod 2, X_\tau \neq x) \\
&\quad = \sum_{x\in\mathbf{X}} \mathsf{P}_G(X_t = x)(1 - \mathsf{P}_G(X = x))^{[\frac{t-1}{2}]} \\
&\quad \leq (1 - \min_x \mathsf{P}_G(X_t = x))^{[\frac{t-1}{2}]}.
\end{aligned}$$

By directly computing the expectation, we obtain $\mathsf{E}_{C_0}\{T0\} < \infty$. ■

*Lemma 3:* Suppose $M_{C_0} = 2$, i.e., $\theta_1 < \theta_2$. Then

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \mathsf{Cond1}(\tau)\}\right\} \leq \lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\mathsf{Cond1}(\tau)\}\right\} < \infty.$$

*Proof:* We define $L_X(t \,|\, \mathsf{Cond1})$ as the empirical distribution of $X_\tau$ at those time instants $\tau \leq t$ for which $\mathsf{Cond1}$ is satisfied. We then have

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\mathsf{Cond1}(\tau)\} \\
&\quad = \sum_{\tau=1}^{t} 1\{\mathsf{Cond1}(\tau), L_X(\tau \,|\, \mathsf{Cond1}) \in \delta-\mathrm{nbd}(G)\} \\
&\qquad + \sum_{\tau=1}^{t} 1\{\mathsf{Cond1}(\tau), L_X(\tau \,|\, \mathsf{Cond1}) \notin \delta-\mathrm{nbd}(G)\}. \qquad (18)
\end{aligned}$$

By Sanov's theorem on finite alphabets (see [24]), each term in the second sum is exponentially upper bounded w.r.t. $\tau$, which implies the bounded expectation of the second sum. For the first sum, we have

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\mathsf{Cond1}(\tau), L_X(\tau \,|\, \mathsf{Cond1}) \in \delta-\mathrm{nbd}(G)\} \\
&\quad \leq \sum_{\tau=1}^{\infty} 1\{\exists i, x, \text{ s.t. } L_i^x(\tau-1) \notin \epsilon-\mathrm{nbd}\,(F_{\theta_i}(\cdot \,|\, x)), \text{and} \\
&\qquad L_X(\tau \,|\, \mathsf{Cond1}) \in \delta-\mathrm{nbd}(G)\} \\
&\quad \leq \sum_{x} \sum_{i=1}^{2} \sum_{\tau=1}^{\infty} 1\{L_i^x(\tau-1) \notin \epsilon-\mathrm{nbd}(F_{\theta_i}(\cdot \,|\, x)), \\
&\qquad L_X(\tau \,|\, \mathsf{Cond1}) \in \delta-\mathrm{nbd}(G)\} \qquad (19)
\end{aligned}$$

[7]There is no need to consider the case $\theta_1 = \theta_2$, since in that case, all allocation rules are optimal.

$$\leq \sum_x \sum_{i=1}^{2} \sum_{\tau'=1}^{\infty} 1\left\{\exists n \geq \left[\frac{\tau' \mathsf{P}_G(X=x)(1-\delta)-1}{2}\right], \right.$$
$$\left. \text{s.t. } \rho\left(L_i^x(n), F_{\theta_i}(\cdot \,|\, x)\right) > \epsilon \right\}. \quad (20)$$

The first inequality follows from extending the finite sum to the infinite sum and the definition of Cond1. The second inequality follows from the union bound. The third inequality follows from the following three steps. First, we change the summation index from the time variable $\tau$ to $\tau'$, which specifies that it is the $\tau'$th time that the condition in (19) is satisfied. (Note: By definition, $\tau \geq \tau'$.) Second, by $L_X(\tau \,|\, \text{Cond1}) \in \delta\text{−nbd}(G)$, there must be at least $\tau' \mathsf{P}_G(X=x)(1-\delta)$ time instants that $X_s = x, s \leq \tau$, which guarantees we have enough access to the bandit machine $x$. Finally, by the definition of Cond1 in Algorithm 3, at the $\tau'$th time of satisfaction, the sample size $n$ must be greater than $[((\tau' \mathsf{P}_{C_0}(X=x)(1-\delta)-1)/2)]$. By slightly abusing the notation $L_i^x(t)$ with $L_i^x(n)$, where $n$ represents the sample size $T_i^x(t)$ rather than the current time $t$, we obtain the third inequality.

*Remark:* This change-of-index transformation will be used extensively throughout the proofs in this section.

By Sanov's theorem on $\mathbb{R}$ (Theorem 9), the probability of each term in (20) is exponentially upper bounded w.r.t. $\tau'$, which implies that the summation has bounded expectation. By (18), the proof of Lemma 3 is then complete. ∎

*Lemma 4:* Suppose $M_{C_0} = 2$, i.e., $\theta_1 < \theta_2$. Then

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \text{Cond2}(\tau)\}\right\} < \infty.$$

*Proof:* By the assumption $\theta_1 < \theta_2$, we have

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \text{Cond2}(\tau)\} \\
&\quad = \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \theta^\alpha = \bar{\theta}\} \\
&\quad = \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \theta^\alpha = \bar{\theta}, \\
&\qquad\qquad L_X(\tau \,|\, \theta^\alpha = \bar{\theta}) \in \delta\text{−nbd}(G)\} \\
&\qquad + \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \theta^\alpha = \bar{\theta}, \\
&\qquad\qquad L_X(\tau \,|\, \theta^\alpha = \bar{\theta}) \notin \delta\text{−nbd}(G)\}.
\end{aligned}$$

By Sanov's theorem on finite alphabets, each term in the second sum is exponentially upper bounded w.r.t. $\tau$, which implies the bounded expectation of the second sum. For the first sum, we have

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \theta^\alpha = \bar{\theta} \\
&\quad L_X(\tau \,|\, \theta^\alpha = \bar{\theta}) \in \delta\text{−nbd}(G)\} \\
&\quad \leq \sum_{\tau=1}^{\infty} 1\left\{\hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \theta^\alpha = \bar{\theta}, \exists x, \right. \\
&\qquad \text{s.t. } \rho\left(L_1^x(\tau-1), F_{\theta_1}(\cdot \,|\, x)\right) > \epsilon, \\
&\qquad\quad \left. L_X(\tau \,|\, \theta^\alpha = \bar{\theta}) \in \delta\text{−nbd}(G)\right\} \\
&\quad \leq \sum_x \sum_{\tau'=1}^{\infty} 1\left\{\exists n \geq [\tau' \mathsf{P}_G(X=x)(1-\delta)-1], \right. \\
&\qquad \left. \text{s.t. } \rho\left(L_1^x(n), F_{\theta_1}(\cdot \,|\, x)\right) > \epsilon\right\}.
\end{aligned}$$

By extending the finite sum to the infinite sum, we obtain the first inequality. By the definition of Cond2 in Algorithm 3 and using exactly the same reasoning used in going from (19) to (20), we obtain the second inequality. By Sanov's theorem, each term in the above sum is exponentially upper bounded w.r.t. $\tau'$. Thus it follows that the expectation of the first sum is also finite, which completes the proof. ∎

*Lemma 5:* Suppose $M_{C_0} = 2$, i.e., $\theta_1 < \theta_2$. Then

$$\begin{aligned}
&\lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \right. \\
&\qquad\qquad \left. \theta^\alpha < \theta^\beta \neq \theta_2, \text{Cond3}(\tau)\}\right\} \\
&\leq \lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \right. \\
&\qquad\qquad \left. \theta^\alpha < \theta^\beta \neq \theta_2, \text{Cond3}(\tau)\}\right\} \\
&< \infty.
\end{aligned}$$

*Proof:* We have

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \theta^\alpha < \theta^\beta \neq \theta_2, \text{Cond3}(\tau)\} \\
&\quad = \sum_{(\theta,\vartheta):\theta<\vartheta\neq\theta_2} \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta, \vartheta), \text{Cond3}(\tau)\} \\
&\quad \leq 2 \sum_{(\theta,\vartheta):\theta<\vartheta\neq\theta_2} \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta, \vartheta), \text{Cond3a}(\tau)\} \\
&\quad = 2 \sum_{(\theta,\vartheta):\theta<\vartheta\neq\theta_2} \sum_x \sum_{\tau=1}^{t} 1\{X_\tau = x, \\
&\qquad\qquad \hat{C}_{\tau-1} = (\theta, \vartheta), \text{Cond3a}(\tau)\} \\
&\quad \leq 2 \sum_{(\theta,\vartheta):\theta<\vartheta\neq\theta_2} \sum_x \sum_{\tau=1}^{\infty} 1\left\{\rho\left(L_2^x(\tau-1), \right.\right. \\
&\qquad\qquad \left.\left. F_{\theta_2}(\cdot \,|\, x)\right) > \epsilon, \text{Cond3a}(\tau)\right\} \\
&\quad \leq 2 \sum_{(\theta,\vartheta):\theta<\vartheta\neq\theta_2} \sum_x \sum_{\tau'=1}^{\infty} 1\left\{\exists n \geq [\tau'-1], \right. \\
&\qquad\qquad \left. \text{s.t. } \rho\left(L_2^x(n), F_{\theta_2}(\cdot \,|\, x)\right) > \epsilon\right\}.
\end{aligned}$$

The first equality follows from conditioning on the event that the exact value of the estimate $\hat{C}_{\tau-1}$ is some configuration pair $(\theta, \vartheta)$. The first inequality follows from the definition of Cond3a in Algorithm 3, where double the number of time instants with odd $\text{ctr}(\hat{C})$ will be larger than the total number of times that Cond3 is satisfied. The second equality follows from conditioning on the value of $X_\tau$. The second inequality follows from the condition that the second coordinate of the estimate, $\vartheta \neq \theta_2$, and then extending the finite sum to the infinite sum. The third inequality follows from the definition of Cond3a and changing the time index to $\tau'$, similar to the reasoning in (19)–(20). By Sanov's theorem, each term is exponentially

upper bounded w.r.t. $\tau'$, and thus the entire sum has bounded expectation. The proof is thus complete. ■

*Corollary 1:* By the symmetry of $\{\phi_\tau\}$, we have

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \theta_1 \neq \theta^\alpha > \theta^\beta, \mathsf{Cond3}(\tau)\}\right\} < \infty.$$

*Lemma 6:* Suppose $M_{C_0} = 2$, i.e., $\theta_1 < \theta_2$. Then

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta) \theta_1 \neq \theta^\alpha < \theta^\beta = \theta_2, \mathsf{Cond3}(\tau)\}\right\} < \infty.$$

*Proof:* We have (21) and (22), as shown at the bottom of the page.

The second equality follows from the fact that the scheme samples the inferior arm only when either Cond3b1a1 or Cond3b1a2 is satisfied. For the first inequality, we condition on $\theta^*$ and extend to the infinite sum. For the last inequality, we change the time index to $\tau'$, which specifies the $\tau'$th satisfaction of Cond3b1a1, so that we can upper bound the first sum of (21). The reason we have a multiplication factor $(4(\tau'+1)^2 - (2\tau')^2)$ in front of the indicator function is in order to upper bound the second sum of (21), concerning Cond3b1a2, simultaneously.

To obtain this result, we note that between the consecutive times $\tau'$ and $\tau'+1$, at which Cond3b1a1 is satisfied and arm 1 is pulled, the number of times that Cond3b1a2 is satisfied and arm 1 is pulled cannot exceed $(2(\tau'+1))^2 - (2\tau')^2 - 1$, which is because of the algorithm involving $\mathrm{ctr}(\hat{C}_t, \theta^*)$ in Line 16. Multiplying the factor $(4(\tau'+1)^2 - (2\tau')^2)$, we simultaneously bound these two sums.

By Sanov's theorem, the expectation of the indicator in (22) is exponentially upper bounded w.r.t. $\tau'$. As a result, the entire sum will have bounded expectation, which in turn completes the proof. ■

*Lemma 7:* Suppose $M_{C_0} = 2$, i.e., $\theta_1 < \theta_2$. Then

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \theta_1 = \theta^\alpha > \theta^\beta, \mathsf{Cond3}(\tau)\}\right\} < \infty.$$

*Proof:* We have

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta), \theta_1 = \theta^\alpha > \theta^\beta, \mathsf{Cond3}(\tau)\} \\
&= \sum_{\vartheta:\theta_1>\vartheta} \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta_1, \vartheta), \mathsf{Cond3}(\tau)\} \\
&\le \sum_{\vartheta:\theta_1>\vartheta} \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \mathsf{Cond3}(\tau)\} \\
&\le \sum_{\vartheta:\theta_1>\vartheta} \left(2\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \mathsf{Cond3b}(\tau)\} + 1\right) \\
&= \#\{\vartheta \in \mathbf{\Theta} : \vartheta < \theta_1\} \\
&\quad + \sum_{\vartheta:\theta_1>\vartheta} \left(2 \sum_{\theta':\theta'>\theta_1} \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b}(\tau)\}\right) \\
&= \#\{\vartheta \in \mathbf{\Theta} : \vartheta < \theta_1\} \\
&\quad + 2 \sum_{\vartheta:\theta_1>\vartheta} \sum_{\theta':\theta'>\theta_1} \sum_{\tau=1}^{t} (1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b}(\tau), L_X(\tau \,|\, \mathsf{Cond3b}) \in \delta\text{−nbd}(G)\} \\
&\quad + 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b}(\tau), L_X(\tau \,|\, \mathsf{Cond3b}) \notin \delta\text{−nbd}(G)\}). \qquad (23)
\end{aligned}$$

The first inequality follows from Line 11 in Algorithm 3, where Cond3b is satisfied once after two times of Cond3 satisfaction. The last two equalities follow from conditioning on

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta),\ \theta_1 \neq \theta^\alpha < \theta^\beta = \theta_2, \mathsf{Cond3}(\tau)\} \\
&\quad = \sum_{\theta:\theta_1\neq\theta<\theta_2} \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta, \theta_2), \mathsf{Cond3}(\tau)\} \\
&\quad = \sum_{\theta:\theta_1\neq\theta<\theta_2} \sum_{\tau=1}^{t} (1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta, \theta_2), \mathsf{Cond3b1a1}(\tau)\} + 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta, \theta_2), \mathsf{Cond3b1a2}(\tau)\}) \\
&\quad \le \sum_{\theta_1\neq\theta<\vartheta=\theta_2} \sum_{\theta':\theta'>\theta_2} \left(\sum_{\tau=1}^{\infty} 1\{\phi_\tau = 1, C_{\tau-1} = (\theta, \theta_2), \theta^* = \theta', \mathsf{Cond3b1a1}(\tau)\}\right. \\
&\qquad \left. + \sum_{\tau=1}^{\infty} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta, \theta_2), \theta^* = \theta', \mathsf{Cond3b1a2}(\tau)\}\right) \qquad (21) \\
&\quad \le \sum_{\theta:\theta_1\neq\theta<\theta_2} \sum_{\theta':\theta'>\theta_2} \sum_{\tau'=1}^{\infty} (4(\tau'+1)^2 - (2\tau')^2) \cdot 1\left\{\exists n \ge [\tau'-1], \text{s.t. } \rho\left(L_1^{x^*(\theta')}(n), F_{\theta_1}(\cdot \,|\, x^*(\theta'))\right) > \epsilon\right\} \qquad (22)
\end{aligned}$$

$\theta^*$ and $L_X(\tau \,|\, \mathsf{Cond3b})$. By Sanov's theorem on finite alphabets, the terms of the second sum in (23) are exponentially upper bounded and the entire sum thus has bounded expectation. For the first sum, we have

$$\begin{aligned}&\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \\ &\qquad \mathsf{Cond3b}\ (\tau), L_X(\tau \,|\, \mathsf{Cond3b}) \in \delta\text{−nbd}(G)\} \\ &\le \frac{1}{\mathsf{P}_G(X = x^*(\theta_1, \vartheta))(1-\delta)} \\ &\quad \cdot \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1}(\tau)\}. \quad (24)\end{aligned}$$

This inequality follows from the fact that once $L_X$ falls into the $\delta-\mathrm{nbd}(G)$, the total number of time instants can be upper bounded by the number of instants when $X_\tau = x^*(\theta_1, \vartheta)$, over $\mathsf{P}_G(X = x^*(\theta_1, \vartheta))(1-\delta)$. To show

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1}\ (\tau)\}\right\} < \infty$$

we further decompose the expectand into

$$\begin{aligned}&\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1}(\tau)\} \\ &\quad = \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1a}(\tau)\} \\ &\qquad + \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1b}(\tau)\}. \quad (25)\end{aligned}$$

For the first sum in (25), under the assumption $\theta_1 > \vartheta$, we can write

$$\begin{aligned}&\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1a}(\tau)\} \\ &\quad \le \sum_{\tau=1}^{\infty} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1a1}(\tau)\} \\ &\qquad + \sum_{\tau=1}^{\infty} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1a2}(\tau)\} \\ &\quad \le 1 + 2\sum_{\tau=1}^{\infty} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1a2}(\tau)\} \\ &\quad \le 1 + 2\sum_{\tau'=1}^{\infty} 1\left\{\rho(\exists n \ge [\tau - 1], \right. \\ &\qquad \left. \text{s.t. } \rho\left(L_2^{x^*(\theta_1,\vartheta)}(n), F_{\theta_2}(\cdot \,|\, x^*(\theta_1, \vartheta))\right) > \epsilon\right\}. \quad (26)\end{aligned}$$

The first inequality follows from conditioning on the sub-conditions $\mathsf{Cond3b1a1}$ and $\mathsf{Cond3b1a2}$, and extending to the infinite sums. Let $\mathsf{SQ}_n$ denote the set of perfectly squared integers in $\{1, \ldots, n\}$. The second inequality is from the definition of $\mathsf{Cond3b1a1}$ in Algorithm 3 and the fact that $\forall n \in \mathbb{N}$, $|\mathsf{SQ}_n|$ is no larger than $1 + |\{1, \ldots, n\} \backslash \mathsf{SQ}_n|$. The third inequality follows from the fact that by definition, under $\mathsf{Cond3b1a2}$, $\phi_\tau = 2$, and changing the time index to $\tau'$, the number of satisfaction times. By Sanov's theorem on $\mathbb{R}$, the above has bounded expectation.

For the second sum of (25), with the condition $\theta_1 > \vartheta$

$$\begin{aligned}&\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1b}(\tau)\} \\ &\quad \le 1 + \sum_{\tau=1}^{t-1} 1\{\hat{C}_{\tau} = (\theta_1, \vartheta), \theta^* = \theta', \\ &\qquad \Lambda_\tau(\hat{C}_\tau, \theta') > \tau(\log\tau)^2\} \\ &\quad \le 1 + \sum_{\tau=1}^{t-1} 1\{\hat{C}_{\tau} = (\theta_1, \vartheta), \Lambda_\tau(\hat{C}_\tau, \theta_2) > \tau(\log\tau)^2\} \\ &\quad \le 1 + \sum_{\tau=1}^{t-1} 1\left\{\prod_{m=1}^{T_2^{x^*(\theta_1,\vartheta)}(\tau)} \frac{F_\vartheta\left(dY^2_{\tau_{x^*}(m)} \,\middle|\, x^*(\theta_1, \vartheta)\right)}{F_{\theta_2}\left(dY^2_{\tau_{x^*}(m)} \,\middle|\, x^*(\theta_1, \vartheta)\right)}\right. \\ &\qquad \left. > \tau(\log\tau)^2, \right\} \\ &\quad \le 1 + \sum_{\tau=1}^{t-1} 1\left\{\exists n \le \tau, \text{s.t.} \prod_{m=1}^{n} \frac{F_\vartheta\left(dY^2_{m\,|\,x^*} \,\middle|\, x^*(\theta_1, \vartheta)\right)}{F_{\theta_2}\left(dY^2_{m\,|\,x^*} \,\middle|\, x^*(\theta_1, \vartheta)\right)}\right. \\ &\qquad \left. > \tau(\log\tau)^2 \right\}\end{aligned}$$

where $Y^2_{m\,|\,x^*}$ is the reward of arm 2 at the $m$-th time that $X_s = x^*(\theta_1, \vartheta)$ and $\phi_s = 2$. The first inequality follows from focusing only on the $\Lambda_{\tau-1}(\hat{C}_{\tau-1}, \theta')$ condition in $\mathsf{Cond3b1b}$ and then shifting the time index $\tau$. The second inequality follows by replacing the minimum achieving $\theta'$ with $\theta_2$. The third inequality follows from expressing $\Lambda_\tau$ using its definition. The fourth inequality follows from the set relationship, where $n$ is $T_2^{x^*(\theta_1,\vartheta)}(\tau)$, the number of time instants that the side information $X_s = x^*(\theta_1, \vartheta)$ and $\phi_s = 2$, for $s \le \tau$.

We first note that

$$\prod_{m=1}^{n} ((F_\vartheta(dY^2_{m\,|\,x^*} \,|\, x^*(\theta_1, \vartheta)))/(F_{\theta_2}(dY^2_{m\,|\,x^*} \,|\, x^*(\theta_1, \vartheta)))$$

is a positive martingale with expectation 1, when being considered under distribution $F_{\theta_2}(\cdot \,|\, x^*(\theta_1, \vartheta))$. By Doob's maximal inequality, we have

$$\mathsf{P}_{C_0}\left(\exists n \le \tau, \prod_{m=1}^{n} \frac{F_\vartheta\left(dY^2_{m\,|\,x^*} \,\middle|\, x^*(\theta_1, \vartheta)\right)}{F_{\theta_2}\left(dY^2_{m\,|\,x^*} \,\middle|\, x^*(\theta_1, \vartheta)\right)} > \tau(\log\tau)^2\right) \le \frac{1}{\tau(\log\tau)^2}$$

and, thus, the expectation is bounded, i.e.,

$$\mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta_1, \vartheta), \theta^* = \theta', \mathsf{Cond3b1b}(\tau)\}\right\} \le 1 + \sum_{\tau=1}^{\infty} \frac{1}{\tau(\log\tau)^2} < \infty. \quad (27)$$

By (23)–(27), Lemma 7 is proved. ■

*Lemma 8:* Suppose $M_{C_0}=2$, i.e., $\theta_1<\theta_2$. Then

$$\limsup_{t\to\infty}\frac{1}{\log t}\cdot \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t}1\{\phi_\tau=1,\hat{C}_{\tau-1}=(\theta^\alpha,\theta^\beta) = (\theta_1,\theta_2),\textsf{Cond3}(\tau)\}\right\} \le \frac{1}{\inf_{\theta>\theta_2}\max_x I(\theta_1,\theta\,|\,x)}.$$

*Proof:* By the definition of $\{\phi_\tau\}$, especially of Cond3b1a, we have

$$\begin{aligned}
&\mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t}1\{\phi_\tau=1,\hat{C}_{\tau-1}=(\theta^\alpha,\theta^\beta) = C_0=(\theta_1,\theta_2),\textsf{Cond3}(\tau)\}\right\}\\
&\le \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t}1\{\hat{C}_{\tau-1}=C_0,\textsf{Cond3b1a}(\tau)\}\right\}\\
&\le \mathsf{E}_{C_0}\left\{\sup\left\{1\le n\le t-1:\ \min_{\theta>\theta_2}\prod_{m=1}^{n}\frac{F_{\theta_1}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}{F_{\theta}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}\le t(\log t)^2\right\}\right\}\\
&\le \mathsf{E}_{C_0}\left\{\sup\left\{1\le n<\infty:\ \min_{\theta>\theta_2}\prod_{m=1}^{n}\frac{F_{\theta_1}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}{F_{\theta}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}\le t(\log t)^2\right\}\right\}\\
&= \mathsf{E}_{C_0}\left\{\max_{\theta>\theta_2}\sup\left\{1\le n<\infty:\ \prod_{m=1}^{n}\frac{F_{\theta_1}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}{F_{\theta}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}\le t(\log t)^2\right\}\right\}\\
&= \mathsf{E}_{C_0}\left\{\max_{\theta>\theta_2}\sum_{s=1}^{\infty}1\left\{\inf_{n\ge s}\prod_{m=1}^{n}\frac{F_{\theta_1}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}{F_{\theta}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}\le t(\log t)^2\right\}\right\}
\end{aligned}$$

where $Y^1_{m\,|\,x^*}$ denotes the reward of the $m$th time that arm 1 of the sub-bandit machine $X_\tau=x^*(C_0)$ is pulled. The first inequality follows because, by definition, only when Cond3b1a is satisfied can $\phi_\tau=1$, given $\hat{C}_{\tau-1}=C_0$. The second inequality is obtained by focusing on the sub-condition $\Lambda_\tau(\hat{C}_\tau,\theta)$ in Cond3b1a, and letting $n=T_1^{x^*}(t-1)$ be the number of time instants when arm 1 is pulled and $X_\tau=x^*(C_0)$. The third inequality follows from extending the upper bound of $n$ from $t-1$ to $\infty$. The equalities follow from rearranging the max and min operators and elementary implications. By applying [12, Lemma 4.3], quoted as Lemma 9 later, we have

$$\begin{aligned}
&\limsup_{t\to\infty}\frac{1}{\log t+2\log\log t}\\
&\quad\cdot\mathsf{E}_{C_0}\left\{\max_{\theta>\theta_2}\sum_{s=1}^{\infty}1\left\{\inf_{n\ge s}\prod_{m=1}^{n}\frac{F_{\theta_1}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}{F_{\theta}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}\le t(\log t)^2\right\}\right\}\\
&\le \frac{1}{\min_{\theta>\theta_2}\mathsf{E}_{C_0}\left\{\log\left(\frac{F_{\theta_1}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}{F_{\theta}\left(dY^1_{m\,|\,x^*}\,\middle|\,x^*(C_0)\right)}\right)\right\}}\\
&= \frac{1}{\min_{\theta>\theta_2}I(\theta_1,\theta\,|\,x^*(C_0))} = \frac{1}{\max_x\inf_{\theta>\theta_2}I(\theta_1,\theta\,|\,x)}\\
&= \frac{1}{\inf_{\theta>\theta_2}\max_x I(\theta_1,\theta\,|\,x)}
\end{aligned}$$

where the equalities come from the existence-of-saddle-points assumption. By noting that $\log t\gg 2\log\log t$, this completes the proof of Lemma 8. ■

By (17) and Lemmas 2–8, it has been proved that for the $\{\phi_\tau\}$ described in Algorithm 3

$$\limsup_{t\to\infty}\frac{\mathsf{E}_{C_0}\{T_{\inf}(t)\}}{\log t}\le\frac{1}{K_{C_0}}\qquad\forall C_0\in\mathbf{\Theta}^2.$$

Lemma 4.3 of [12] is quoted as follows.

*Lemma 9 ([12, Lemma 4.3]):* Suppose $Y_1,Y_2,\ldots$ are i.i.d. r.v.'s taking values in a finite set $\mathbf{Y}$, with marginal mass function $p(y)$. Let $f^\theta\colon\mathbf{Y}\to\mathbb{R}$ be such that $0<\mathsf{E}_p\{f^\theta(Y_1)\}<\infty,\forall\theta\in\mathbf{\Theta}$, where $\mathbf{\Theta}$ is a finite set. Define $S_t^\theta=\sum_{\tau=1}^{t}f^\theta(Y_\tau)$, $L_A^\theta=\sum_{\tau=1}^{\infty}1\{\inf_{t\ge\tau}S_t^\theta\le A\}$, and $L_A=\max_{\theta\in\mathbf{\Theta}}L_A^\theta$. Then

$$\limsup_{A\to\infty}\frac{\mathsf{E}_p\{L_A\}}{A}\le\frac{1}{\min_{\theta\in\mathbf{\Theta}}\mathsf{E}_p\{f^\theta(Y_1)\}}.\tag{28}$$

Note: By incorporating Cramér's theorem during the proof of this lemma in [12], it can be extended to continuous r.v.'s $Y_1,Y_2,\ldots$, provided $\mathsf{E}_p\{|f^\theta(Y_1)|\}$ and $\mathsf{E}_p\{|f^\theta(Y_1)|^2\}$ are finite for all $\theta$.

## Appendix VI
## Proof of Theorems 7 and 8

*Proof of Theorem 7 ($\log t$ Lower Bound):* This proof is basically a variation of that for Theorem 5, with the major difference being that the competing configuration $C'=(\theta,\theta_2)$ is now from a different set: $\{\theta:\exists x_0,\mu_\theta(x_0)>\mu_{\theta_2}(x_0)\}$. We can first follow line by line in the proof of Theorem 5, and replace (16) with the following inequality:

$$\begin{aligned}
&\mathsf{E}_{C'}\{T_{\inf}(t)\}\\
&\ge \mathsf{E}_{C'}\left\{\sum_{\tau=1}^{t}1\{\phi_\tau=2,M_{C'}(X_\tau)=1\}\right\}\\
&= \mathsf{E}_{C'}\left\{\sum_{\tau=1}^{t}1\{M_{C'}(X_\tau)=1\} - \sum_{\tau=1}^{t}1\{\phi_\tau=1,M_{C'}(X_\tau)=1\}\right\}
\end{aligned}$$

$$
\begin{aligned}
&\geq \mathsf{E}_{C'}\left\{\sum_{\tau=1}^{t} 1\{M_{C'}(X_\tau)=1\} - \sum_{\tau=1}^{t} 1\{\phi_\tau = 1\}\right\}\\
&\overset{(b)}{=} \mathsf{E}_{C'}\{\pi t - T_1(t)\}\\
&\overset{(c)}{\geq} \left(\pi t - \frac{\log t}{(1+2\delta)K_{C'}}\right)\mathsf{P}_{C'}(A_1)\\
&\overset{(d)}{\geq} \left(\pi t - \frac{\log t}{(1+2\delta)K_{C'}}\right)\mathsf{P}_{C'}(A_1 \cap A_2)\\
&\overset{(e)}{\geq} \left(\pi t - \frac{\log t}{(1+2\delta)K_{C'}}\right) e^{-\frac{(1+\delta)\log t}{1+2\delta}}\mathsf{P}_{C_0}(A_1 \cap A_2)\\
&\overset{(f)}{=} \mathcal{O}\left(t^{\frac{\delta}{1+2\delta}}\right)
\end{aligned}
$$

where the first inequality follows from dropping the other half of the events where $\{\phi_\tau = 1, M_{C'}(X_\tau) = 2\}$. The second inequality follows from dropping the condition $M_{C'}(X_\tau) = 1$. With $\pi := \mathsf{P}_G(M_{C'}(X_\tau) = 1) > 0$, recalling that $\theta'$ satisfies that $\exists x_0$, such that $M_{C'}(x_0) = 1$, we obtain $(b)$. $(c)$–$(f)$ follow from the same reasoning as discussed in connection with (16). From the contradiction of the uniformly good rule assumption, we have

$$\lim_{t\to\infty} \mathsf{P}_{C_0}\left(T_1(t) \geq \frac{(1-\epsilon)\log t}{K_{C'}}\right) = 1 \qquad \forall \epsilon > 0.$$

By choosing the $\theta$ in $C' = (\theta, \theta_2)$ with the minimizing configuration $\inf_{\{\theta:\exists x, \text{s.t. } \mu_\theta(x) > \mu_{\theta_2}(x)\}} \sup_x I(\theta_1, \theta \,|\, x)$, the proof of the first statement in Theorem 7 follows. The second statement in Theorem 7 can be obtained by simply applying Markov's inequality and the first statement. ∎

*Proof of Theorem 8 (Bound-Achieving Scheme):* Following the same path as in the proof of Theorem 6, we first decompose the inferior sampling time instants into disjoint subsequences, each of which will be discussed separately

$$
\begin{aligned}
&T_{\inf}(t)\\
&\quad= \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau)\}\\
&\quad= \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \mathsf{Cond0}(\tau)\}\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \mathsf{Cond1}(\tau)\}\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \mathsf{Cond2}(\tau)\}\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \mathsf{Cond2.5}(\tau)\},\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta)\\
&\qquad\qquad \theta^\alpha \prec \theta^\beta \neq \theta_2, \mathsf{Cond3}(\tau)\},\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta)\\
&\qquad\qquad \theta_1 \neq \theta^\alpha \succ \theta^\beta, \mathsf{Cond3}(\tau)\},\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta)\\
&\qquad\qquad \theta_1 \neq \theta^\alpha \prec \theta^\beta = \theta_2, \mathsf{Cond3}(\tau)\},\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta)\\
&\qquad\qquad \theta_1 = \theta^\alpha \succ \theta^\beta \neq \theta_2, \mathsf{Cond3}(\tau)\},\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau)\\
&\qquad\qquad \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta) = C_0 = (\theta_1, \theta_2), \mathsf{Cond3}(\tau)\}.
\end{aligned}
\tag{29}
$$

By exactly the same analysis as in Lemmas 2 and 3, the first two sums in (29), concerning Cond0 and Cond1, have bounded expectations. Let $\bar{\theta}$ denote the configuration satisfying $\forall x_0, \mu_{\bar{\theta}}(x_0) \leq \mu_{\bar{\theta}}(x_0)$. For the sum concerning Cond2, $\{\phi_\tau \neq M_{C_0}(X_\tau), \mathsf{Cond2}(\tau)\}$ implies it is either $\theta^\alpha = \bar{\theta} \neq \theta_1$ or $\theta^\beta = \bar{\theta} \neq \theta_2$, where $(\theta^\alpha, \theta^\beta) = \hat{C}_{\tau-1}$. Both of the previous cases are discussed in Lemma 4 and are proved to have finite expectations.

For future reference, we denote the five different sums concerning Cond3 as term3a, term3b, term3c, term3d, and term3e, in order. By Lemma 5 and Corollary 1, both term3a and term3b have bounded expectations.

If the underlying $C_0$ is not implicitly revealing, by Lemmas 6 and 7, term3c and term3d have bounded expectation. And by Lemma 8, $\limsup_t((\mathsf{E}_{C_0}\{\mathsf{term3e}\})/(\log t)) \leq K_{C_0}$.

If the underlying $C_0$ is implicitly revealing, term3e $= 0$. For term3c and term3d, we have

$$
\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta),\\
&\quad \theta_1 \neq \theta^\alpha \prec \theta^\beta = \theta_2, \mathsf{Cond3}(\tau)\}\\
&\qquad+ \sum_{\tau=1}^{t} 1\left\{\phi_\tau \neq M_{C_0}(X_\tau), \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta),\right.\\
&\qquad\quad \left.\theta_1 = \theta^\alpha \succ \theta^\beta \neq \theta_2, \mathsf{Cond3}(\tau)\right\}\\
&\leq \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta),\\
&\quad \theta_1 \neq \theta^\alpha \prec \theta^\beta = \theta_2, \mathsf{Cond3}(\tau)\}\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau = 2, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta),\\
&\qquad\quad \theta_1 \neq \theta^\alpha \prec \theta^\beta = \theta_2, \mathsf{Cond3}(\tau)\}\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau = 1, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta),\\
&\qquad\quad \theta_1 = \theta^\alpha \succ \theta^\beta \neq \theta_2, \mathsf{Cond3}(\tau)\}\\
&\qquad+ \sum_{\tau=1}^{t} 1\{\phi_\tau = 2, \hat{C}_{\tau-1} = (\theta^\alpha, \theta^\beta),\\
&\qquad\quad \theta_1 = \theta^\alpha \succ \theta^\beta \neq \theta_2, \mathsf{Cond3}(\tau)\}
\end{aligned}
\tag{30}
$$

which is obtained by replacing the condition $\phi_\tau \neq M_{C_0}(X_\tau)$ with either $\phi_\tau = 1$ or $\phi_\tau = 2$. By Lemma 6, both the first and the fourth sums in (30) have bounded expectations.

By Lemma 7, both the second and the third sums in (30) also have bounded expectations.

Note: In the proofs of Lemmas 6–8, there are summations or minima taken on the set $\{\theta > \theta^{\beta}\}$. All those sets could be replaced by $\{\theta\colon \exists x_0, \text{s.t. } \mu_{\theta}(x_0) > \mu_{\theta^{\beta}}(x_0)\}$ and the rest of the proofs still follow.

We have discussed all sub-sums in (29) except the sum regarding Cond2.5. It remains to show that the sum concerning Cond2.5 has bounded expectation, which is addressed in the following lemma. ■

*Lemma 10:* Consider the $\{\phi_\tau\}$ described in Algorithm 4. For all possible $C_0$, we have

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \textsf{Cond2.5}(\tau)\}\right\} < \infty.$$

*Proof:*

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\phi_\tau \neq M_{C_0}(X_\tau), \textsf{Cond2.5}(\tau)\}\\
&\leq \sum_{(\theta,\vartheta):(\theta,\vartheta)\neq C_0} \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta,\vartheta), \textsf{Cond2.5}(\tau)\}\\
&= \sum_{(\theta,\vartheta):(\theta,\vartheta)\neq C_0} \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta,\vartheta), \textsf{Cond2.5}(\tau),\\
&\quad L_X(\tau \,|\, \textsf{Cond2.5}) \in \delta\text{−nbd}(G)\}\\
&\qquad + \sum_{(\theta,\vartheta):(\theta,\vartheta)\neq C_0} \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta,\vartheta), \textsf{Cond2.5}(\tau),\\
&\qquad\quad L_X(\tau \,|\, \textsf{Cond2.5}) \notin \delta\text{−nbd}(G)\}. \qquad (31)
\end{aligned}$$

By Sanov's theorem on finite alphabets, each term in the second sum is exponentially upper bounded w.r.t. $\tau$, which implies that the second sum has finite expectation. For the first sum, we have

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta,\vartheta), \textsf{Cond2.5}(\tau),\\
&\quad L_X(\tau \,|\, \textsf{Cond2.5}) \in \delta\text{−nbd}(G)\}\\
&\leq \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta,\vartheta), \theta \neq \theta_1, \textsf{Cond2.5}(\tau),\\
&\quad L_X(\tau \,|\, \textsf{Cond2.5}) \in \delta\text{−nbd}(G)\}\\
&\qquad + \sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta,\vartheta), \vartheta \neq \theta_2, \textsf{Cond2.5}(\tau),\\
&\qquad\quad L_X(\tau \,|\, \textsf{Cond2.5}) \in \delta\text{−nbd}(G)\} \qquad (32)
\end{aligned}$$

which is obtained by considering whether $\theta \neq \theta_1$ or $\vartheta \neq \theta_2$, recalling that $(\theta,\vartheta) \neq C_0$. Since these two sums are symmetric, henceforth we show only the finite expectation of the first sum in (32). The finite expectation of the second sum then follows by symmetry

$$\begin{aligned}
&\sum_{\tau=1}^{t} 1\{\hat{C}_{\tau-1} = (\theta,\vartheta), \theta \neq \theta_1, \textsf{Cond2.5}(\tau),\\
&\quad L_X(\tau \,|\, \textsf{Cond2.5}) \in \delta\text{−nbd}(G)\}\\
&\leq \sum_{\tau=1}^{t} 1\{\exists x, \text{s.t.} M_{(\theta,\vartheta)}(x) = 1, \rho\left(L_1^x(\tau-1), F_{\theta_1}(\cdot \,|\, x)\right) > \epsilon,\\
&\quad \textsf{Cond2.5}(\tau), L_X(\tau \,|\, \textsf{Cond2.5}) \in \delta\text{−nbd}(G)\}\\
&\leq \sum_{x:M_{(\theta,\vartheta)}(x)=1} \sum_{\tau'=1}^{\infty} 1\{\exists n \geq [\tau' \mathsf{P}_G(X=x)(1-\delta)],\\
&\quad \text{s.t.} \rho\left(L_1^x(n), F_{\theta_1}(\cdot \,|\, x)\right) > \epsilon\}. \qquad (33)
\end{aligned}$$

The first inequality follows from the definition of Cond2.5: since $\hat{C}_{\tau-1} = (\theta,\vartheta)$ is implicitly revealing, there must be an $x$ s.t. $M_{\hat{C}_{\tau-1}} = 1$. And since the estimate $\theta \neq \theta_1$, for that specific $x$, the distance between $L_1^x$ and $F_{\theta_1}(\cdot \,|\, x)$ must be greater than $\epsilon$. The second inequality follows from changing the time index to $\tau'$, the time instants at which $X_s = x$ and Cond2.5 is satisfied, and extending the summation to infinity. [This change of the time index is similar to the one described in (19) and (20)].

Thus, by Sanov's theorem on $\mathbb{R}$, the expectation of each term in (33) is exponentially upper bounded w.r.t. $\tau'$, which implies finite expectation of the entire sum in (33). By the discussions on (31)–(33), Lemma 10 is proved. ■

From the aforementioned discussion of the sub-sums in (29), we conclude that the modified scheme, $\{\phi_\tau\}$ in Algorithm 4, has bounded $\mathsf{E}_{C_0}\{T_{\inf}(t)\}$ if the underlying $C_0$ is implicitly revealing. If $C_0$ is not implicitly revealing, the $\{\phi_\tau\}$ in Algorithm 4 achieves the new $\log t$ lower bound (4). ■

## REFERENCES


[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, pp. 527–535, 1952.

[2] K. Adam, "Learning while searching for the best alternative," *J. Econ. Theory*, vol. 101, pp. 252–280, 2001.

[3] D. A. Berry, "A Bernoulli two-armed bandit," *Ann. Math. Stat.*, vol. 43, no. 3, pp. 871–897, Jun. 1972.

[4] H. Chernoff, *Sequential Analysis and Optimal Design*. Philadelphia, PA: SIAM, 1972.

[5] B. Ghosh and P. K. Sen, *Handbook of Sequential Analysis*. New York: Marcel Dekker, 1991.

[6] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Royal Stat. Soc. B*, vol. 41, no. 2, pp. 148–177, 1979.

[7] ——, "A dynamic allocation index for the discounted multiarmed bandit problem," *Biometrika*, vol. 66, no. 3, pp. 561–565, Dec. 1979.

[8] T. L. Lai and H. Robbins, "Asymptotically optimal allocation of treatments in sequential experiments," in *Design of Experiments: Ranking and Selection*, T. J. Santner and A. C. Tamhane, Eds. New York: Marcel Dekker, 1984.

[9] ——, "Asymptotically efficient allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[10] T. L. Lai and S. Yakowitz, "Machine learning and nonparametric bandit theory," *IEEE Trans. Autom. Control*, vol. 40, no. 7, pp. 1199–1209, Jul. 1995.

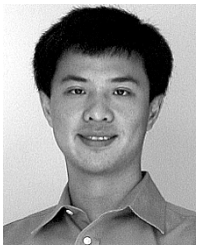
[11] R. Agrawal, M. V. Hegde, and D. Teneketzis, "Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost," *IEEE Trans. Autom. Control*, vol. 33, no. 10, pp. 899–906, Oct. 1988.

[12] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space," *IEEE Trans. Autom. Control*, vol. 34, no. 3, pp. 258–267, Mar. 1989.

[13] ——, "Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space," *IEEE Trans. Autom. Control*, vol. 34, no. 12, pp. 1249–1259, Dec. 1989.

[14] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part I: I.i.d. rewards," *IEEE Trans. Autom. Control*, vol. AC-32, no. 11, pp. 968–976, Nov. 1987.

[15] ——, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part II: Markovian rewards," *IEEE Trans. Autom. Control*, vol. AC-32, no. 11, pp. 977–982, Nov. 1987.
[16] M. N. Katehakis and H. Robbins, "Sequential choice from several populations," in *Proc. Nat. Acad. Sci.*, vol. 92, Sep. 1995, pp. 8584–8585.
[17] S. R. Kulkarni and G. Lugosi, "Finite-time lower bounds for the two-armed bandit problem," *IEEE Trans. Autom. Control*, vol. 45, no. 4, pp. 711–714, Apr. 2000.
[18] M. K. Clayton, "Covariate models for Bernoulli bandits," *Seq. Anal.*, vol. 8, no. 4, pp. 405–426, 1989.
[19] S. R. Kulkarni, "On bandit problems with side observations and learnability," in *Proc. 31st Allerton Conf. Communications, Control, Computing*, Sep. 1993, pp. 83–92.
[20] J. Sarkar, "One-armed bandit problems with covariates," *Ann. Statist.*, vol. 19, no. 4, pp. 1978–2002, 1991.
[21] M. Woodroofe, "A one-armed bandit problem with a concomitant variable," *J. Amer. Stat. Assoc.*, vol. 74, no. 368, pp. 799–806, Dec. 1979.
[22] T. Zoubeidi, "Optimal allocations in sequential tests involving two populations with covariates," *Commun. Statist.: Theory Meth.*, vol. 23, no. 4, pp. 1215–1225, 1994.
[23] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*. New York: Wiley, 1990.
[24] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*. New York, NY: Springer-Verlag, 1998.



**Chih-Chun Wang** received the B.E. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999. He is currently working toward the Ph.D. degree in electrical engineering at Princeton University, Princeton, NJ.

He was with COMTREND Corporation, Taipei, Taiwan, from 1999 to 2000, and spent the summer of 2004 with Flarion Technologies, Bedminster, NJ. His research interests are in optimal control, information theory, and coding theory, especially on iterative decoding of LDPC codes.

**Sanjeev R. Kulkarni** (M'91–SM'96–F'04) received the B.S. degree in mathematics, the B.S. degree in electrical engineering, and the M.S. degree in mathematics from Clarkson University, Potsdam, NY, in 1983, 1984, and 1985, respectively, the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1985, and the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1991.

From 1985 to 1991, he was a Member of the Technical Staff at MIT's Lincoln Laboratory, working on the modeling and processing of laser radar measurements. Since 1991, he has been with Princeton University, Princeton, NJ, where he is currently Professor of Electrical Engineering. He spent January 1996 as a Research Fellow at the Australian National University, Canberra, 1998 with Susquehanna International Group, BalaCynwyd, PA, and summer 2001 with Flarion Technologies, Bedminster, NJ. His research interests include statistical pattern recognition, nonparametric estimation, learning and adaptive systems, information theory, wireless networks, and image/video processing.

Dr. Kulkarni received an Army Research Office Young Investigator Award in 1992, a National Science Foundation Young Investigator Award in 1994, and several teaching awards at Princeton University. He has served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY.

**H. Vincent Poor** (S'72–M'77–SM'82–F'87) received the Ph.D. degree in electrical engineering and computer science from Princeton University, Princeton, NJ, in 1977.

He is currently the George Van Ness Lothrop Professor in Engineering at Princeton University. From 1977 to 1990, he was a faculty member at the University of Illinois at Urbana-Champaign. His research interests are primarily in the areas of stochastic analysis and statistical signal processing, with applications in wireless communications and related areas. Among his publications in this area is the recent book *Wireless Networks: Multiuser Detection in Cross-Layer Design* (New York: Springer-Verlag, 2005).

Dr. Poor is a Member of the National Academy of Engineering, and is a Fellow of the Institute of Mathematical Statistics, the Optical Society of America, and other organizations. In 1990, he served as the President of the IEEE Information Theory Society and he is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION THEORY. Recent recognition of his work includes the Joint Paper Award of the IEEE Communications and Information Theory Societies (2001), the National Science Foundation Director's Award for Distinguished Teaching Scholars (2002), a Guggenheim Fellowship (2002–2003), and the IEEE Education Medal (2005).