

ARBITRARY SIDE OBSERVATIONS IN BANDIT PROBLEMS

BY CHIH-CHUN WANG, SANJEEV R. KULKARNI, AND H. VINCENT POOR

Princeton University

A bandit problem with side observations is an extension of the traditional two-armed bandit problem, in which the decision maker has access to side information before deciding which arm to pull. In this paper, essential properties of the side observations that allow achievability results with respect to optimal regret are extracted and formalized. The sufficient conditions for good side information obtained here admit various types of random processes as special cases, including i.i.d sequences, Markov chains, deterministic periodic sequences, etc. A simple necessary condition for optimal regret is given, providing further insight into the nature of bandit problems with side observations. A game-theoretic approach simplifies the analysis and justifies the viewpoint that the side observation serves as an index of different sub-bandit machines.

1. Introduction. The classical two-armed bandit problem can be described in the context of finding the optimal choice between two slot machines, in which the reward distribution of each machine is unknown. Let $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$ denote the respective reward sequences from machines 1 and 2. The reward function is then defined as follows,

$$W_\phi(t) = \sum_{\tau=1}^t (1_{\{\phi_\tau=1\}}Y_\tau^1 + 1_{\{\phi_\tau=2\}}Y_\tau^2),$$

where ϕ_τ , taking values in $\{1, 2\}$, is the player's strategy at time τ and depends only on the history before τ . The learning task must be accomplished by letting ϕ_t sample both arms while also serving the mission of maximizing $E\{W_\phi(t)\}$. Due to the inherent nature of coordinated learning and control, bandit problems have drawn much attention in various areas of statistics, control, learning, and economics, as in [Ada01, Ber72, Che72, GP91, Git79a, Git79b, LR84, LR85, LY95, Rob52].

Bandit problems with side information involve another process X_t , which is observed before the decision ϕ_t is made. That is, at the current time t , in addition to the history strictly before time t , ϕ_t is also a function of X_t . This additional information helps our decision as long as the random processes $\{X_\tau\}$ and $(\{Y_\tau^1\}, \{Y_\tau^2\})$

AMS 2000 subject classifications. Primary 93E35; Secondary 93E20, 93E10.

Key words and phrases. Two-armed bandit, arbitrary, side information, regret, allocation rule, asymptotic, efficient, adaptive, evenly distributed.

This work was supported by the National Science Foundation under Grant Number ECS-9873451, the Office of Naval Research under Grant Number N00014-03-1-0102, the Army Research Office under Contract Number DAAD19-00-1-0466, and the New Jersey Center for Pervasive Information Technologies.

are correlated. This idea was first introduced by Woodroffe [Woo79], where an independent and identically distributed (i.i.d.) $\{X_\tau\}$ was considered. Woodroffe proved that a myopic approach is asymptotically optimal under a simple relationship between the i.i.d. X_t and (Y_t^1, Y_t^2) , namely $Y_t^i = \theta_i + X_t + N_t$, where $\{N_\tau\}$ is an i.i.d. standard Gaussian random sequence. Sarkar [Sar91] extended the simple relationship discussed in [Woo79] to exponential families. In [WKP04], by using the side observations X_t as the index of different bandit machines with common parameter pair, different levels of asymptotic efficiency improvement were found for i.i.d. $\{X_\tau\}$ and several types of relationships between $\{X_\tau\}$ and $(\{Y_\tau^1\}, \{Y_\tau^2\})$, which includes the previous results as special cases. Other approaches regarding side observations can be found in [Cla89, Kul93, Zou94].

Previous work on bandit problems with side observations has considered only i.i.d. $\{X_\tau\}$. However, the results of [Woo79, Sar91, WKP04] also suggest that the benefits of side observations on bandit problems is not due to the *random* appearance of all values x of an i.i.d. $\{X_\tau\}$, but rather is due to the *evenly* distributed appearance of all possible x . In this paper, we extract the essential properties of “evenly distributed appearance” and investigate their effects on the attainable results. Thus, our results generalize the benefit of side observations to a wide range of non-i.i.d. processes.

To motivate this approach, we consider an example within the parametric, asymptotic setting explored by [LR85, AHT88, ATA89a, ATA89b, AVW87a, AVW87b, KR95, KL00].

EXAMPLE: Suppose Y_t^1 and Y_t^2 are two Bernoulli random variables, which denote the success of transmitting a single information block over a communication channel at time t , under different modulation techniques MD1 and MD2. Each time we transmit an information block, only one modulation technique can be applied. The channel properties depend on a parameter pair denoted by $C_0 = (\theta_1, \theta_2) \in \Theta^2$, where we assume Y_t^1 depends only on θ_1 , Y_t^2 depends only on θ_2 , and θ_1, θ_2 take values in Θ . X_t is some observation at time t taking values in a finite set \mathbf{X} . $\{X_\tau\}$ is not necessarily i.i.d. and may or may not reveal information about C_0 , but is correlated with Y_t^1 and Y_t^2 . For example, X_t might be a noisy measurement of the parameter pair (θ_1, θ_2) , or geographical information about the receiver (affecting Y_t^1 and Y_t^2), or X_t may be a pair containing both of the above. Our goal is to design a transmission scheme that minimizes the growth rate of the expected number of uses of the inferior modulation scheme.

This scenario is analogous to a two-armed bandit machine: $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$ are associated with the rewards from pulling arms 1 and 2 respectively, where the distributions depend on the configuration pair $C_0 = (\theta_1, \theta_2)$. To formally describe this bandit problems, we define some necessary notation and several quantities of interest in TABLE 1, and it is assumed throughout that all necessary expectations exist and are finite.

With the side observation X_t , the goal of selecting the better modulation scheme is equivalent to simultaneously minimizing the growth rate of the sampling of the

TABLE 1
Glossary

Not'n	Description
Θ, Θ^2	$\Theta \subseteq \mathbb{R}$ is the set of all possible θ ; Θ^2 is the space of parameter pairs.
\mathbf{X}	The set of possible X_t 's, and is assumed to be finite.
$G_{t_1, \dots, t_k C_0}(x_{t_1}, \dots, x_{t_k})$	The finite dimensional distribution of $\{X_\tau\}$, under the configuration C_0 .
$F_{\theta_i}(y x)$	The conditional distribution of arm i , Y_t^i , given $X_t = x$ and the unknown parameter θ_i .
$\mu_\theta(x)$	The conditional expectation of the reward, $\mu_\theta(x) = \mathbf{E}_\theta\{Y x\} = \int y F_\theta(dy x)$.
$1(C_0), 2(C_0)$	The first and the second coordinates of the configuration pair C_0 , i.e., if $C_0 = (\theta_1, \theta_2)$, then $1(C_0) = \theta_1$, $2(C_0) = \theta_2$. For example: $\mu_{1(C_0)}(x) = \mu_{\theta_1}(x)$.
$M_{C_0}(x)$	The index of the preferred arm, namely, $\arg \max_{i=1,2} \{\mu_{i(C_0)}(x)\}$.
ϕ_t	The decision rule taking values $\{1, 2\}$ and depending only on the past outcomes and the current side information X_t .
$T_i(t)$	The total number of samples taken from 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 : $I(P, Q) = \mathbf{E}_P \left\{ \log \left(\frac{dP}{dQ} \right) \right\}$.
$I(\theta_1, \theta_2 x)$	The conditional K-L information number: $I(\theta_1, \theta_2 x) = I(F_{\theta_1}(\cdot x), F_{\theta_2}(\cdot x))$.

inferior arm, denoted by $T_{inf}(t)$, over all C_0 , where

$$T_{inf}(t) = \sum_{\tau=1}^t 1_{\{\phi_\tau \neq M_{C_0}(X_\tau)\}}.$$

Following [LR85], we define the notion of simultaneously minimizing the growth rate over all possible C_0 by *uniformly good rules* as follows.

DEFINITION 1.1 UNIFORMLY GOOD RULES. *An allocation rule is uniformly good¹ if for all $C_0 = (\theta_1, \theta_2)$, $\mathbf{E}_{C_0}\{T_{inf}(t)\} = o(t^\alpha)$, $\forall \alpha > 0$.*

Henceforth we consider only uniformly good rules and regard other rules as uninteresting.

The traditional two-armed bandit without side observations X_t is simply a degenerated case and is equivalent to having $\mathbf{X} = \{x_0\}$ containing only a single element x_0 . Suppose the reward sequences $\{Y_\tau^i\}$ are i.i.d. A $\log t$ lower bound has

¹In the literature on bandit problems, the “regret” is more frequently used than $T_{inf}(t)$ when defining uniformly good rules. For traditional two-armed bandits, the regret of a strategy $\{\phi_\tau\}$ is defined as

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

been proved for all uniformly good rules [AVW87a, LR84, LR85], which is quoted as follows.

THEOREM 1.1 *log t LOWER BOUND.* For any uniformly good rule $\{\phi_\tau\}$, $T_{inf}(t)$ satisfies

$$(1.1) \quad \lim_{t \rightarrow \infty} \mathbb{P}_{C_0} \left(T_{inf}(t) \geq \frac{(1 - \epsilon) \log t}{K_{C_0}} \right) = 1$$

and $\liminf_{t \rightarrow \infty} \frac{\mathbb{E}_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}},$

where $\epsilon > 0$ is an arbitrary constant and K_{C_0} is a constant depending on C_0 . If $M_{C_0}(x_0) = 2$, then $T_{inf}(t) = T_1(t)$ and K_{C_0} is defined as follows.

$$(1.2) \quad K_{C_0} = \inf\{I(\theta_1, \theta|x_0) : \forall \theta, \mu_\theta(x_0) > \mu_{\theta_2}(x_0)\}.$$

The expression for K_{C_0} in the case that $M_{C_0}(x_0) = 1$ can be obtained by symmetry.

The asymptotic tightness of the above lower bound is also obtained in these papers:

THEOREM 1.2 *ASYMPTOTIC TIGHTNESS.* Under certain regularity conditions², the above lower bound is asymptotically tight. Formally stated, given the distribution family $\{F_\theta(\cdot|x)\}$, there exists a decision rule $\{\phi_\tau\}$ such that for all $C_0 = (\theta_1, \theta_2) \in \Theta^2$,

$$\limsup_{t \rightarrow \infty} \frac{\mathbb{E}_{C_0}\{T_{inf}(t)\}}{\log t} \leq \frac{1}{K_{C_0}},$$

where K_{C_0} is the same as in THEOREM 1.1.

When \mathbf{X} contains more than one element, the extent of the improvement by observing X_t in advance depends on the relationship of $\{X_\tau\}$ to both the configuration C_0 and the reward random processes $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$. Under certain scenarios, the $\log t$ lower bound for traditional bandit problems can be surpassed. To explicitly characterize the correlation between C_0 , $\{X_\tau\}$, $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$, the probability distribution of the two-armed bandit with side observations is modelled as follows. At time t_1, \dots, t_k , the joint probability distribution of $(X_{t_i}, Y_{t_i}^1, Y_{t_i}^2)_{i=1, \dots, k}$ is

$$G_{t_1, \dots, t_k | C_0}(x_{t_1}, \dots, x_{t_k}) \prod_{i=1}^k F_{\theta_1}(y_{t_i}^1 | x_{t_i}) F_{\theta_2}(y_{t_i}^2 | x_{t_i}).$$

the difference between the best possible reward and the expected reward using $\{\phi_\tau\}$. The linear relationship between the regret and $T_{inf}(t)$ is as follows.

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

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

²If the parameter space is finite, THEOREM 1.2 always holds. If Θ is continuous, the required regularity conditions are on the unboundedness and the continuity of $\mu_\theta(x_0)$ w.r.t. θ and on the continuity of $I(\theta_1, \theta|x_0)$ w.r.t. $\mu_\theta(x_0)$.

In this parametric setting, both the families of the distributions $\{G_{\dots|C_0}\}_{C_0}$ and $\{F_{\theta}(\cdot|x)\}_{\theta}$ are known to the decision maker, but the true values of the corresponding θ_1 and θ_2 must be learned through experiments. Note that the concept of the i.i.d. bandit is now extended to the assumption that conditioning on the sequence $\{X_{\tau}\}$, $\{Y_{\tau}^i\}$ is independently distributed. However, the player is now facing $F_{\theta_i}(x)$, the conditional distribution of Y_t^i , which is a function of the observed side information X_t . This is the problem to be examined in this paper.

This paper is organized as follows. In Section 2, we provide formal definitions of several “even distribution” properties, examples of each property, and relationships among them. In Sections 3 through 6, we provide results for different relationships among $\{X_{\tau}\}$, $\{Y_{\tau}^1\}$, and $\{Y_{\tau}^2\}$ with the satisfaction of the “even distribution” properties in Section 2. All results in [WKP04], obtained under the assumption of i.i.d. $\{X_{\tau}\}$, hold as special cases of this new general framework presented. The new framework includes many other other side observation processes (e.g. Markov chains and periodic sequences) as well. Section 7 provides a simple necessary condition concerning the extent of the benefit from observing $\{X_{\tau}\}$.

For the sake of readability, formal statements of our results are provided in each section, while details of the proofs are included in the appendix. In the following development, we will make use of three different levels of required conditions, which are named as follows.

- Ch1,Ch2,...: *Characterization conditions* specify the relationships between X_t and (Y_t^1, Y_t^2) .
- R1,R2,...: *Regularity conditions* are general enough to be satisfied for most cases, and may be removed by adding more complexity in the proof/analysis.
- A1,A2,...: *Assumptions* are the conditions required in the proof/analysis, which are not stringent but may not be as general as regularity conditions.

2. Even Distribution Properties Our goal is to extract the essential properties of a side observation process that are helpful to the improvement of uniformly good rules.

Suppose X_t takes value in a finite state set \mathbf{X} . The following properties of “even distribution” among all different values of x are defined by the use of the relative frequency of x up to time t , which is denoted as $f_r(x, t) = \left(\sum_{\tau=1}^t 1\{X_{\tau} = x\}\right) / t$.

DEFINITION 2.1 EVENLY DISTRIBUTED IN L^1 . $\{X_{\tau}\}$ is evenly distributed in L^1 if

$$\forall x \in \mathbf{X}, \quad \pi(x) := \liminf_{t \rightarrow \infty} \mathbf{E}\{f_r(x, t)\} > 0.$$

DEFINITION 2.2 EVENLY DISTRIBUTED IN PROBABILITY SERIES. $\{X_{\tau}\}$ is evenly distributed “in probability series” if there exists a strictly positive mapping $\pi(\cdot) > 0$, such that the duration of the event $\{f_r(x, t) < \pi(x)\}$ being satisfied has a

finite expectation. That is,

$$\forall x \in \mathbf{X}, \mathbb{E} \left\{ \sum_{\tau=1}^{\infty} 1\{f_r(x, \tau) < \pi(x)\} \right\} < \infty.$$

This property automatically implies that $\forall x, \liminf_{t \rightarrow \infty} f_r(x, t) \geq \pi(x)$ almost surely.

DEFINITION 2.3 UNIFORMLY STRONGLY EVENLY (U.S.E.) DISTRIBUTED IN L^1 . $\{X_\tau\}$ is u.s.e. distributed in L^1 , if for any stopping time T , the conditional expectation of the first hitting time of x after T has a global upper bound. That is, $\exists B < \infty$ such that

$$\forall T, \forall x \in \mathbf{X}, \mathbb{E}\{H_T(x)|T\} \leq B,$$

where $H_T(x) \triangleq \inf\{l > 0 | X_{T+l} = x\}$.

The following examples demonstrate that the above properties are quite general and can be applied to many interesting random processes.

- *Example 1:* If $\{X_\tau\}$ is an i.i.d. random process with strictly positive probability on each x , then by the results of the large deviations theorems, $\{X_\tau\}$ is evenly distributed in L^1 , evenly distributed in probability series, and u.s.e. distributed in L^1 .
- *Example 2:* If $\{X_\tau\}$ is a finite Markov chain with strictly positive entries in its transition matrix, then by similar reasoning as the i.i.d. case, $\{X_\tau\}$ is evenly distributed in L^1 , evenly distributed in probability series, and u.s.e. distributed in L^1 .
- *Example 3:* If $\{X_\tau\}$ is a deterministic periodic sequence, (e.g., $\{1, 2, 2, 3, 1, 2, 2, 3, 1, \dots\}$), then by definition, $\{X_\tau\}$ is evenly distributed in L^1 , evenly distributed in probability series, and u.s.e. distributed in L^1 .

Remark: Let $A_{u.s.e.}$ denote the family of $\{X_\tau\}$ being u.s.e. distributed in L^1 , and $A_{p.s.}$ and A_{L^1} denote the corresponding families when considering $\{X_\tau\}$ being evenly distributed in probability series and evenly distributed in L^1 . It can be shown that both $A_{u.s.e.}$ and $A_{p.s.}$ are proper subsets of A_{L^1} , and $A_{p.s.} \setminus A_{u.s.e.} \neq \emptyset$. The proof of these implications and some further notes on these properties are provided in APPENDIX E.

3. Direct Information

We now consider the implications of the first of several ways in which the side observations relate to the reward sequences.

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

- Direct Information(Ch1): If $C_0 \neq C'_0$, $\exists t_1, \dots, t_k$, such that $G_{t_1, \dots, t_k|C_0} \neq G_{t_1, \dots, t_k|C'_0}$.

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

3.2. *Scheme of Separating Learning and Control* Since we are able to obtain information about C_0 from $\{X_\tau\}$, it is natural to sample only the seemingly better arm while leaving the learning task to $\{X_\tau\}$. A corresponding control scheme ϕ_t is as follows.

- Step 1: At time³ t , obtain an estimate \hat{C}_t based on the side observations X_1, \dots, X_t .
- Step 2: Set $\phi_t = M_{\hat{C}_t}(X_t)$.

To give an upper bound on the performance of this scheme, we need the following condition.

CONDITION 3.1 (A1). *For any fixed C_0 and any convergent sequence $\{\hat{C}_\tau\} \rightarrow C_0$, there exists τ_0 such that $\forall x \in \mathbf{X}, \tau > \tau_0, M_{\hat{C}_\tau}(x) = M_{C_0}(x)$. Or equivalently,*

$$\inf \{ \rho(G_{C_0}, G_C) : C \in \Theta^2, \exists x, \text{ s.t. } M_C(x) \neq M_{C_0}(x) \} > 0,$$

for some metric ρ on the space of distributions.

- *Example 4:* With the assumptions that \mathbf{X} is finite, and $\forall x \in \mathbf{X}, \mu_\theta(x)$ is continuous with respect to θ , A1 is satisfied.
- *Example 5:* If $F_\theta(\cdot|x) \sim \mathcal{N}(\theta x, 1)$, a standard Gaussian distribution with mean θx , then A1 is satisfied.

THEOREM 3.1. *Suppose both Ch1 and A1 hold. For all C_0 and any sequence of estimates $\{\hat{C}_\tau\}$, there exists $\epsilon > 0$ such that the inferior sampling time $T_{inf}(t)$ of the above scheme satisfies*

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}_{C_0}\{T_{inf}(t)\}}{\sum_{\tau=1}^t \mathbb{P}_{C_0}(|\hat{C}_\tau - C_0| > \epsilon)} \leq 1.$$

A detailed proof is given in APPENDIX A.

The above theorem provides an upper bound on the best achievable expected inferior sampling time.

³“At time t ” means after observing X_t but before the decision ϕ_t is made. It is basically the moment when we are decoding the value of ϕ_t .

COROLLARY 3.1. *If for all $C_0, \epsilon > 0, \lim_{t \rightarrow \infty} \sum_{\tau=1}^t \mathbb{P}_{C_0}(|\hat{C}_\tau - C_0| > \epsilon)$ is finite, then $\forall C_0, \lim_{t \rightarrow \infty} \mathbb{E}_{C_0}\{T_{inf}(t)\}$ is finite as well.*

- *Example 6:* If $\{X_\tau\}$ is an i.i.d. sequence with marginal distribution G_{C_0} , and the mapping from C_0 to G_{C_0} is one-to-one, then by the large deviation theorem on finite alphabets, there exists $\{\hat{C}_\tau\}$ such that $\forall C_0, \epsilon > 0, \lim_{t \rightarrow \infty} \sum_{\tau=1}^t \mathbb{P}_{C_0}(|\hat{C}_\tau - C_0| > \epsilon) < \infty$. Hence $\forall C_0, \lim_{t \rightarrow \infty} \mathbb{E}_{C_0}\{T_{inf}(t)\} < \infty$, and the proposed scheme is uniformly good.
- *Example 7:* If $\{X_\tau\}$ is a finite Markov chain with transition matrix A_{C_0} , and the mapping from C_0 to A_{C_0} is one-to-one, then by the same reasoning as in the i.i.d. case, there exists a uniformly good rule such that $\forall C_0, \lim_{t \rightarrow \infty} \mathbb{E}_{C_0}\{T_{inf}(t)\} < \infty$.
- *Example 8:* Consider the case in which $\{X_\tau\}$ is a deterministic sequence denoted by $\{x_\tau\}_{C_0}$. If the mapping from C_0 to $\{x_\tau\}_{C_0}$ is one-to-one, and Θ is finite, then there exists $\{\hat{C}_\tau\}$ such that $\forall C_0, \epsilon > 0, \lim_{t \rightarrow \infty} \sum_{\tau=1}^t \mathbb{P}_{C_0}(|\hat{C}_\tau - C_0| > \epsilon) < \infty$. Hence $\forall C_0, \lim_{t \rightarrow \infty} \mathbb{E}_{C_0}\{T_{inf}(t)\} < \infty$, and the proposed scheme is uniformly good.

4. Best Arm As a Function of X_t

We now turn to another formulism for the interaction of X_t with the bandits. In particular, here and in the following two sections, we consider the following case.

- Ch2: 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} , the distribution of $\{X_\tau\}$, is not a function of C_0 , and we thus can use $G := G_{C_0}$ as shorthand.

Three cases of further refinements concerning the relationships between $M_C(x)$ and x will be discussed separately (each in one section).

4.1. *Formulation* In this section, we assume that the side observation X_t is *always* able to change the preference order, which is formally stated as follows and illustrated in FIG. 1.

- Best arm as a function of X_t (Ch3): $\forall C \in \Theta^2, \exists x_1, x_2 \in \mathbf{X}$, such that $M_C(x_1) = 1$ and $M_C(x_2) = 2$.

The necessary regularity conditions are as follows.

- R1: \mathbf{X} is finite.
- R2: $\forall \theta_1, \theta_2, x$, the K-L information number $I(\theta_1, \theta_2 | x)$ is finite and strictly positive.
- R3: $\Theta \subseteq \mathbb{R}$, and $\forall x, \mu_\theta(x)$ is continuous as a function of θ .

R1 embodies the idea of treating X_t as the index of several different bandit problems, which also simplifies our proof. R2 ensures that all these different bandit problems are non-trivial, i.e., with *non-identical* arms.

Example:

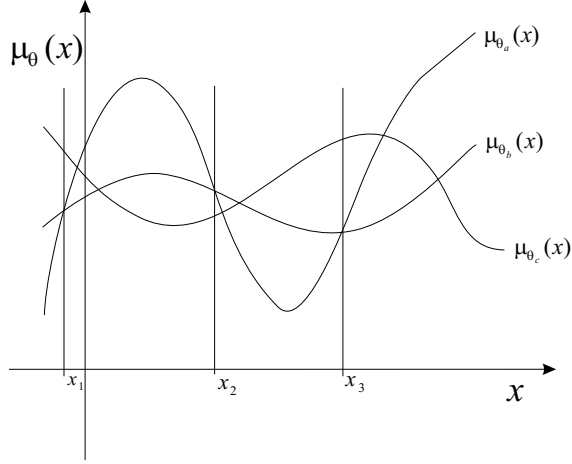


FIG. 1. The best arm at time t is a function of the side observation X_t . That is, for any possible pair $C = (\theta_1, \theta_2)$, the two curves, $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ (as a function of x) always intersect each other.

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

4.2. Scheme with Bounded $\lim_t \mathbb{E}_{C_0} \{T_{inf}(t)\}$

THEOREM 4.1. Suppose Ch2, Ch3, R1, R2, and R3 are satisfied. If the side observation sequence $\{X_\tau\}$ is evenly distributed in probability series, then there exists an allocation rule $\{\phi_\tau\}$ such that $\forall C_0$,

$$\lim_{t \rightarrow \infty} \mathbb{E}_{C_0} \{T_{inf}(t)\} < \infty.$$

Such a rule is uniformly good and surpasses the $\log t$ lower bound for traditional bandit problems.

Remark: 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 evenly distributed X_t takes on different values 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 ($\lim_t \mathbb{E}_{C_0} \{T_{inf}(t)\} < \infty$) is obtained over the $\log t$ lower bound in THEOREM 1.1.

We construct a scheme with $\lim_t \mathbb{E}_{C_0} \{T_{inf}(t)\} < \infty$ as follows.

- Step 1: We 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 we also define $x_i^* := \arg \max_x \{T_i^x(t)\}$ and $T_i^{x^*}(t) := \max_x \{T_i^x(t)\}$.

- Step 2: Sample the arms in a round robin fashion until $t = 6$, i.e., $\phi_1 = 1$, $\phi_2 = 2$, $\phi_3 = 1$, $\phi_4 = 2$, $\phi_5 = 1$, $\phi_6 = 2$.
- Step 3: At time $t + 1$, construct the following set,

$$\mathbf{C}_t = \left\{ C = (\theta_1, \theta_2) \in \Theta^2 : \sigma(C, t) \leq \inf\{\sigma(C, t) : C \in \Theta^2\} + \frac{1}{t} \right\},$$

where

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

$L_i^x(t)$ is the empirical measure of rewards sampled from arm i at those time instants $\tau \leq t$ when $X_\tau = x$, and $\rho(P, Q)$ is the Prohorov metric over distributions on \mathbb{R} . After constructing \mathbf{C}_t , arbitrarily choose $\hat{C}_t \in \mathbf{C}_t$.

- Step 4: At time $t + 1$, if $\exists i$ such that $T_i(t) < \sqrt{t+1}$, then we set $\phi_{t+1} = i$. Otherwise, $\phi_{t+1} = M_{\hat{C}_t}(X_{t+1})$. (Note that Step 2 guarantees that there is at most one i such that $T_i(t) < \sqrt{t+1}$.)

In the above scheme, since the probability of having a wrong estimate \hat{C}_t decreases exponentially with respect to the sample size, which is guaranteed by the forced sampling mechanism to be greater than $O(t^{\frac{1}{2}})$, the expected duration of $\{|\hat{C}_t - C_0| > \epsilon\}$ is bounded. When considering the traditional bandit setting, this forced sampling mechanism will inevitably result in $O(t^{\frac{1}{2}})$ inferior samplings, which is an undesired result. However, in the current setting, the alternating nature described in Ch3 and the even distribution property of $\{X_\tau\}$ make the myopic approach $\phi_{t+1} = M_{\hat{C}_t}(X_{t+1})$ automatically sample both arms evenly. Thus $T_1(t)$ and $T_2(t)$ grow linearly with t , and the forced sampling mechanism will terminate quickly. Since inferior sampling results either from incorrect estimates or from forced sampling, the resulting $\mathbb{E}_{C_0}\{T_{inf}(t)\}$ is finite.

A detailed proof is provided in APPENDIX B.

5. Best Arm Is Not A Function of X_t

5.1. *Formulation* Following Section 4, we assume Ch2: $G_{C_0} = G$, the distribution of $\{X_\tau\}$ does not vary with respect to C_0 . In this section, we consider the case that $\forall C_0, X_t$ never changes the preference order. This setting is illustrated in FIG. 2 and is formally stated as follows.

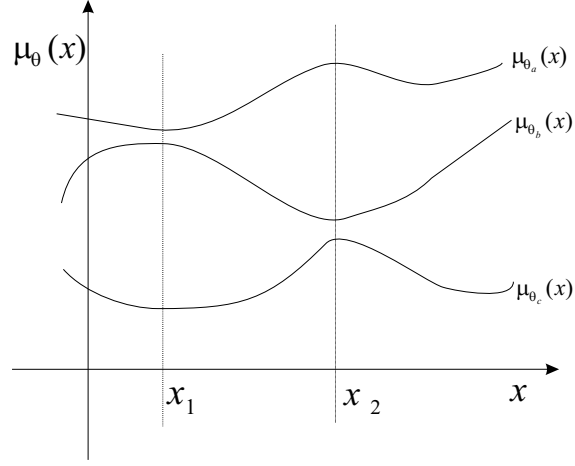


FIG. 2. *The best arm at time t is not a function of the side observation X_t . That is, for any possible pair, (θ_1, θ_2) , the two curves, $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$, do not intersect each other. In this case, we can postpone our sampling until the most informative time instants.*

- Best arm is not a function of X_t (Ch4): $\forall C = (\theta_1, \theta_2)$, $\theta_1 \neq \theta_2$, the preferred arm $M_C(x)$ is constant for all possible $x \in \mathbf{X}$. That is, we can use M_C as shorthand for $M_C(x)$.

Within the same two regularity conditions R1 and R2 in Section 4:

- R1: \mathbf{X} is finite, and
- R2: $\forall \theta_1, \theta_2, x$, the K-L information number $I(\theta_1, \theta_2 | x)$ is finite and strictly positive,

we can still have improvements over the traditional bandit problems. To simplify the notation, we assume further that

- R4: The parameter space $\Theta \subseteq \mathbb{R}$ can be relabelled, so that $\forall x$, the conditional expected reward $\mu_\theta(x)$ is strictly increasing in θ .

This relabelling gives us the convenience that the order of $(\mu_{\theta_1}(x), \mu_{\theta_2}(x))$ is the same as that of (θ_1, θ_2) .

Example:

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

5.2. Lower Bound In [WKP04], bandit problems with i.i.d. side observation $\{X_\tau\}$ are discussed, and a $\log t$ lower bound is provided under the same characterizations Ch2, Ch4, and regularity conditions R1, R2, R4. Since the proof of the $\log t$ lower bound remains exactly the same even when considering arbitrary random processes $\{X_\tau\}$, we restate the $\log t$ lower bound theorem [Theorem 5, p.13, [WKP04]] without proof.

THEOREM 5.1 [THEOREM 5, P.13, [WKP04]]. *Under Ch2, Ch4, R1, R2, and R4 for any uniformly good rule $\{\phi_\tau\}$, $T_{inf}(t)$ satisfies*

$$(5.1) \quad \lim_{t \rightarrow \infty} \mathbb{P}_{C_0} \left(T_{inf}(t) \geq \frac{(1-\epsilon) \log t}{K_{C_0}} \right) = 1,$$

and $\liminf_{t \rightarrow \infty} \frac{\mathbb{E}_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}},$

where $\epsilon > 0$ is an arbitrary constant and 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} is different than that of the traditional bandit problem and can be expressed as follows.

$$(5.2) \quad K_{C_0} = \inf_{\theta: \theta > \theta_2} \sup_{x \in \mathbf{X}} \{I(\theta_1, \theta|x)\}.$$

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

Note that if the decision maker is not able to access the side observation X_t at time t , the player will then face the *unconditional* reward distribution $\int G_{t,C_0}(x) F_{\theta_i}(y|x) dx$ rather than $F_{\theta_i}(y|x)$. 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

$$\sup_x I(\theta_1, \theta|x) \geq \int I(\theta_1, \theta|x) G_{t,C_0}(x) dx \geq I(\theta_1, \theta).$$

The above shows that the new constant in front of $\log t$, in (5.2), is smaller than the corresponding constant in (1.2), and the additional side information thus X_t improves the decision in the bandit problem, which of course it must.

5.3. *Scheme Achieving the Lower Bound* To construct a tractable scheme achieving the lower bound (5.1), we need the following assumptions.

- A2: The parameter space Θ is finite.
- A3: The side observations $\{X_\tau\}$ are u.s.e. distributed in L^1 .
- A4: The value of the game,

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

exists. A sufficient condition for the existence of the value of the game is that θ the dominating factor over x in the conditional distributions $F_\theta(\cdot|x)$, which is illustrated in the following example.

- *Example 9:* Consider the case in which $\Theta = \{1, 2, 3\}$, $\mathbf{X} = \{1, 2\}$ and the conditional distributions of Y_t^i given X_t can be written as $F_\theta(\cdot|x) \sim \mathcal{N}(\theta + xa_{\theta,x}, 1)$ for some constant coefficients $a_{\theta,x}$. If $a_{\theta,x} \in (-0.1, 0.1)$, $\forall \theta, x$, then the dominance of θ over x in the conditional distributions $F_\theta(\cdot|x)$ guarantees the existence of the value of the game.

In many cases of interest, the parameter plays a more critical role in determining the distribution than the side observation x . Therefore in these cases, condition A4 is generally satisfied and is a reasonable assumption.

To construct a $\log t$ -lower-bound achieving scheme, we first consider a specific type of decision rule for traditional bandit problems, which was introduced in [ATA89a] for finite parameter space Θ . This type of decision rule possesses the following properties.

1. After time t , an estimate $\hat{C}_t = (\hat{\theta}_1, \hat{\theta}_2)$ is found and used to make the decision ϕ_{t+1} . To be more explicit, \hat{C}_t is generated by the results for $\tau \in [1, t]$, and ϕ_{t+1} is a function of \hat{C}_t .
2. The expected duration over which $\hat{C}_t \neq C_0$ is finite⁴, namely,

$$\lim_{t \rightarrow \infty} \mathbb{E} \left\{ \sum_{\tau=1}^t \mathbf{1}\{\hat{C}_\tau \neq C_0\} \right\} < \infty.$$

3. The expected duration over which $\hat{C}_t = C_0$ and $\phi_t \neq M_{C_0}(X_t)$ is upper bounded by $\frac{\log t}{K_{C_0}}$, namely,

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E} \left\{ \sum_{\tau=1}^t \mathbf{1}\{\hat{C}_\tau = C_0, \phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\} \right\}}{\log t} < \frac{1}{K_{C_0}},$$

where $K_{C_0} = \inf_{\theta > \theta_2} I(\theta_1, \theta)$ if $M_{C_0} = 2$.

DEFINITION 5.1 TIGHT ϕ_t . *A decision rule ϕ_t , for a traditional bandit problem (without side observations) with finite Θ , is tight if it possesses the above three properties.*

Obviously, a *tight* rule ϕ_t is uniformly good and meets the $\log t$ lower bound on $T_{inf}(t)$ in THEOREM 1.1.

The detailed construction of a *tight* ϕ_t can be found in [ATA89a], and we construct a new decision rule Φ_t (for the side-observation-aided bandits) based on such a ϕ_t for the traditional bandits. Suppose $|\mathbf{X}| = k$. For different values of X_t , we can group the rewards Y_t^1 and Y_t^2 into sequences corresponding to k different sub-bandit machines, denoted by $\{\mathbf{B}_x\}_{x \in \mathbf{X}}$ such that \mathbf{B}_x corresponds to those time instants when $X_t = x$. For example, if $X_1 X_2 X_3 X_4 \cdots = x_a x_b x_a x_c \cdots$, then after time $t = 4$, we have 2 samples in \mathbf{B}_{x_a} , 1 sample in \mathbf{B}_{x_b} , and 1 sample in \mathbf{B}_{x_c} . Let $\hat{C}_{x,t}$ denote the corresponding estimates generated from the samples in \mathbf{B}_x after time t , and $\{\phi_{x,\tau}\}$ denote the tight decision rule for \mathbf{B}_x such that $\phi_{x,t+1}$, the decision at time $t + 1$, is a function of $\hat{C}_{x,t}$. This composite Φ_t for the side-observation-aided bandits is constructed as in **Algorithm 1**.

⁴In this paper, we use the convention that $\{\hat{C}_t \neq C_0\}$ represents both the cases that \hat{C}_t does not exist, and that \hat{C}_t exists but does not equal C_0 .

Algorithm 1 Φ_{t+1} , the decision at time $t + 1$

```

1: if not all  $\hat{C}_{x,t}$  are identical, then
2:    $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$ .
3: else
4:   Denote  $\hat{C}_t = (\hat{\theta}_1, \hat{\theta}_2)$  as the common estimate for all  $B_x$ . Without loss of
     generality, we may assume  $M_{\hat{C}_t} = 2$ . The case that  $M_{\hat{C}_t} = 1$  can be obtained
     by symmetry.
5:   if  $X_{t+1} \neq x^* := \arg \max_x \inf_{\{\theta: \theta > \hat{\theta}_2\}} I(\hat{\theta}_1, \theta|x)$ , then
6:      $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ .
7:   else
8:      $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$ .
9:   end if
10: end if

```

To perform a rigorous analysis, the constituent $\phi_{x,t}$ must be fully encapsulated. Namely, only those samples obtained from performing $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$, lines 2 and 8 in **Algorithm 1**, can be counted as valid samples for $\phi_{x,t}$. In other words, the time instants when we skip $\phi_{x,t}$ by setting $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ must be excluded from the computation of $\hat{C}_{x,t}$ and $\phi_{x,t+1}$, since otherwise it may spoil the *tightness* of the original $\phi_{x,t+1}$. For example, suppose $X_1 X_2 X_3 X_4 \cdots = x_a x_b x_a x_c \cdots$, but only at time instants 1 and 2 are lines 2 and 8 executed. Then from the individual B_x point of view, we only have 1 sample in B_{x_a} , 1 sample in B_{x_b} , and 0 sample in B_{x_c} , and only those samples can be used to generate the corresponding value of $\hat{C}_{x,t}$ and $\phi_{x,t+1}$. With this construction, we have the following result.

THEOREM 5.2 ASYMPTOTIC OPTIMALITY. *With Ch2, Ch4, R1, R2, R4, A2, A3, and A4, the scheme Φ_t described in **Algorithm 1** achieves the log t lower bound (5.1), and is thus asymptotically optimal.*

The intuition behind **THEOREM 5.2** is as follows. This situation is like having several related bandit machines, whose reward distributions are all determined by the common configuration pair (θ_1, θ_2) . The information obtained from one machine is thus 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 pull the seemingly better arm most of the time, and postpone our forced sampling (learning) to the most informative machine $X_t = x^*$. With the assumption of the existence of the value of the game, this composite Φ_t achieves the new constant in the log t lower bound.

A detailed analysis is provided in **APPENDIX C**.

6. Mixed Case In Sections 4 and 5, we dealt with the cases in which the distribution of X_t remains constant with respect to C_0 . The main difference between Sections 4 and 5 is that in one case, X_t *always* changes the preference order, in the

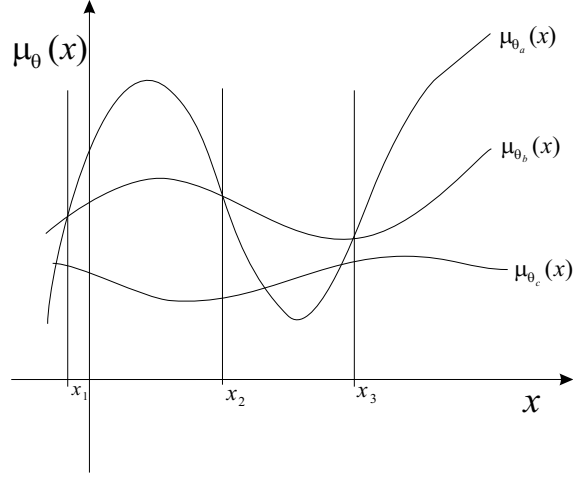


FIG. 3. For some (θ_1, θ_2) the best arm is a function of x , i.e. $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ intersect each other as in Section 4. For the remaining (θ_1, θ_2) the best arm is not a function of x , i.e. $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ do not intersect each other as first described in Section 5.

other, X_t never changes the order. A much more general case is a mixture of these two cases. In this section, we consider this mixed case, which leads to the main result of this paper.

6.1. *Formulation* Following Section 4, we assume Ch2: $G_{C_0} = G$, the distribution of $\{X_\tau\}$ does not vary with respect to C_0 . The distinction of the setting in this section is formally stated as follows.

- Best arm as a function of X_t (Ch5): As illustrated in FIG. 3, for some $C \in \Theta^2$, $M_C(x)$ is not a function of x , i.e., $M_C(x) := M_C$. 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*.

Example:

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

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 4 and 5, we would like to find a scheme that has bounded expected inferior sampling time when being applied to an unknown but *implicitly revealing* C_0 , and achieves the $\log t$ lower bound when the unknown C_0 is not *implicitly revealing*. Within the same two regularity conditions R1 and R2 as in Sections 4 and 5:

- R1: \mathbf{X} is finite, and

- R2: $\forall \theta_1, \theta_2, x$, the K-L information number $I(\theta_1, \theta_2|x)$ is finite and strictly positive,

we can achieve this goal.

6.2. *Lower Bound* Similar to THEOREM 5.1, we obtain a $\log t$ lower bound for the $T_{inf}(t)$ for uniformly good rules. However, under the more general framework in this section, we may have to perform more forced sampling for distinguishing a C_0 that is not implicitly revealing, i.e., $\forall x, M_{C_0}(x) = 2$, from a $C' = (\theta, \theta_2)$ such that $\exists x, M_{C'}(x) = 1$. The resulting $\log t$ lower bound is formally stated as follows.

THEOREM 6.1. *Under Ch2, Ch5, R1, and R2, for any uniformly good rule, if the authentic parameter pair C_0 is not implicitly revealing and the side observation sequence $\{X_\tau\}$ is evenly distributed in L^1 , then $T_{inf}(t)$ satisfies*

$$(6.1) \quad \lim_{t \rightarrow \infty} \mathbb{P}_{C_0} \left(T_{inf}(t) \geq \frac{(1 - \epsilon) \log t}{K_{C_0}} \right) = 1,$$

and $\liminf_{t \rightarrow \infty} \frac{E_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}},$

where $\epsilon > 0$ is an arbitrary constant and K_{C_0} is a constant depending on C_0 . If $M_{C_0} = 2$, $T_{inf}(t) = T_1(t)$. The constant K_{C_0} is different from that of THEOREM 5.1 and 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} in the case that $M_{C_0} = 1$ can be obtained by symmetry.

In [WKP04], it has been proved that THEOREM 6.1 holds for i.i.d. $\{X_\tau\}$. The only property of the i.i.d. $\{X_\tau\}$ being used in the proof there is that for any $A_X \subseteq \mathbf{X}$, there exists a $\pi > 0$ such that

$$\mathbb{E} \left\{ \sum_{\tau=1}^t 1\{X_\tau \in A_X\} \right\} \geq \pi t.$$

As a result, for any $\{X_\tau\}$ evenly distributed in L^1 , THEOREM 6.1 still holds and follows the same proof.

6.3. *Scheme Achieving the Lower Bound* To have a tractable scheme having bounded expected inferior sampling time when being applied to *implicitly revealing* C_0 or otherwise achieving the $\log t$ lower bound in THEOREM 6.1, we need the following assumptions similar to those in Section 5.

- A2: The parameter space Θ is finite.
- A3: The side observations $\{X_\tau\}$ are u.s.e. distributed in L^1 .

Algorithm 2 Φ_{t+1} , the decision at time $t + 1$

```

1: if not all  $\hat{C}_{x,t}$  are identical, then
2:    $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$ .
3: else
4:   Denote  $\hat{C}_t = (\hat{\theta}_1, \hat{\theta}_2)$  as the common estimate for all  $\mathbf{B}_x$ .
5:   if  $\hat{C}_t$  is implicitly revealing, then
6:     if  $\check{C}_t \neq \hat{C}_t$  (including the cases that  $\check{C}_t$  does not exist), then
7:       if  $\text{ctr}(X_{t+1}, \hat{C}_t, \check{C}_t)$  is even, then
8:          $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$ .
9:       else
10:         $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ .
11:      end if
12:       $\text{ctr}(X_{t+1}, \hat{C}_t, \check{C}_t) \leftarrow \text{ctr}(X_{t+1}, \hat{C}_t, \check{C}_t) + 1$ .
13:    else
14:       $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ .
15:    end if
16:  else
17:    Without loss of generality, we may assume  $M_{\hat{C}_t} = 2$ . The case that  $M_{\hat{C}_t} = 1$ 
    can be obtained by symmetry.
18:    if  $X_{t+1} \neq x^* := \arg \max_x \inf_{\{\theta: \mu_\theta(x) > \mu_{\hat{\theta}_2}(x)\}} I(\hat{\theta}_1, \theta|x)$ , then
19:       $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ .
20:    else
21:       $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$ .
22:    end if
23:  end if
24: end if

```

- A5: The value of the game,

$$\inf_{\{\theta: \exists x_0, \mu_\theta(x_0) > \mu_{\hat{\theta}_2}(x_0)\}} \sup_{x \in \mathbf{X}} \{I(\theta_1, \theta|x)\} = \sup_{x \in \mathbf{X}} \inf_{\{\theta: \mu_\theta(x) > \mu_{\hat{\theta}_2}(x)\}} \{I(\theta_1, \theta|x)\},$$

exists.

An adaptive control scheme Φ_t is constructed in **Algorithm 2** with the help of the *tight* decision rules ϕ_t for the traditional bandits. Some notation of Section 5.3 will be reused here, and we also define a number of counters (actually $|X| \cdot |\Theta|^2$ counters), which are named $\text{ctr}(x, C', C'')$ and initially set to zero.

In **Algorithm 2**, $\hat{C}_{x,t}$ and $\phi_{x,t+1}$ come from applying the adaptive scheme for traditional bandit problems to each sub-bandit \mathbf{B}_x , and are generated only from those samples when $\Phi_{t+1} \leftarrow \phi_{x,t+1}$ is active, namely, when lines 2, 8, and 21 are executed.

The \check{C}_t used in **Algorithm 2** is an estimate of C_0 generated from the time instants when $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ is active, namely, when lines 10, 14, and 19 are executed. And we will prove that any “good” \check{C}_t will result in a bound-achieving Φ_t . Formally speaking, we need the following definition.

DEFINITION 6.1 GOOD ESTIMATE \check{C}_t . \check{C}_t is a good estimate if $\forall i = 1, 2, \exists a_i, b_i > 0$ such that $\mathbb{P}_{C_0}(i(\hat{C}_t) \neq i(C_0)) \leq a_i \exp(-b_i N_i)$, where N_i denotes the number of samples of arm i , based on which we generate \check{C}_t .

By the large deviation principle and the regularity condition R2, a good estimate \check{C}_t always exists.

The following example demonstrates how to generate $\hat{C}_{x,t}$ and \check{C}_t respectively.

If $X_1 X_2 X_3 X_4 \dots = x_a x_b x_a x_c \dots$, consider the case in which at time instants 1 and 2, lines 2, 8, and 21 are executed, and at time instants 3 and 4, lines 10, 14, and 19 are executed. So when $t = 4$, we have 1 sample in B_{x_a} to generate $\hat{C}_{x_a,4}$, 1 sample in B_{x_b} for $\hat{C}_{x_b,4}$, and 0 sample in B_{x_c} for $\hat{C}_{x_c,4}$. At the same time, we only have 1 sample in B_{x_a} , 0 sample in B_{x_b} and 1 sample in B_{x_c} , all of which will be combined to generate \check{C}_4 .

THEOREM 6.2 ASYMPTOTIC OPTIMALITY. Suppose the constituent $\phi_{x,t}$ are tight, and the estimate \check{C}_t is good. With Ch2, Ch5, R1, R2, A2, A3, and A5, the Φ_t described in **Algorithm 2** either has bounded inferior sampling time, or achieves the $\log t$ lower bound, depending on whether or not the underlying configuration pair C_0 is implicitly revealing.

A detailed proof is given in APPENDIX D.

7. A Simple Necessary Condition In Sections 3 through 6, we have discussed the benefits of having side observations under various situations. The main results are summarized in TABLE 2. Since all of the given conditions are sufficient, the question naturally arises as to what conditions are necessary for achieving these performance improvements.

From THEOREM 3.1 of Section 3, having estimates of C_0 from $\{X_\tau\}$ with appropriate convergence speed provides an upper bound on the attainable expected inferior sampling time, which can help surpassing the $\log t$ lower bound in [LR85]. However, even without a good estimate, if all possible C_0 are *implicitly revealing*, we are still able to obtain $\lim \mathbf{E}\{T_{inf}(t)\} < \infty$ as described in Section 4. Hence, having estimates with appropriate convergence speed is not a necessary condition to surpass the traditional $\log t$ lower bound.

For the case in which $\{X_\tau\}$ reveals no information about C_0 , Sections 4 to 6 concentrate on performing both learning and control on Y_t^i with the help of X_t . We provide a simple necessary condition showing that for a side observation sequence to be beneficial, the even distribution of all possible values of $\{X_\tau\}$ is necessary .

THEOREM 7.1 COMMON NECESSARY CONDITION. For the achievability results in THEOREMS 4.1, 5.2, AND 6.2 to hold for all distribution families $\{F_\theta(\cdot|x)\}$ satisfying the characterization and regularity conditions, we must have

$$\forall x, \mathbb{P}(\exists \tau, \text{ s.t. } X_\tau = x) > 0.$$

TABLE 2
Summary of the Relationships between X_t and Y_t^i .

Characterization	Regularity Cond.	Even Distr. Cond.	Results for all $C_0 \in \Theta^2$
$\forall C_1 \neq C_2,$ $G_{C_1} \neq G_{C_2}$	As $\hat{C}_t \rightarrow C_0, \forall x,$ $M_{\hat{C}_t}(x) = M_{C_0}(x).$		$\exists \{\phi_\tau\}$ such that $\lim \frac{\mathbb{E}_{C_0}\{T_{inf}(t)\}}{\sum \mathbb{P}(\hat{C}_\tau - C_0 > \epsilon)} \leq 1.$
All $C_0 \in \Theta^2$ have $G_{C_0} = G,$ and are <i>implicitly revealing.</i>	(i) \mathbf{X} is finite, (ii) $\forall \theta_1 \neq \theta_2, x,$ $I(\theta_1, \theta_2 x) > 0,$ (iii) $\forall x, \mu_\theta(x)$ is conti. w.r.t. $\theta.$	$\{X_\tau\}$ is evenly distr. in prob. series.	$\exists \{\phi_\tau\}$ such that $\lim \mathbb{E}_{C_0}\{T_{inf}(t)\} < \infty.$
$\forall C_0 \in \Theta^2,$ $G_{C_0} = G.$ No $C_0 \in \Theta^2$ is <i>implicitly revealing.</i>	(i) \mathbf{X} is finite, (ii) $\forall \theta_1 \neq \theta_2, x,$ $I(\theta_1, \theta_2 x) > 0.$		The performance of any uniformly good $\{\phi_\tau\}$ is lower bounded by $\lim \frac{\mathbb{E}_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}},$ where $K_{C_0} \triangleq$ $\inf_\theta \sup_x I(\theta_1, \theta x).$
	(i), (ii), and (iii) Θ is finite. (iv) The existence of the value of the game.	$\{X_\tau\}$ is u.s.e. distributed in $L^1.$	$\exists \{\phi_\tau\}$ such that $\lim \frac{\mathbb{E}_{C_0}\{T_{inf}(t)\}}{\log t} \leq \frac{1}{K_{C_0}},$ namely $\{\phi_\tau\}$ achieves the lower bound.
$\forall C_0 \in \Theta^2,$ $G_{C_0} = G.$ In $\Theta^2,$ some C_0 are <i>implicitly revealing</i> and some are not.	(i) \mathbf{X} is finite, (ii) $\forall \theta_1 \neq \theta_2, x,$ $I(\theta_1, \theta_2 x) > 0.$	$\{X_\tau\}$ is evenly distributed in $L^1.$	If C_0 is not <i>implicitly revealing,</i> the performance of any uniformly good $\{\phi_\tau\}$ is lower bounded by $\lim \frac{\mathbb{E}_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}},$ where $K_{C_0} \triangleq$ $\inf_\theta \sup_x I(\theta_1, \theta x).$
	(i), (ii), and (iii) Θ is finite. (iv) The existence of the value of the game.	$\{X_\tau\}$ is u.s.e. distributed in $L^1.$	$\exists \{\phi_\tau\}$ s.t. if C_0 is <i>implicitly revealing,</i> $\lim \mathbb{E}_{C_0}\{T_{inf}(t)\} < \infty.$ If C_0 is not <i>i.r.,</i> $\{\phi_\tau\}$ achieves the lower bound: $\lim \frac{\mathbb{E}_{C_0}\{T_{inf}(t)\}}{\log t} \leq \frac{1}{K_{C_0}}.$

Note that the condition $\forall x, \mathbb{P}(\exists \tau, \text{ s.t. } X_\tau = x) > 0$ is the weakest evenly distributedness property we have introduced.

The intuition behind THEOREM 7.1 is that if there exists some x_0 such that $\mathbb{P}(\exists \tau, \text{ s.t. } X_\tau = x_0) = 0$ (equivalently, $\mathbb{P}(\forall \tau, X_\tau \neq x_0) = 1$), the benefit of the characterization properties (helpful structure between X_t, Y_t^i) may degenerate to another case with new support $\mathbf{X}' = \mathbf{X} \setminus \{x_0\}$, which severely affects the attainable results. We demonstrate the necessary condition by providing several examples,

where we assume $F_\theta(\cdot|x) \sim \mathcal{N}(\mu_{\theta,x}, 1)$, $\Theta = \{1, 2, 3\}$ or $\Theta = \{1, 2, 3, 4\}$, $\mathbf{X} = \{1, 2, 3\}$, and $\{X_\tau\}$ is i.i.d. but may have zero probability for some x .

- *Example 10:* If

$$(\mu_{\theta,x}) = \begin{pmatrix} -1 & -2 & 1 \\ 0 & 0 & 0 \\ 1 & 2 & -1 \end{pmatrix},$$

it is obvious that this example satisfies the characterization of Section 4. However, if $\mathbb{P}(\exists \tau, \text{ s.t. } X_\tau = 3) = 0$, it degenerates to $\mathbf{X}' = \{1, 2\}$, and

$$(\mu_{\theta,x})' = \begin{pmatrix} -1 & -2 \\ 0 & 0 \\ 1 & 2 \end{pmatrix}$$

falls within the case of Section 5. According to the result in Section 5, there exists a $\log t$ lower bound for all uniformly good rules $\{\phi_\tau\}$ so that no decision rule can achieve bounded $\lim_t \mathbb{E}_{C_0} \{T_{inf}(t)\}$ for all C_0 , although all configuration pairs are implicitly revealing.

- *Example 11:* If

$$(\mu_{\theta,x}) = \begin{pmatrix} -1 & -2 & -3 \\ 0 & 0 & 0 \\ 1 & 2 & 3 \end{pmatrix},$$

and the underlying $C_0 = (\theta_1, \theta_2) = (1, 2)$, the assumptions of Section 5 are satisfied. However, if $\mathbb{P}(\exists \tau, \text{ s.t. } X_\tau = 3) = 0$, $(\mu_{\theta,x})$ degenerates to

$$(\mu_{\theta,x})' = \begin{pmatrix} -1 & -2 \\ 0 & 0 \\ 1 & 2 \end{pmatrix},$$

and by the $\log t$ lower bound in THEOREM 5.1, the achievable constant in front of $\log t$ is lower bounded by $1/(\inf_{\theta > 2} \sup_{x \leq 2} I(\theta_1, \theta|x)) = 1/2$. This in turn shows that the constant $1/(\inf_{\theta > 2} \sup_x I(\theta_1, \theta|x)) = 2/9$ specified by the non-degenerate case cannot be achieved by any uniformly good rules. The asymptotic optimality result does not hold.

- *Example 12:* If

$$(\mu_{\theta,x}) = \begin{pmatrix} -1 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 1 & 1 \\ 2 & 3 & 4 \end{pmatrix},$$

the problem falls within the setting of Section 6. Suppose the underlying $C_0 = (\theta_1, \theta_2) = (2, 3)$, which is not *implicitly revealing*. As a result, the constant in front of the $\log t$ lower bound is

$$\frac{1}{\inf_{\{\theta: \exists x_0, \mu_\theta(x_0) > \mu_{\theta_2}(x_0)\}} \sup_x I(\theta_1, \theta|x)} = \frac{2}{25}.$$

However, when $\mathbb{P}(\exists \tau, \text{ s.t. } X_\tau = 3) = 0$, $(\mu_{\theta,x})$ degenerates to

$$(\mu_{\theta,x})' = \begin{pmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 3 \end{pmatrix}.$$

The achievable constant is no longer $2/25$, but is lower bounded by $1/(\inf_{\theta > 3} \sup_x I(\theta_1, \theta|x)) = 2/9$. Thus the asymptotic optimality result, concerning C_0 that are not *implicitly revealing*, does not hold. Even for the *implicitly revealing* $C_0 = (\theta_1, \theta_2) = (1, 2)$, in the degenerate case, the expected inferior sampling time is lower bounded by $(\log t)/2$. The uneven distribution of $\{X_\tau\}$ thus spoils the possible benefits of observing the side information in advance before making decisions. The conclusions in THEOREM 6.2 do not hold in this example.

8. Conclusion It has been shown in [WKP04] that observing additional side information can definitely improve sequential decisions in bandit problems. To further explore the origins of this improvement, in this paper we have extracted basic properties of the side observation process and proved their efficacy for bandit problems. With a scheme separating the learning and control tasks by observing $\{X_\tau\}$ for learning, and playing arm $M_{\hat{C}_t}(X_t)$ for control, we have proved that the order of growth rate of the inferior sampling time is the same as the order of $\sum_{\tau=1}^t \mathbb{P}(|\hat{C}_\tau - C_0| > \epsilon)$. For many types of $\{X_\tau\}$, this summation is bounded for all C_0 , and thus we can achieve the bounded expected inferior sampling time as well.

If the side observation does not provide information about the configuration C_0 , three cases have been considered: (1) the best arm is a function of X_t , as in Section 4, (2) the best arm is not a function of X_t , as in Section 5, and (3) the mixed case as in Section 6. For any $\{X_\tau\}$, *regular/even* appearances of all $x \in \mathbf{X}$ guarantee that we can fully use the beneficial structure/relationship between the side observation $\{X_\tau\}$ and the reward process $\{Y_\tau^i\}$. It has been shown in [WKP04] that for i.i.d. $\{X_\tau\}$ and for all C_0 , (1) leads to bounded expected inferior sampling time, (2) leads to asymptotically tight $\log t$ lower bound, and (3) leads to $\log t$ lower bound for some C_0 , and bounded expected inferior sampling time for other C_0 . And in Sections 4 through 6, these results have been successfully generalized to arbitrary side observation sequences $\{X_\tau\}$ possessing different levels of “regular/even appearance” properties. Consequently, a much more general class of side observation sequences, which includes Markov chains, and arbitrary deterministic periodic sequences, has the same impact on bandit problems as those of i.i.d. sequences. The idea of using X_t as an index of sub-bandit-machines has been implemented in this paper by introducing a composite decision rule and assuming the existence of the value of a game on Kullback-Leibler divergence.

Finally, we have also provided a simple necessary condition, namely $\forall x, \mathbb{P}(\exists \tau, \text{ s.t. } X_\tau = x) > 0$, which is essential for a side observation sequence to fully exploit the inherent structure between X_t and Y_t^i .

A. Proof of THEOREM 3.1

PROOF OF THEOREM 3.1. For any underlying configuration pair $C_0 = (\theta_1, \theta_2)$, define the error set \mathbf{C}_e as follows.

$$(A.1) \quad \mathbf{C}_e := \bigcup_{x \in \mathbf{X}} \{C \in \Theta^2 : M_C(x) \neq M_{C_0}(x)\}.$$

Let $\bar{\mathbf{C}}_e$ denote the closure of \mathbf{C}_e . By condition A1, we have that C_0 is not in $\bar{\mathbf{C}}_e$ and there exists $\epsilon > 0$ such that $\bar{\mathbf{C}}_e \subseteq \{C : |C - C_0| > \epsilon\}$. For any $t \geq 1$,

$$\begin{aligned} & \mathbb{P}_{C_0}(\phi_t \neq M_{C_0}(X_t)) \\ &= \mathbb{P}_{C_0}(M_{\hat{C}_t}(X_t) \neq M_{C_0}(X_t)) \\ &\leq \mathbb{P}_{C_0}(\exists x, M_{\hat{C}_t}(x) \neq M_{C_0}(x)) \\ &= \mathbb{P}_{C_0}(\hat{C}_t \in \mathbf{C}_e) \\ &\leq \mathbb{P}_{C_0}(\hat{C}_t \in \bar{\mathbf{C}}_e) \\ &\leq \mathbb{P}_{C_0}(|\hat{C}_t - C_0| > \epsilon), \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_{C_0}\{T_{inf}(t)\} &= \sum_{\tau=1}^t \mathbb{E}_{C_0}\{1\{\phi_\tau \neq M_{C_0}(X_\tau)\}\} \\ &\leq \sum_{\tau=1}^t \mathbb{P}_{C_0}(|\hat{C}_\tau - C_0| > \epsilon). \end{aligned}$$

This completes the proof.

B. Proof of THEOREM 4.1 We define \mathbf{C}_e similarly to (A.1). The necessary [Lemma 1, p.23, [WKP04]] is quoted as follows.

LEMMA B.1 [LEMMA 1, P.23, [WKP04]]. *With the regularity conditions specified in Section 4, $\exists a_1, a_2 > 0$ such that $\mathbb{P}_{C_0}(\hat{C}_t \in \mathbf{C}_e) \leq a_1 \exp(-a_2 \min\{T_1^{x^*}(t), T_2^{x^*}(t)\})$.*

ANALYSIS OF THE SCHEME. By the definition of \mathbf{C}_e , when \hat{C}_t is not in \mathbf{C}_e , the estimate is accurate enough that the myopic decision is simply the optimal decision, namely, $\forall x, M_{\hat{C}_t}(x) = M_{C_0}(x)$. Hence we have

$$\begin{aligned} \{\phi_{t+1} \neq M_{C_0}(X_{t+1})\} &= \{\phi_{t+1} \neq M_{C_0}(X_{t+1}), \hat{C}_t \in \mathbf{C}_e\} \\ &\quad \cup \{\phi_{t+1} \neq M_{C_0}(X_{t+1}), \hat{C}_t \notin \mathbf{C}_e\} \\ &\subseteq \{\hat{C}_t \in \mathbf{C}_e\} \cup \{\phi_{t+1} \neq M_{C_0}(X_{t+1}), \hat{C}_t \notin \mathbf{C}_e\} \\ (B.1) \quad &\stackrel{\Delta}{=} A_{t+1} \cup B_{t+1}. \end{aligned}$$

By the definition of the allocation rule and induction on t , it can be shown that $\forall i \in \{1, 2\}, \forall t \geq 6, T_i(t) \geq \sqrt{t}$, so that $\min_i T_i^{x^*}(t) \geq \frac{\sqrt{t}}{|\mathbf{X}|}$. By LEMMA B.1, we have $\mathbb{P}_{C_0}(A_{t+1}) \leq a_1 \exp(-a_2 \sqrt{t}/k)$, and hence $\sum_{t+1=7}^{\infty} \mathbb{P}_{C_0}(A_{t+1}) < \infty$.

For B_{t+1} , we have

$$\begin{aligned} B_{t+1} &= \{\phi_{t+1} \neq M_{C_0}(X_{t+1}), \hat{C}_t \notin \mathbf{C}_e\} \\ &= \{\phi_{t+1} = 1 \neq M_{C_0}(X_{t+1}), \hat{C}_t \notin \mathbf{C}_e\} \cup \{\phi_{t+1} = 2 \neq M_{C_0}(X_{t+1}), \hat{C}_t \notin \mathbf{C}_e\} \\ &\triangleq B_{t+1}^1 \cup B_{t+1}^2, \end{aligned}$$

where B_{t+1}^1 and B_{t+1}^2 , correspond to $\phi_{t+1} = 1, 2$ separately. We then have

$$\begin{aligned} \text{(B.2)} \quad B_{t+1}^1 &= \left\{ \exists s \in [\sqrt{t}, t-1] \text{ s.t. } \hat{C}_s \in \mathbf{C}_e, \phi_{t+1} = 1 \neq M_{C_0}(X_{t+1}), \hat{C}_t \notin \mathbf{C}_e \right\} \\ &\quad \cup \left\{ \forall s \in [\sqrt{t}, t], \hat{C}_s \notin \mathbf{C}_e, \phi_{t+1} = 1 \neq M_{C_0}(X_{t+1}) \right\} \\ &\subseteq \left\{ \exists s \in [\sqrt{t}, t-1] \text{ s.t. } \hat{C}_s \in \mathbf{C}_e \right\} \\ &\quad \cup B^{1.1}. \end{aligned}$$

This inequality comes from modifying the first term of the union and using $B^{1.1}$ as shorthand. To further bound $B^{1.1}$, we need some new notation:

$$\begin{aligned} N_1 &:= \sum_{s \in [1, t]} 1\{M_{C_0}(X_s) = 1\} \\ N_{1 \rightarrow 2} &:= \sum_{s \in [1, t]} 1\{M_{C_0}(X_s) = 1, \phi_s = 2\} \\ \text{and } N_{2 \rightarrow 1} &:= \sum_{s \in [1, t]} 1\{M_{C_0}(X_s) = 2, \phi_s = 1\}. \end{aligned}$$

From the definition, we have $T_1(t) = N_1 - N_{1 \rightarrow 2} + N_{2 \rightarrow 1}$. Suppose $\forall s \in [\sqrt{t}, t], \hat{C}_s \notin \mathbf{C}_e$, which is the first condition of $B^{1.1}$, and we notice the following inequalities,

$$\begin{aligned} N_{1 \rightarrow 2} + N_{2 \rightarrow 1} &= \sum_{s \in [1, \sqrt{t}]} 1\{\phi_s \neq M_{C_0}(X_s)\} + \sum_{s \in [\sqrt{t}+1, t]} 1\{\phi_s \neq M_{C_0}(X_s)\} \\ &\leq \sqrt{t} + \sum_{s \in [\sqrt{t}+1, t]} 1\{\phi_s \neq M_{\hat{C}_{s-1}}(X_s)\} \\ \text{(B.3)} \quad &\leq 2\sqrt{t} + 1. \end{aligned}$$

The equality is obvious and the first inequality is true since $\forall s \in [\sqrt{t}, t], \hat{C}_s \notin \mathbf{C}_e$ and thus $M_{\hat{C}_s}(\cdot) = M_{C_0}(\cdot)$. The second inequality follows from the fact that the total number of forced samples up to time t cannot be greater than $\sqrt{t} + 1$, so the number of times $\phi_s \neq M_{\hat{C}_{s-1}}(X_s)$ is smaller than $\sqrt{t} + 1$.

If the second condition of $B^{1.1}$, $\phi_{t+1} = 1 \neq M_{\hat{C}_t}(X_{t+1})$, is satisfied, it implies that the player performs the forced sampling at time $t+1$, or equivalently $T_1(t) < \sqrt{t} + 1$.

Since $\forall i, T_i(t) \geq \sqrt{t}$, it follows that $T_1(t) = N_1 - N_{1 \rightarrow 2} + N_{2 \rightarrow 1} = \sqrt{t}$. Combining the result in (B.3), we conclude that

$$\begin{aligned} B^{1,1} &\subseteq \{N_1 \leq 3\sqrt{t} + 1\} \\ &= \left\{ \sum_{s \in [1, t]} 1\{M_{C_0}(X_s) = 1\} \leq 3\sqrt{t} + 1 \right\}. \end{aligned}$$

Let $\mathbf{X}_{C_0}^1 := \{x \in \mathbf{X} : M_{C_0}(x) = 1\}$ denote the set of the possible values of the side observation such that arm 1 is favorable. From (B.2) we have

$$\begin{aligned} \text{(B.4)} \quad &P(B_{t+1}^1) \\ &\leq \left(\sum_{s \in [\sqrt{t}, t-1]} P(\hat{C}_s \in \mathbf{C}_e) \right) + P(B^{1,1}) \\ &\leq \sum_{s \in [\sqrt{t}, t-1]} a_1 e^{-a_2 \sqrt{s}} + P\left(\frac{\sum_{s \in [1, t]} 1\{X_s \in \mathbf{X}_{C_0}^1\}}{t} \leq \frac{3\sqrt{t} + 1}{t} \right), \end{aligned}$$

where the second inequality follows from the application of LEMMA B.1 to the first term, while the second term can be restated as an equivalent form. By simple algebra, we have

$$\text{(B.5)} \quad \sum_{t+1=7}^{\infty} \sum_{s \in [\sqrt{t}, t-1]} a_1 e^{-a_2 \sqrt{s}} < \infty.$$

And by the assumption that $\{X_\tau\}$ is evenly distributed in probability series, we have

$$\text{(B.6)} \quad \sum_{t+1=7}^{\infty} P\left(\frac{\sum_{s \in [1, t]} 1\{X_s \in \mathbf{X}_{C_0}^1\}}{t} \leq \frac{3\sqrt{t} + 1}{t} \right) < \infty.$$

From (B.4), (B.5), and (B.6), we conclude

$$\sum_{t+1=7}^{\infty} P(B_{t+1}) \leq \sum_{t+1=7}^{\infty} (P(B_{t+1}^1) + P(B_{t+1}^2)) < \infty,$$

and by (B.1),

$$\lim_{t \rightarrow \infty} \mathbb{E}\{T_{inf}(t)\} \leq 6 + \sum_{t+1=7}^{\infty} (P(A_{t+1}) + P(B_{t+1})) < \infty,$$

which completes the analysis.

C. Proof of THEOREM 5.2 We need the following lemma for the later proof.

LEMMA C.1. *We consider a random process $\{X_\tau\}$ and a sequence of stopping time pairs $\{(S_j, T_j)\}$, where for all $j \in \mathbb{N}$, $S_j < T_j \leq S_{j+1}$ are stopping times taking values in \mathbb{N} . We denote*

$$\text{sum} := \sum_{j=1}^{\infty} (T_j - S_j + 1)$$

$$\text{and } U := \sup\{j \in \mathbb{N} | S_j < \infty\}.$$

If both S_j, T_j , are ∞ , we define $T_j - S_j + 1 = 0$.

Suppose for some $B < \infty$ and $K < \infty$, we have $\mathbf{E}\{U\} \leq K$, and $\forall j$, $\mathbf{E}\{T_j - S_j + 1 | S_j\} \leq B$. It follows that $\mathbf{E}\{\text{sum}\} \leq B \cdot K < \infty$.

PROOF. The proof is similar to that of Wald's Lemma. Using the convention that $0 \cdot \infty = 0$, we rewrite sum in the following form:

$$\begin{aligned} \text{sum} &= \sum_{j=1}^{\infty} 1\{S_j < \infty\} (T_j - S_j + 1) \\ \implies \mathbf{E}\{\text{sum}\} &= \sum_{j=1}^{\infty} \mathbf{E}\{1\{S_j < \infty\} \cdot \mathbf{E}\{T_j - S_j + 1 | S_j\}\} \\ &\leq \sum_{j=1}^{\infty} B \cdot \mathbf{E}\{1\{S_j < \infty\}\} \\ &= B \sum_{j=1}^{\infty} \mathbf{P}(U \geq j) = B \cdot K. \end{aligned}$$

With the help of LEMMA C.1, we prove THEOREM 5.2 by proving the following arguments.

- ARGUMENT 1: The expected duration over which \hat{C}_t does not exist is finite, i.e.,

$$\lim_{t \rightarrow \infty} \mathbf{E} \left\{ \sum_{\tau=1}^t 1\{\hat{C}_\tau \text{ does not exist}\} \right\} < \infty.$$

For simplicity, we use $1\{\hat{C}_t\} = 0$ as shorthand that \hat{C}_t does not exist.

- ARGUMENT 2: The expected duration over which $\hat{C}_t \neq C_0$ is finite, i.e.,

$$\lim_{t \rightarrow \infty} \mathbf{E} \left\{ \sum_{\tau=1}^t 1\{\hat{C}_\tau \neq C_0\} \right\} < \infty.$$

- ARGUMENT 3: The expected duration over which $\hat{C}_t = C_0$ and $\Phi_{t+1} \neq M_{C_0}(X_{t+1})$ is upper bounded by $\frac{\log t}{K_{C_0}}$, i.e.,

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E} \left\{ \sum_{\tau=1}^t 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\} \right\}}{\log t} \leq \frac{1}{K_{C_0}},$$

where $K_{C_0} = \inf_{\theta > \theta_2} \sup_x I(\theta_1, \theta|x)$ if $M_{C_0} = 2$.

PROOF OF ARGUMENT 1. To discuss stopping times, we first define the filtration \mathcal{F}_t in an explicit way, that is, \mathcal{F}_t is the σ -algebra generated by the past outcomes of the rewards $1\{\Phi_\tau = 1\}Y_\tau^1 + 1\{\Phi_\tau = 2\}Y_\tau^2$ for $\tau \in [1, t]$, and the observations X_τ for $\tau \in [1, t+1]$. For instance, by definition we have $\hat{C}_t \in \mathcal{F}_t$, $X_{t+1} \in \mathcal{F}_t$ and $\phi_{t+1} \in \mathcal{F}_t$.

For any $x \in \mathbf{X}$, we iteratively define the stopping time pairs $S_{x,j}$ and $T_{x,j}$ as follows.

$$S_{x,j} := \inf \left\{ t > S_{x,j-1} : X_t = x, 1\{\hat{C}_t\} = 0, \right. \\ \left. \text{and either } 1\{\hat{C}_{t-1}\} = 1 \text{ or } \mathbf{X} = \bigcup_{s \in (S_{x,j-1}, t)} \{X_s\} \right\},$$

and

$$T_{x,j} := \inf \left\{ t > S_{x,j} : \text{either } 1\{\hat{C}_t\} = 1 \text{ or } \mathbf{X} = \bigcup_{s \in (S_{x,j}, t]} \{X_s\} \right\},$$

where $S_{x,0} = 0$. Note that $S_{x,j}$ and $T_{x,j}$ are basically dividing the duration over which $1\{\hat{C}_t\} = 0$ into disjoint⁵ intervals, while x specifying the value of the side observation X_t at the leading time instant $S_{x,j}$. We then have

$$\sum_{\tau=1}^{\infty} 1\{1\{\hat{C}_\tau\} = 0\} \leq \sum_x \sum_{j \in \mathbb{N}} (T_{x,j} - S_{x,j} + 1).$$

Since

$$T_{x,j} \leq \inf \left\{ t > S_{x,j} : \mathbf{X} = \bigcup_{s \in (S_{x,j}, t]} \{X_s\} \right\},$$

and by the assumption that $\{X_\tau\}$ is u.s.e. distributed in L^1 , there exists $B < \infty$ such that $\forall x, j, \mathbb{E}\{T_{x,j} - S_{x,j} + 1 | S_{x,j}\} < B$. If we can show

$$(C.1) \quad \forall x, \mathbb{E}\{\sup\{j \in \mathbb{N} : S_{x,j} < \infty\}\} < \infty,$$

then by LEMMA C.1, we have $\mathbb{E}\left\{\sum_{t=1}^{\infty} 1\{1\{\hat{C}_t\} = 0\}\right\} < \infty$.

We prove Eq. (C.1) by case study. For any x, j , and time $t := S_{x,j}$, since $1\{\hat{C}_t\} = 0$ and $X_t = x$, we must have one of the following two cases.

⁵In some cases, the intervals may overlap with each other, but the overlap many only happens at the end points, which does not affect the validity of the proof.

- $\hat{C}_{x,t} \neq C_0$:
 - If $1\{\hat{C}_{t-1}\} = 0$, then $\Phi_t \leftarrow \phi_{x,t}$. By the assumption that the constituent $\phi_{x,t}$ is tight, the expected duration of the event $\{X_t = x, \Phi_t \leftarrow \phi_{x,t}, \hat{C}_{x,t} \neq C_0\}$ must be finite. So this case can only contribute finite expectation.
 - If $1\{\hat{C}_{t-1}\} = 1$, the only condition resulting in $1\{\hat{C}_t\} = 0$ is that $\hat{C}_{x,t-1}$ is destroyed after time t , which in turn implies $X_t = x$ and $\Phi_t \leftarrow \phi_{x,t}$. By the assumption of tight $\phi_{x,t}$, the expected duration of the event $\{X_t = x, \Phi_t \leftarrow \phi_{x,t}, \hat{C}_{x,t} \neq C_0\}$ must be finite. So this case can only contribute finite expectation.

- $\hat{C}_{x,t} = C_0$:
By observing $\sup\{j \in \mathbb{N} : S_{x,j} < \infty\} \leq \sup\{j \in \mathbb{N} : T_{x,j} < \infty\} + 1$, we choose to show the latter has bounded expectation.

Suppose $T_{x,j} < \infty$, and note that $1\{\hat{C}_t\} = 0$ implies there exists $x' \neq x$ such that $\hat{C}_{x',t} \neq C_0$. There are only two sub-cases as follows.

- $\exists t' \in (S_{x,j}, T_{x,j}]$ such that $X_{t'} = x'$ and $\hat{C}_{x',t'-1} \neq C_0$.
- $X_{T_{x,j}} = x$ and $\hat{C}_{x,T_{x,j}} \neq \hat{C}_{x,t} = C_0$.

The reason why there are only two sub-cases follows because if there exists no such t' , then $\hat{C}_{x',s}$ remains unchanged within the interval $(S_{x,j}, T_{x,j}]$. So the only situation in which $T_{x,j} < \infty$ is when $\hat{C}_{x,t}$ is destroyed at $T_{x,j}$. Since for all $s \in (S_{x,j}, T_{x,j}]$ the decision rule is $\Phi_s \leftarrow \phi_{X_s,s}$, we then have

$$\begin{aligned} & \sup\{j \in \mathbb{N} : T_{x,j} < \infty\} \\ & \leq \sum_{\tau=1}^{\infty} 1\{X_{\tau} = x, \Phi_{\tau} \leftarrow \phi_{x,\tau}, \hat{C}_{x,\tau} \neq C_0\} \\ & \quad + \sum_{x':x' \neq x} \sum_{\tau=1}^{\infty} 1\{X_{\tau+1} = x', \Phi_{\tau+1} \leftarrow \phi_{x',\tau+1}, \hat{C}_{x',\tau} \neq C_0\}. \end{aligned}$$

By the assumption of tight constituent $\phi_{x,t}$, the above must have finite expectation.

From the previous discussions, we have proved $\mathbf{E}\{\sup\{j \in \mathbb{N} : S_{x,j} < \infty\}\} < \infty$ and ARGUMENT 1.

PROOF OF ARGUMENT 2. Consider a fixed $C' := (\theta'_1, \theta'_2) \neq C_0$ and set $x^* := \arg \max_x \inf_{\theta: \theta > \theta'_2} I(\theta'_1, \theta|x)$. We then iteratively define the stopping time pairs $S_{C',j}$ and $T_{C',j}$ as follows.

$$\begin{aligned} S_{C',j} & := \inf \left\{ t > S_{C',j-1} : \hat{C}_t = C', \right. \\ & \quad \left. \text{and either } 1\{\hat{C}_{t-1}\} = 0, \text{ or } \hat{C}_{t-1} \neq C', \text{ or } X_t = x^* \right\}, \end{aligned}$$

and

$$T_{C',j} := \inf \left\{ t > S_{C',j} : \text{either } 1\{\hat{C}_t\} = 0, \text{ or } \hat{C}_t \neq C', \text{ or } X_t = x^* \right\},$$

where $S_{C',0} = 0$. Note that $S_{C',j}$ and $T_{C',j}$ are basically dividing the duration of the event $\{\hat{C}_t \neq C_0\}$ into disjoint intervals while C' is specifying the value of the common estimate \hat{C}_t during those intervals. Then we have

$$\sum_{t=1}^{\infty} 1\{\hat{C}_t \neq C_0\} \leq \sum_{C' \neq C_0} \sum_{j \in \mathbb{N}} (T_{C',j} - S_{C',j} + 1).$$

Since

$$T_{C',j} \leq \inf \{t > S_{C',j} : X_{t+1} = x^*\},$$

and by the assumption that $\{X_\tau\}$ is u.s.e. distributed in L^1 , there exists $B < \infty$ such that $\forall x, j, \mathbf{E} \{T_{C',j} - S_{C',j} + 1 | S_{C',j}\} < B$. If we can show

$$\forall x, \mathbf{E} \{\sup\{j \in \mathbb{N} : S_{C',j} < \infty\}\} < \infty,$$

then by LEMMA C.1, we have $\mathbf{E} \left\{ \sum_{t=1}^{\infty} 1\{\hat{C}_t \neq C_0\} \right\} < \infty$.

By observing $\sup\{j \in \mathbb{N} : S_{C',j} < \infty\} \leq \sup\{j \in \mathbb{N} : T_{C',j} < \infty\} + 1$, we choose to show the latter has bounded expectation. We first observe that there is some redundancy in the definition of $T_{C',j}$ since when \hat{C}_t exists, the only possible situation under which \hat{C}_t will change is when $X_t = x^*$. So $T_{C',j}$ can be rewritten as follows.

$$T_{C',j} := \inf \{t > S_{C',j} : X_t = x^*\}.$$

By this new definition, if $T_{C',j} < \infty$, we have $X_{T_{C',j}} = x^*$, $\hat{C}_{x^*, T_{C',j}-1} = C' \neq C_0$, and $s \in (S_{C',j}, T_{C',j}]$, $\Phi_{T_{C',j}} s \leftarrow \phi_{x^*, T_{C',j}}$. Using these facts, we have

$$\begin{aligned} & \sup\{j \in \mathbb{N} : T_{C',j} < \infty\} \\ & \leq \sum_{t=1}^{\infty} 1\{X_{t+1} = x^*, \Phi_{t+1} \leftarrow \phi_{x^*, t+1}, \hat{C}_{x^*, t} \neq C_0\}. \end{aligned}$$

And by the assumption of tight constituent $\phi_{x,t}$, the above has finite expectation and we have proved ARGUMENT 2.

PROOF OF ARGUMENT 3. Suppose $C_0 = (\theta_1, \theta_2)$. Without loss of generality, we may assume $M_{C_0} = 2$ and let $x^* = \arg \max_x \inf_{\theta: \theta > \theta_2} I(\theta_1, \theta | x)$. We then have

$$\begin{aligned} & \sum_{\tau=1}^t 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\} \\ & = \sum_{\tau=1}^t 1\{\hat{C}_\tau = \hat{C}_{x^*, \tau} = C_0, X_{\tau+1} = x^*, \Phi_{\tau+1} \leftarrow \phi_{x^*, \tau+1} \neq M_{C_0}(X_{\tau+1})\} \\ & \leq \sum_{\tau=1}^t 1\{\hat{C}_{x^*, \tau} = C_0, X_{\tau+1} = x^*, \Phi_{\tau+1} \leftarrow \phi_{x^*, \tau+1} \neq M_{C_0}(X_{\tau+1})\}. \end{aligned}$$

By the assumptions of tight constituent $\phi_{x,t}$ and the existence of the value of the game, we have

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E} \left\{ \sum_{\tau=1}^t 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\} \right\}}{\log t} \leq \frac{1}{K_{C_0}},$$

where

$$K_{C_0} = \inf_{\theta > \theta_2} I(\theta_1, \theta | x^*) = \inf_{\theta > \theta_2} \sup_x I(\theta_1, \theta | x).$$

The proof of ARGUMENT 3, and thus that of Theorem 5.2, is complete.

D. Proof of THEOREM 6.2 With the help of LEMMA C.1, we prove THEOREM 6.2 by proving the following arguments.

- ARGUMENT 1: The expected duration over which \hat{C}_t does not exist is finite, namely,

$$\lim_{t \rightarrow \infty} \mathbb{E} \left\{ \sum_{\tau=1}^t 1\{\hat{C}_\tau \text{ does not exist}\} \right\} < \infty.$$

Again we use $1\{\hat{C}_t\} = 0$ as shorthand for the situation in which \hat{C}_t does not exist.

- ARGUMENT 2: The expected duration over which $\hat{C}_t \neq C_0$ is finite, namely,

$$\lim_{t \rightarrow \infty} \mathbb{E} \left\{ \sum_{\tau=1}^t 1\{\hat{C}_\tau \neq C_0\} \right\} < \infty.$$

- ARGUMENT 3: If C_0 is *implicitly revealing*, the expected duration over which $\hat{C}_t = C_0$ and $\Phi_{t+1} \neq M_{C_0}(X_{t+1})$ is finite, namely,

$$\lim_{t \rightarrow \infty} \mathbb{E} \left\{ \sum_{\tau=1}^t 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\} \right\} \leq \infty.$$

- ARGUMENT 4: If C_0 is not *implicitly revealing*, the expected duration over which $\hat{C}_t = C_0$ and $\Phi_{t+1} \neq M_{C_0}(X_{t+1})$ is upper bounded by $\frac{\log t}{K_{C_0}}$, namely,

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E} \left\{ \sum_{\tau=1}^t 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\} \right\}}{\log t} \leq \frac{1}{K_{C_0}},$$

where $K_{C_0} = \inf_{\{\theta: \exists x_0, \mu_\theta(x_0) > \mu_{\theta_2}(x_0)\}} \sup_x I(\theta_1, \theta | x)$ if $M_{C_0} = 2$.

With the above four arguments, it is straightforward to show that the Φ_t described in **Algorithm 2** satisfies the statements in THEOREM 6.2.

PROOF OF ARGUMENT 1. This proof follows word by word the proof of ARGUMENT 1 in APPENDIX C.

PROOF OF ARGUMENT 2. Since

$$\sum_{t=1}^{\infty} 1\{\hat{C}_t \neq C_0\} = \sum_{C' \neq C_0} \sum_{t=1}^{\infty} 1\{\hat{C}_t = C' \neq C_0\},$$

we would like to prove that for any $C' \neq C_0$, $\sum_{t=1}^{\infty} 1\{\hat{C}_t = C' \neq C_0\}$ has finite expectation. For those C' that are not *implicitly revealing*, the proof follows word by word the proof of ARGUMENT 2 in APPENDIX C.

So we may assume that C' is *implicitly revealing*, and by conditioning on whether or not $\hat{C}_t = \check{C}_t$, we have

$$\begin{aligned} \sum_{t=1}^{\infty} 1\{\hat{C}_t = C' \neq C_0\} &= \sum_{t=1}^{\infty} 1\{\hat{C}_t = C' \neq C_0, \check{C}_t \neq \hat{C}_t\} \\ &\quad + \sum_{t=1}^{\infty} 1\{\hat{C}_t = C' \neq C_0, \check{C}_t = \hat{C}_t\}. \end{aligned}$$

These two summations will be considered separately. Note that when considering the estimate $\hat{C}_t \neq C'$, there are always the situations in which an estimate \hat{C}_t does not exist or the case in which \hat{C}_t exists but does not equal C' . In the following proof, $\{\hat{C}_t \neq C'\}$ is used as shorthand for both of these situations.

Let $C'' \neq C'$ denote another *implicitly revealing* parameter pair, and construct the stopping time pairs $S_{x,C',C'',j}$ and $T_{x,C',C'',j}$ iteratively as follows.

$$\begin{aligned} S_{x,C',C'',j} &:= \inf \left\{ t > S_{x,C',C'',j-1} : X_{t+1} = x, \hat{C}_t = C', \check{C}_t = C'', \right. \\ &\quad \left. \text{and either } \hat{C}_{t-1} \neq C' \text{ or } \check{C}_{t-1} \neq C'' \text{ or } X_t \neq x \right\}, \end{aligned}$$

and

$$T_{x,C',C'',j} := \inf \left\{ t > S_{x,C',C'',j} : \text{either } \hat{C}_t \neq C', \text{ or } \check{C}_t \neq C'', \text{ or } X_{t+1} = x \right\},$$

where $S_{x,C',C'',0} = 0$. Note that $S_{x,C',C'',j}$ and $T_{x,C',C'',j}$ are basically dividing the duration over which $\{\hat{C}_t = C', \check{C}_t = C''\}$ into disjoint intervals when x specifies the value of the side observation X_{t+1} at the leading time instant of those intervals. Thus we have

$$\begin{aligned} \sum_{t=1}^{\infty} 1\{\hat{C}_t = C' \neq C_0, \check{C}_t \neq \hat{C}_t\} &= \sum_{C''} \sum_{t=1}^{\infty} 1\{\hat{C}_t = C', \check{C}_t = C''\} \\ &\leq \sum_{x,C''} \sum_{j \in \mathbb{N}} (T_{x,C',C'',j} - S_{x,C',C'',j} + 1). \end{aligned}$$

Since

$$T_{x,C',C'',j} \leq \inf \{t > S_{x,C',C'',j} : X_{t+1} = x\},$$

and by the assumption that $\{X_\tau\}$ is u.s.e. distributed in L^1 , there exists a $B < \infty$ such that $\forall x, j, \mathbf{E} \{T_{x,C',C'',j} - S_{x,C',C'',j} + 1 | S_{x,C',C'',j}\} < B$. It we can show that

$$\forall x, C'', \exists K, \mathbf{E} \{\sup\{j \in \mathbb{N} : S_{x,C',C'',j}\}\} < K,$$

and thus by LEMMA C.1, we have $\mathbb{E} \left\{ \sum_{t=1}^{\infty} 1\{\hat{C}_t = C' \neq C_0, \ddot{C}_t \neq \hat{C}_t\} \right\} < \infty$.

Let $t := S_{x,C',C'',j}$. By the definition of **Algorithm 2**, for odd j the decision rule results in $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t}$ (since at time t , $\text{ctr}(x, C', C'') = j - 1$). Thus we have

$$\begin{aligned} \sup \{j \in \mathbb{N} : S_{x,C',C'',j} < \infty\} &= \sum_{j=1}^{\infty} 1\{S_{x,C',C'',j} < \infty\} \\ &\leq 2 \sum_{\tau=1}^{\infty} 1\{X_{t+1} = x, \hat{C}_t = C' \neq C_0, \ddot{C}_t = C'', \Phi_{t+1} \leftarrow \phi_{x,t+1}\}. \end{aligned}$$

By the assumption of tight $\phi_{x,t}$, the above right-hand side has finite expectation.

For the case in which $\hat{C}_t = \ddot{C}_t = C' \neq C_0$, we construct the stopping time pairs as follows.

$$S_{x,C',j} := \inf \left\{ t > S_{x,C',j-1} : X_{t+1} = x, \hat{C}_t = \ddot{C}_t = C', \right.$$

$$\left. \text{and either } \hat{C}_{t-1} \neq C' \text{ or } \ddot{C}_{t-1} \neq C' \text{ or } \{1, 2\} = \bigcup_{s \in (S_{x,C',j-1}, t]} \{M_{C'}(X_s)\} \right\},$$

and

$$T_{x,C',j} := \inf \left\{ t > S_{x,C',j} : \text{either } \hat{C}_t \neq C', \text{ or } \ddot{C}_t \neq C', \text{ or } \{1, 2\} = \bigcup_{s \in (S_{x,C',j}, t]} \{M_{C'}(X_s)\} \right\},$$

where $S_{x,C',0} = 0$. We then have

$$\sum_{t=1}^{\infty} 1\{\hat{C}_t = \ddot{C}_t = C' \neq C_0\} \leq \sum_{x \in \mathbf{X}} \sum_{j \in \mathbb{N}} (T_{x,C',j} - S_{x,C',j} + 1).$$

Since

$$T_{x,C',j} \leq \inf \left\{ t > S_{x,C',j} : \mathbf{X} = \bigcup_{s \in (S_{x,C',j}, t]} \{X_s\} \right\},$$

and by the assumption that $\{X_\tau\}$ is u.s.e. distributed in L^1 , there exists a $B < \infty$ such that $\forall x, C', j, \mathbb{E} \{T_{x,C',j} - S_{x,C',j} + 1 | S_{x,C',j}\} \leq B$. If we can show

$$(D.1) \quad \forall x \in \mathbf{X}, \mathbb{E} \{ \sup \{j \in \mathbb{N} : S_{x,C',j} < \infty\} \} < \infty,$$

then by LEMMA C.1, we have $\mathbb{E} \left\{ \sum_{t=1}^{\infty} 1\{\hat{C}_t = \ddot{C}_t = C' \neq C_0\} \right\} < \infty$.

We prove Eq. (D.1) by case study. Without loss of generality, we may assume $M_{C'}(x) = 1$, for any fixed x and C' . Consider the cases as follows.

- $1(C') \neq 1(C_0)$: Since after $t = S_{x,C',j}$, $\Phi_{t+1} = M_{\hat{C}_t}(X_{t+1}) = M_{C'}(x) = 1$, we then have

$$\sup \{j \in \mathbb{N} : S_{x,C',j} < \infty\} = \sum_{j=1}^{\infty} 1\{S_{x,C',j} < \infty\}$$

$$\leq \sum_{\tau=1}^{\infty} 1\{X_{\tau+1} = x, 1(\hat{C}_\tau) = 1(C') \neq 1(C_0), \Phi_{\tau+1} \leftarrow M_{C'}(x) = 1\} \triangleq D_1.$$

Since every time the event $\{X_{t+1} = x, 1(\hat{C}_t) = 1(C') \neq 1(C_0), \Phi_{t+1} \leftarrow M_{C'}(x) = 1\}$ occurs, the effective sample size of arm 1 (used to generate \hat{C}_t) increases by one. Since \hat{C}_t is a *good* estimate, the expectation of D_1 must be bounded. Thus, this case can at most contribute finite expectation.

- $1(C') = 1(C_0)$: This implies that $2(C') \neq 2(C_0)$. By noting that $\sup\{j \in \mathbb{N} : S_{x, C', j} < \infty\} \leq \sup\{j \in \mathbb{N} : T_{x, C', j} < \infty\} + 1$, we prove that the latter can have at most finite expectation. If $T_{x, C', j} < \infty$, it follows that it is either $M_{C'}(X_{T_{x, C', j}}) = 2$ or $1(\hat{C}_{T_{x, C', j}}) \neq 1(C_0)$. As a result, we have

$$\begin{aligned} & \sup\{j \in \mathbb{N} : T_{x, C', j} < \infty\} = \sum_{j=1}^{\infty} 1\{T_{x, C', j} < \infty\} \\ & \leq \sum_{\tau=1}^{\infty} 1\{X_\tau = x, 1(\hat{C}_\tau) \neq 1(C_0), \hat{C}_\tau = C', \Phi_\tau \leftarrow M_{C'}(x) = 1\} \\ & \quad + \sum_{x': M_{C'}(x')=2} \sum_{\tau=1}^{\infty} 1\{X_{\tau+1} = x', 2(\hat{C}_\tau) \neq 2(C_0), \hat{C}_\tau = C', \Phi_{\tau+1} \leftarrow M_{C'}(x) = 2\}. \end{aligned}$$

Since the estimate \hat{C}_t is *good*, each infinite sum in the above equation has finite expectation. Thus we have proved that this case can contribute at most finite expectation.

From our treatment of the three cases: \hat{C}_t is not *implicitly revealing*, \hat{C}_t is *implicitly revealing* but $\hat{C}_t \neq \check{C}_t$, and $\hat{C}_t = \check{C}_t$ is *implicitly revealing*, the proof of ARGUMENT 2 is complete.

PROOF OF ARGUMENT 3. When $\hat{C}_t = C_0$, the only situation of sampling the inferior arm is $\check{C}_t \neq \hat{C}_t = C_0$. For any fixed $C' \neq C_0$, construct the stopping time pairs as follows.

$$\begin{aligned} S_{C', j} &:= \inf \left\{ t > S_{C', j-1} : \hat{C}_t = C_0, \check{C}_t = C', \right. \\ & \quad \text{and either } \hat{C}_{t-1} \neq C_0 \text{ or } \check{C}_{t-1} \neq C', \\ & \quad \left. \text{or } \{1, 2\} = \bigcup_{s \in \mathbf{S}_{j-1, t-1}} \{M_{C_0}(X_s)\} \right\}, \end{aligned}$$

where $S_{C', 0} = 0$ and

$$\mathbf{S}_{j-1, t-1} := \{s \in (S_{C', j-1}, t-1] : \text{the line } \Phi_s \leftarrow M_{C_0}(X_s) \text{ is active}\}.$$

For $T_{C', j}$, we have

$$T_{C', j} := \inf \left\{ t > S_{C', j} : \text{either } \hat{C}_t \neq C_0, \text{ or } \check{C}_t \neq C', \text{ or } \{1, 2\} = \bigcup_{s \in \mathbf{S}_{j, t}} \{M_{C_0}(X_s)\} \right\}.$$

Since $S_{C',j}$ and $T_{C',j}$ partition the duration over which $\{\hat{C}_t = C_0, \check{C}_t = C'\}$ into disjoint intervals, we then have

$$\begin{aligned} \sum_{t=1}^{\infty} 1\{\hat{C}_t = C_0 \neq \check{C}_t\} &\leq \sum_{C' \neq C_0} \sum_{t=1}^{\infty} 1\{\hat{C}_t = C_0, \check{C}_t = C'\} \\ &\leq \sum_{C' \neq C_0} \sum_{j \in \mathbb{N}} (T_{C',j} - S_{C',j} + 1). \end{aligned}$$

By line 7 in **Algorithm 2**, for any $X_{t+1} = x$, $\hat{C}_t = C_0$, $\check{C}_t = C'$, the decision rule Φ_{t+1} is alternating between $\phi_{x,t}$ and $M_{C_0}(x)$. As a result, we have

$$T_{C',j} \leq \inf\{t > S_{C',j} : \forall x \in \mathbf{X}, \exists s_1 \neq s_2 \in (S_{C',j}, t] \text{ s.t. } X_{s_1} = X_{s_2} = x\}.$$

By the assumption that $\{X_\tau\}$ is u.s.e. distributed in L^1 , there exists a $B < \infty$ such that $\forall C', j, \mathbf{E}\{T_{C',j} - S_{C',j} + 1 | S_{C',j}\} \leq B$. If we can show

$$\forall x \in \mathbf{X}, C', \mathbf{E}\{\sup\{j \in \mathbb{N} : S_{C',j} < \infty\}\} < \infty,$$

then by LEMMA C.1, we have $\mathbf{E}\left\{\sum_{t=1}^{\infty} 1\{\hat{C}_t = C_0 \neq \check{C}_t\}\right\} < \infty$

Since $\sup\{j : S_{C',j} < \infty\} \leq \sup\{j : T_{C',j} < \infty\} + 1$, equivalently, we can focus on proving $\mathbf{E}\{\sup\{j \in \mathbb{N} : T_{C',j} < \infty\}\} < \infty$. For any $j \in \mathbb{N}$, let $t' := T_{C',j} < \infty$. Then one of the following situations must be true.

- $\Phi_{t'} \leftarrow \phi_{X_{t'}, t'}$: The only situation under which we can have $\Phi_{t'} \leftarrow \phi_{X_{t'}, t'}$ is $\hat{C}_t \neq C_0$. Since the constituent $\phi_{x,t}$ is tight, this part contributes at most bounded expectation.
- $\Phi_{t'} \leftarrow M_{C_0}(X_{t'})$: There are two ways in which the interval will end in this situation:
 - $\{1, 2\} = \bigcup_{s \in \mathbf{S}_{j,t'}} \{M_{C_0}(X_s)\}$: In this case, both the samples of arm 1 and arm 2 used by \check{C}_t must have increased by 1. Since \check{C}_t is a good estimate, this portion contributes at most bounded expectation.
 - $\check{C}_{t'} \neq C'$: Without loss of generality, we may assume $\{1\} = \bigcup_{s \in \mathbf{S}_{j,t'}} \{M_{C_0}(X_s)\}$. Two sub-cases are as follows:
 - * $1(C') \neq 1(C_0)$: Since $\mathbf{S}_{j,t'}$ is not empty, there exists an s such that $\Phi_s \leftarrow M_{C_0}(X_s) = 1$. Thus the number of samples from arm 1 used to generate \check{C}_t must increase by 1 during the interval $[S_{C',j}, T_{C',j}]$. By the assumption that \check{C}_t is good, that portion contributes at most finite expectation.
 - * $1(C') = 1(C_0)$: First we observe that in this case, $t' \in \mathbf{S}_{j,t'}$, which implies $\Phi_{t'} \leftarrow M_{C_0}(X_{t'})$. For each j , the number of samples of arm 1 (used by \check{C}_t) increases by at least one. We also note that $\check{C}_{t'} \neq \check{C}_{t'-1} = C'$ and $1(\check{C}_{t'}) \neq 1(C_0)$. Combining the above observations and the assumption that \check{C}_t is good, this portion can contribute at most bounded expectation.

From the above discussions, we have

$$\mathbb{E} \left\{ \sum_{t=1}^{\infty} 1\{\hat{C}_t = C_0 \neq \check{C}_t\} \right\} < \infty.$$

PROOF OF ARGUMENT 4. Suppose $C_0 = (\theta_1, \theta_2)$, $M_{C_0} = 2$ and let $x^* = \arg \max_x \inf_{\{\theta: \mu_\theta(x) > \mu_{\theta_2}(x)\}} I(\theta_1, \theta|x)$. We then have

$$\begin{aligned} & \sum_{\tau=1}^t 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\} \\ &= \sum_{\tau=1}^t 1\{\hat{C}_\tau = \hat{C}_{x^*, \tau} = C_0, X_{\tau+1} = x^*, \Phi_{\tau+1} \leftarrow \phi_{x^*, \tau+1} \neq M_{C_0}(X_{\tau+1})\} \\ &\leq \sum_{\tau=1}^t 1\{\hat{C}_{x^*, \tau} = C_0, X_{\tau+1} = x^*, \Phi_{\tau+1} \leftarrow \phi_{x^*, \tau+1} \neq M_{C_0}(X_{\tau+1})\}. \end{aligned}$$

By the assumptions of tight constituent $\phi_{x,t}$ and existence of the value of the game, we have

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E} \left\{ \sum_{\tau=1}^t 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\} \right\}}{\log t} \leq \frac{1}{K_{C_0}},$$

where

$$\begin{aligned} K_{C_0} &= \inf_{\{\theta: \mu_\theta(x^*) > \mu_{\theta_2}(x^*)\}} I(\theta_1, \theta|x^*) \\ &= \inf_{\{\theta: \exists x_0, \mu_\theta(x_0) > \mu_{\theta_2}(x_0)\}} \sup_x I(\theta_1, \theta|x). \end{aligned}$$

The proof of ARGUMENT 4 is then complete.

E. Relationships Among Evenly Distribution Properties Let $A_{u.s.e.}$ denote the family of $\{X_\tau\}$ that are u.s.e. distributed in L^1 , and $A_{p.s.}$ and A_{L^1} denote the corresponding families when $\{X_\tau\}$ is evenly distributed in probability series and evenly distributed in L^1 , respectively. We have the following proposition.

PROPOSITION E.1. *Both $A_{u.s.e.}$ and $A_{p.s.}$ are proper subsets of A_{L^1} .*

PROOF. First we prove that $A_{p.s.}$ is a subset of A_{L^1} . For any $\{X_\tau\} \in A_{p.s.}$, it follows that $\exists \pi(\cdot) > 0$ such that $\forall x$,

$$\begin{aligned} & \lim_{t \rightarrow \infty} \mathbb{P}(f_r(x, t) < \pi(x)) = 0 \\ \iff & \lim_{t \rightarrow \infty} \mathbb{P}(f_r(x, t) \geq \pi(x)) = 1. \end{aligned}$$

And since

$$\mathbb{E}\{f_r(x, t)\} \geq \pi(x) \cdot \mathbb{P}(f_r(x, t) \geq \pi(x)),$$

we have

$$\liminf_{t \rightarrow \infty} \mathbf{E}\{f_r(x, t)\} \geq \pi(x),$$

and thus $\{X_\tau\} \in A_{L^1}$.

Since $\{X_\tau\} \in A_{u.s.e.}$, $\exists B < \infty$ such that $\forall x, T, \mathbf{E}\{H_T(x)|T\} \leq B$. By setting the stopping time $T = 2(n-1)B$, we have $\mathbf{E}\{H_{2(n-1)B}(x)\} \leq B$. Thus

$$\begin{aligned} B &\geq \mathbf{E}\{H_{2(n-1)B}(x)\} \\ &\geq 2B \cdot \mathbf{P}(H_{2(n-1)B}(x) > 2B). \end{aligned}$$

As a result, $1/2 \leq \mathbf{P}(H_{2(n-1)B}(x) \leq 2B) = \mathbf{P}(\exists \tau \in (2(n-1)B, 2nB], \text{ s.t. } X_\tau = x)$, and

$$\mathbf{E}\{f_r(x, (2(n-1)B, 2nB])\} \geq \frac{1}{2} \cdot \frac{1}{2B}.$$

Since the above is true for all intervals $\{(2(n-1)B, 2nB]\}_{n \in \mathbb{N}}$, we have

$$\liminf_{t \rightarrow \infty} \mathbf{E}\{f_r(x, t)\} \geq \frac{1}{4B}.$$

To prove that both $A_{p.s.}$ and $A_{u.s.e.}$ are *proper* subsets of A_{L^1} , we will construct a specific $\{X_\tau\}$ that is in A_{L^1} but not in $A_{p.s.} \cup A_{u.s.e.}$. Suppose $\mathbf{X} = \{0, 1\}$ and consider the following two deterministic sequences \mathbf{u} and \mathbf{v} such that $\forall \tau \in \mathbb{N}$, $u_\tau = \tau \bmod 2$ and $v_\tau = 0$. Let $\{X_\tau\}$ equal \mathbf{u} and \mathbf{v} , each with probability $1/2$. It is easy to verify that $X \in A_{L^1}$ but not in $A_{p.s.} \cup A_{u.s.e.}$.

PROPOSITION E.2. $A_{p.s.} \setminus A_{u.s.e.} \neq \emptyset$.

PROOF. We prove this result by explicitly constructing an $\{X_\tau\} \in A_{p.s.} \setminus A_{u.s.e.}$. Suppose $\mathbf{X} = \{0, 1\}$. Let \mathbf{u} be a deterministic sequence, and \mathbf{u} can be expressed as $\mathbf{u} = 0^{s_1}1^{t_1}0^{s_2}1^{t_2}\dots$, where 0^s (1^s) represents a sequence of all zeros (ones) with length s , and $\forall i, s_i, t_i > 0$. For instance, $\mathbf{u} = 0^11^20^21^3\dots = 01100111\dots$. We can then sequentially determine s_i and t_i as follows, and thus a specific \mathbf{u}_0 is constructed.

$$\begin{aligned} s_i &= \min \left\{ s > 0 : \frac{s + \sum_{j=1}^{i-1} s_j}{s + \sum_{j=1}^{i-1} s_j + t_j} > 2/3 \right\} \\ t_i &= \min \left\{ t > 0 : \frac{\sum_{j=1}^i s_j}{t + s_i + \sum_{j=1}^{i-1} s_j + t_j} < 1/3 \right\}. \end{aligned}$$

Namely, the resulting $f_r(0, t)$ keeps growing until it hits $2/3$, then starts to decrease until it hits $1/3$, then keeps growing again, and repeats this cycle indefinitely. Based on this construction, we can show that $\mathbf{u}_0 \in A_{p.s.}$ but not in $A_{u.s.e.}$.

It is worth mentioning that if $\{X_\tau\}$ is u.s.e. distributed in L^1 then the hitting time after any fixed time t has bounded expectation. However, the converse statement is not true. This relationship is formally stated as follows.

CONDITION E.1. $H_t(x)$, the first hitting time of x after a deterministic time t , satisfies the following condition:

$$\exists B < \infty, \text{ such that } \forall t, \mathbf{E}\{H_t(x)\} \leq B.$$

PROPOSITION E.3. Being u.s.e. distributed in L^1 implies CONDITION E.1, but not vice versa.

PROOF. “ \rightarrow ” is easy, and we construct an example to show that the converse is not true. Let $\mathbf{u}_{(i)}$ denote a collection of deterministic periodic sequences with period equal to $i + 1$. In each period, $\mathbf{u}_{(i)}$ contains a length $(i + 1)$ interval starting with a 1, and followed by all 0’s. This is illustrated by the examples for $i = 1, 2, 3$:

$$\mathbf{u}_{(1)} = \{1, 0, 1, 0, 1, 0, 1, 0, 1, 0, \dots\}$$

$$\mathbf{u}_{(2)} = \{1, 0, 0, 1, 0, 0, 1, 0, 0, 1, \dots\}$$

$$\mathbf{u}_{(3)} = \{1, 0, 0, 0, 1, 0, 0, 0, 1, 0, \dots\}.$$

Then we construct the $\{X_\tau\}$ with the distribution $\mathbf{P}(\{X_\tau\} = \mathbf{u}_{(i)}) = K \cdot \frac{1}{i^3}$, where K is a normalization factor such that $\sum_i \mathbf{P}(\{X_\tau\} = \mathbf{u}_{(i)}) = 1$.

According to the above construction, it is easy to check that $\{X_\tau\}$ satisfies CONDITION E.1. However, for any $B < \infty$, we can choose our stopping time T as the first time instant in which we have found B consecutive 0’s, followed by a 1. Then we have

$$\forall T < \infty, \mathbf{E}\{H_T(1)|T\} \geq B + 1.$$

Remarks:

- The example is nontrivial since $\mathbf{P}(T < \infty) > 0$ for all B .
- The above is a good example to illustrate why the definition of the desired even distribution involves stopping times when considering the benefit of observing side information $\{X_\tau\}$. In particular, for k large, the unevenly distributed sequence $\mathbf{u}_{(k)}$ will cause a significant amount of inferior sampling times for the decision scheme.

REFERENCES

- [Ada01] K. Adam, *Learning while searching for the best alternative*, Journal of Economic Theory **101** (2001), 252–280.
- [AHT88] R. Agrawal, M. V. Hegde, and D. Teneketzis, *Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost*, IEEE Trans. Automat. Contr. **33** (1988), no. 10, 899–906.
- [ATA89a] R. Agrawal, D. Teneketzis, and V. Anantharam, *Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space*, IEEE Trans. Automat. Contr. **34** (1989), no. 3, 258–267.
- [ATA89b] ———, *Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space*, IEEE Trans. Automat. Contr. **34** (1989), no. 12, 1249–1259.

- [AVW87a] 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. Automat. Contr. **32** (1987), no. 11, 968–976.
- [AVW87b] ———, *Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part II: Markovian rewards*, IEEE Trans. Automat. Contr. **32** (1987), no. 11, 977–982.
- [Ber72] D. A. Berry, *A Bernoulli two-armed bandit*, Ann. Math. Stat. **43** (1972), no. 3, 871–897.
- [Che72] H. Chernoff, *Sequential analysis and optimal design*, Society for Industrial and Applied Mathematics, Philadelphia, 1972.
- [Cla89] M. K. Clayton, *Covariate models for Bernoulli bandits*, Sequential Analysis **8** (1989), no. 4, 405–426.
- [Git79a] J. C. Gittins, *Bandit processes and dynamic allocation indices*, J. Royal Statistical Society. Series B (Methodological) **41** (1979), no. 2, 148–177.
- [Git79b] ———, *A dynamic allocation index for the discounted multiarmed bandit problem*, Biometrika **66** (1979), no. 3, 561–565.
- [GP91] B.K. Ghosh and P.K.Sen, *Handbook of sequential analysis*, Dekker, New York, 1991.
- [KL00] S. R. Kulkarni and G. Lugosi, *Finite-time lower bounds for the two-armed bandit problem*, IEEE Trans. Automat. Contr. **45** (2000), no. 4, 711–714.
- [KR95] M. N. Katehakis and H. Robbins, *Sequential choice from several populations*, Proc. Nat. Acad. Sci. USA, vol. 92, Sept. 1995, pp. 8584–8585.
- [Kul93] S. R. Kulkarni, *On bandit problems with side observations and learnability*, Proc. 31st Allerton Conf. Commun. Contr. Comp., Sept. 1993, pp. 83–92.
- [LR84] T. L. Lai and H. Robbins, *Asymptotically optimal allocation of treatments in sequential experiments*, Design of Experiments : Ranking and Selection, by Thomas J. Santner, Ajit C. Tamhane, New York: Dekker, 1984.
- [LR85] ———, *Asymptotically efficient allocation rules*, Adv. Appl. Math. **6** (1985), no. 1, 4–22.
- [LY95] T. L. Lai and S. Yakowitz, *Machine learning and nonparametric bandit theory*, IEEE Trans. Automat. Contr. **40** (1995), no. 7, 1199–1209.
- [Rob52] H. Robbins, *Some aspects of the sequential design of experiments*, Bull. Am. Math. Soc. **58** (1952), 527–535.
- [Sar91] J. Sarkar, *One-armed bandit problems with covariates*, Ann. Statist. **19** (1991), no. 4, 1978–2002.
- [WKP04] C. C. Wang, S. R. Kulkarni, and H. V. Poor, *Bandit problems with side observations*, IEEE Trans. Automat. Contr. (2004), to appear.
- [Woo79] M. Woodroofe, *A one-armed bandit problem with a concomitant variable*, J. Amer. Stat. Assoc. **74** (1979), no. 368, 799–806.
- [Zou94] T. Zoubeidi, *Optimal allocations in sequential tests involving two populations with covariates*, Commun. Statist.: Theory and Methods **23** (1994), no. 4, 1215–1225.

DEPARTMENT OF ELECTRICAL ENGINEERING
 PRINCETON UNIVERSITY
 PRINCETON, NJ 08544.
 EMAIL: {CHIHW, KULKARNI, POOR}@PRINCETON.EDU