Available online at www.sciencedirect.com
SCIENCE DIRECT

Advances in Applied Mathematics 34 (2005) 903–938

ADVANCES IN Applied Mathematics

www.elsevier.com/locate/yaama

Arbitrary side observations in bandit problems ☆

Chih-Chun Wang *, Sanjeev R. Kulkarni, H. Vincent Poor

Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

Received 19 March 2004; accepted 11 October 2004

Abstract

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 specifying different sub-bandit machines.

Keywords: 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.

* Corresponding author.

E-mail addresses: chihw@princeton.edu (C.-C. Wang), kulkarni@princeton.edu (S.R. Kulkarni), poor@princeton.edu (H.V. Poor).

0196-8858/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.aam.2004.10.004

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_t^1 and Y_t^2 denote the respective reward sequences at time t from machines 1 and 2. The reward function is then defined as follows,

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

where $1_{\{\cdot\}}$ is the indicator function, ϕ_τ, taking values in $\{1,2\}$, is the player's strategy at time τ, and $\{\alpha_\tau\}$ is a predefined discount sequence.

With the assumption that the distributions of $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$ are unknown to the player, the knowledge of which arm yields higher reward can only be gathered from sampling both arms often enough. However, this task of learning both arms inevitably limits the opportunity of pulling the more rewarding arm. Our goal is to maximize $W_\phi(t)$ under various conditions and discount sequences. Due to this inherent nature of coordinated learning and control, bandit problems have drawn much attention in various areas of statistics, control, learning, and economics, as in [1,7,9,14,18–21].

Typical optimality criteria include maximizing the expected reward $\lim_{t\to\infty} \mathsf{E}\{W_\phi(t)\}$ or maximizing the averaged expected reward $\lim_{t\to\infty} \mathsf{E}\{W_\phi(t)\}/t$. The former optimality condition is usually considered either within a finite horizon setting: $\alpha_\tau = 1_{\{\tau \leqslant t_0\}}$, or with an infinite geometric discount sequence: $\alpha_\tau = r^\tau$, $r < 1$ [12,13], while the latter is more appropriate to situations with no discounting, namely, $\alpha_\tau = 1$, $\forall \tau \in \mathbb{N}$. The unknown reward distribution is often parametrized as F_θ, where the rewards $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$ are governed by F_{θ_1} and F_{θ_2}. The decision maker has complete knowledge of the entire family $\{F_\theta\}_{\theta\in\boldsymbol{\Theta}}$, but the underlying parameter pair (θ_1, θ_2), taking values in $\boldsymbol{\Theta}^2$, is unknown. Dynamic programming is the central technique for solving these problems. Further discussions can be found in [7].

In this paper, we will focus on maximizing the averaged expected reward with no discounting, i.e., $\alpha_\tau = 1$, $\forall \tau \in \mathbb{N}$. Let μ_1 and μ_2 denote the expected returns of arms 1 and 2 under distributions F_{θ_1} and F_{θ_2}. By Wald's lemma, $\mathsf{E}\{W_\phi(t)\}$ can be rewritten as:

$$\mathsf{E}\big\{W_\phi(t)\big\} = t \cdot \max\{\mu_1, \mu_2\} - |\mu_1 - \mu_2| \cdot \mathsf{E}\big\{T_{\inf}(t)\big\}, \tag{1.1}$$

where $T_{\inf}(t)$ is the total number of samples taken on the inferior arm up to time t. More explicitly, $T_{\inf}(t) = \sum_{\tau=1}^{t} 1\{\phi_\tau \neq \arg\max(\mu_1, \mu_2)\}$. Since the term $|\mu_1 - \mu_2| \cdot \mathsf{E}\{T_{\inf}(t)\}$ represents the expected cost of not knowing the preference between μ_1 and μ_2, it is often called the "regret", and is commonly considered in the literature of bandit problems. For notational simplicity, in this paper, we will study the inferior sampling time $T_{\inf}(t)$. It is worth noting that all expectations used in this paper depend on the unknown F_{θ_1} and F_{θ_2} and thus are functions of the parameter pair $C_0 = (\theta_1, \theta_2)$. Hence the terms $\mathsf{E}\{T_{\inf}(t)\}$ and $\mathsf{E}_{C_0}\{T_{\inf}(t)\}$ are used interchangeably.

1.1. Uniformly good rules

Lai and Robbins [19] considered bandit problems using a non-Bayesian min–max approach with no discounting (i.e., $\forall\tau,\ \alpha_\tau = 1$), and in which the objective function is $\max_\phi \min_{\theta_1,\theta_2} \mathsf{E}_{C_0}\{W_\phi(t)\}$. Recasting this problem in terms of minimizing the regret rather than maximizing the rewards, this formalism leads to the following useful definition of uniformly good rules.

Definition 1.1 (*Uniformly good rules* [19]). An allocation rule is uniformly good if for every possible (θ_1,θ_2) pair, $\mathsf{E}_{C_0}\{T_{\inf}(t)\} = o(t^\alpha)$, $\forall\alpha > 0$.

A $\log t$ lower bound on achievable regret has been proved for uniformly good rules under various settings [5,18,19], and this is quoted as follows.

Theorem 1.1 ($\log t$ lower bound). *For any uniformly good rule $\{\phi_\tau\}$, $T_{\inf}(t)$ satisfies*

$$\lim_{t\to\infty} \mathsf{P}_{C_0}\left(T_{\inf}(t) \geqslant \frac{(1-\varepsilon)\log t}{K_{C_0}}\right) = 1, \quad \forall\varepsilon > 0, \quad \textit{and}$$
$$\liminf_{t\to\infty} \frac{\mathsf{E}_{C_0}\{T_{\inf}(t)\}}{\log t} \geqslant \frac{1}{K_{C_0}}, \tag{1.2}$$

where K_{C_0} is a constant depending on C_0. If $\arg\max(\mu_1,\mu_2) = 2$, *then $T_{\inf}(t) = T_1(t)$ and K_{C_0} is defined*[1] *as*:

$$K_{C_0} = \inf\bigl\{I(\theta_1,\theta)\colon\ \forall\theta,\ \mu_\theta > \mu_{\theta_2}\bigr\}, \tag{1.3}$$

where $I(\theta_1,\theta) = \mathsf{E}_{\theta_1}\log(\mathrm{d}F_{\theta_1}/\mathrm{d}F_\theta)$ is the Kullback–Leibler (*K-L*) *information number between F_{θ_1} and F_θ, and μ_θ is the expected reward under F_θ. The expression for K_{C_0} for the case in which* $\arg\max(\mu_1,\mu_2) = 1$ *can be obtained by symmetry.*

The asymptotic sharpness of the above lower bound is also proved in the above papers:

Theorem 1.2 (Asymptotic sharpness). *Under certain regularity conditions,*[2] *the above lower bound is asymptotically sharp. That is, given the family of possible distributions $\{F_\theta\}$, there exists a decision rule $\{\phi_\tau\}$ such that for all $C_0 = (\theta_1,\theta_2)$,*

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

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

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

[2] If the parameter space is finite, Theorem 1.2 always holds. If $\boldsymbol{\Theta}$ is continuous, the required regularity conditions concern the unboundedness and the continuity of μ_θ w.r.t. θ and the continuity of $I(\theta_1,\theta)$ w.r.t. μ_θ.

Remark. For arbitrary rules, one possible situation is that the decision rule samples the inferior arm a *finite* number of times (depending on the sample path), and thereafter sticks to the seemingly superior arm indefinitely. For such rules, we may have $\lim_{t\to\infty} T_{\inf}(t) < \infty$ almost surely for certain values of C_0. However, since no forced sampling is performed after a finite amount of time, one can prove that $\mathsf{E}_{C_0'}\{T_{\inf}(t)\}$ grows linearly for some other C_0', and these rules are thus *not* uniformly good. A uniformly good rule, on the other hand, must always be skeptical, and keeps sampling the other arm infinitely often. Theorem 1.1 guarantees that the probability of the inferior sampling time $T_{\inf}(t)$ exceeding $\log t/K_{C_0}$ converges to one as t tends to infinity. In other words, the forced sampling times must grow at least on the order of $\log(t)$ with the minimum constant $1/K_{C_0}$.

Henceforth we consider only uniformly good rules. As discussed, by limiting our focus to uniformly good rules, the possibility of almost sure finiteness of $T_{\inf}(t)$ is sacrificed,[3] but acceptable performance is guaranteed for all possible C_0. Further results on uniformly good rules within slightly different settings can be found in [2–6,15,16,19].

1.2. Bandit problems with side information

A common scenario in practice is that before making the decision ϕ_t (at time t), another random variable X_t, taking values in $\mathbf{X}$, can be observed. Suppose at time instant t, $X_t = x$. The rewards (Y_t^1, Y_t^2) are then governed by the conditional distributions[4] $F_{\theta_1}(\cdot|X_t = x)$ and $F_{\theta_2}(\cdot|X_t = x)$, and have conditional expected return $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$. Additional gain is expected once this new structure is exploited. It is worth noting that under this new framework, the inferior sampling time $T_{\inf}(t)$ is defined slightly differently as

$$T_{\inf}(t) = \sum_{\tau=1}^{t} 1\bigl\{\phi_\tau \neq \arg\max\bigl(\mu_{\theta_1}(X_\tau), \mu_{\theta_2}(X_\tau)\bigr)\bigr\},$$

and the traditional two-armed bandit without side observations X_t can be viewed as a degenerate case in which the range of X_t contains only one element: $\mathbf{X} = \{x_0\}$.

This idea was first introduced by Woodroofe [24], where an independent and identically distributed[5] (i.i.d.) $\{X_\tau\}$ was considered. Contrary to the traditional bandit problems (without side information), Woodroofe proved that even a myopic approach becomes asymptotically optimal, assuming the governing conditional distributions $F_{\theta_i}(\cdot|X_t)$ are Gaussian with means $\theta_i + X_t$ and variances 1. Sarkar [22] extended the simple relationship in [24] to exponential families. [23] focused solely on i.i.d. $\{X_\tau\}$, and various levels of asymptotic

[3] It will be shown in this paper that under certain scenarios, the almost sure finiteness of $T_{\inf}(t)$ can be recovered by exploiting the side information.

[4] The term "side observation" implies that the distribution of X_t depends on the upcoming rewards Y_t^1 and Y_t^2. Nevertheless, since X_t is observed before deciding which arm to pull, it is more convenient to reverse the conditional probability using Bayes' formula and view X_t as the basic quantity while letting the distributions of Y_t^1 and Y_t^2 depend on the value of X_t. Formal description of this underlying relationship among $\{X_\tau\}$, $\{Y_\tau^1\}$, $\{Y_\tau^2\}$, and (θ_1, θ_2) can be found in Section 2.1.

[5] In the literature of bandit problems, the commonly used term "i.i.d. side observation $\{X_\tau\}$" refers to a marginally i.i.d. $\{X_\tau\}$. Namely, after averaging over $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$, $\{X_\tau\}$ becomes an i.i.d. random process.

efficiency were proved after abstracting the relationship between $\{X_\tau\}$ and $(\{Y_\tau^1\}, \{Y_\tau^2\})$ into four separate categories, which included the results in [22,24] as special cases. Other approaches regarding side observations can be found in [10,17,25].

Results of [22–24] suggest that the benefits of side observations for bandit problems are not due to the *random* appearance of all values x of the i.i.d. $\{X_\tau\}$, but rather are due to the *evenly* distributed appearance of all possible x. In this paper, we further extract the essential properties of "evenly distributed appearance" and investigate their effects on the attainable results. Our results generalize the benefit of side observations to a wide range of non-i.i.d. processes.

1.3. Examples of uniformly good rules and side information

Here we provide several examples illustrating the benefits of side information.

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. The channel characteristics depend on an unknown parameter pair $C_0 = (\theta_1, \theta_2)$, which might represent the propagation coefficients, the number of paths in a multipath channel, the K-factors of Rician channels, etc. The side information X_t (not necessarily i.i.d.) might be a noisy measurement of the parameter pair (θ_1, θ_2), or geographical information about the receiver, or X_t could be a pair containing both of these types of information. In the following examples, the range of θ and x are simplified as $\{1, 2, 3, 4\}$ or $\{1, 2, 3\}$, and the governing conditional distributions $F_\theta(\cdot|X_t = x)$ are Bernoulli with success probability $p_{\theta,x}$. The entire family of conditional distributions can then be specified by a matrix $(p_{\theta,x})$, and we will discuss the following three examples.

Example 1.

$$(p_{\theta,x}) = \begin{pmatrix} 0.4 & 0.3 & 0.6 \\ 0.5 & 0.5 & 0.5 \\ 0.6 & 0.7 & 0.4 \end{pmatrix}.$$

Example 2.

$$(p_{\theta,x}) = \begin{pmatrix} 0.4 & 0.3 & 0.2 \\ 0.5 & 0.5 & 0.5 \\ 0.6 & 0.7 & 0.8 \end{pmatrix}.$$

Example 3.

$$(p_{\theta,x}) = \begin{pmatrix} 0.4 & 0.4 & 0.5 \\ 0.5 & 0.5 & 0.4 \\ 0.6 & 0.6 & 0.6 \\ 0.7 & 0.8 & 0.9 \end{pmatrix}.$$

Suppose $\{X_\tau\}$ is an i.i.d. (see footnote 5) sequence with its marginal uniformly distributed among $\{1, 2, 3\}$. For any parameter θ, if we ignore the side information X_t, the player is then facing a Bernoulli distribution with parameter $p_{\theta-} := (p_{\theta,1} + p_{\theta,2} + p_{\theta,3})/3$. Suppose the true parameter pair $C_0 = (\theta_1, \theta_2)$ equals $(1, 2)$ (unknown to the player). By Theorem 1.1, $\lim \mathsf{E}_{C_0}\{T_{\mathrm{inf}}(t)\}/\log t \geqslant 1/K_{C_0}$, where K_{C_0} is $0.0358 = I(p_{1-}, p_{3-})$ for Example 1, $0.3389 = I(p_{1-}, p_{3-})$ for Example 2, and $0.0564 = I(p_{1-}, p_{3-})$ for Example 3.

[23] shows that by exploiting X_t, these $\log t$ lower bounds can be surpassed, and have different levels of improvement. For Example 1, there exists a uniformly good rule ϕ_t achieving bounded expected rewards: $\lim_{t\to\infty} \mathsf{E}\{T_{\mathrm{inf}}(t)\} < \infty$. For Example 2, the performance is still $\log t$ lower bounded, but a smaller constant $1/K'_{C_0}$ can be achieved: $\lim \mathsf{E}_{C_0}\{T_{\mathrm{inf}}(t)\}/\log t \geqslant 1/K'_{C_0}$ with $K'_{C_0} = I(p_{1,3}, p_{3,3}) = 0.8318$. For Example 3, there exists a uniformly good rule admitting bounded $\lim_{t\to\infty} \mathsf{E}\{T_{\mathrm{inf}}(t)\}$.

[23] also demonstrates that the amount of improvement may depend on the unknown value of (θ_1, θ_2). Within the setting of Example 3, if the unknown (θ_1, θ_2) equals $(2, 3)$ instead of $(1, 2)$ (contrary to the previous discussion), it can be proved that no rule can achieve bounded $\mathsf{E}\{T_{\mathrm{inf}}(t)\}$ and the minimum regret is still $\log t$ lower bounded. The best achievable constant in front of $\log t$ becomes $1/K'_{C_0}$ with $K'_{C_0} = I(p_{2,3}, p_{4,3}) = 0.7507$. For comparison, the traditional $\log t$ lower bound (ignoring side observations) is $\log t/K_{C_0}$, $K_{C_0} = I(p_{2-}, p_{4-}) = 0.2716$.

These three examples possess different internal structures and thus the side observations provide different improvements. In Sections 3 through 6, we will show that these improvements over traditional bandit problems can be achieved with a more general class of X_t's, including but not limited to i.i.d. sequences, Markov chains, and fixed periodic sequences.

This paper is organized as follows. In Section 2, we provide a rigorous formulation of side-observation-aided bandit problems and give formal definitions of several "even distribution" properties, examples of each such property, and relationships among them. In Sections 3 through 6, we provide results for various relationships among $\{X_\tau\}$, $\{Y_\tau^1\}$, and $\{Y_\tau^2\}$ with the satisfaction of the "even distribution" properties defined in Section 2.2. All results in [23], obtained under the assumption of i.i.d. $\{X_\tau\}$, hold as special cases under this new framework, which includes many other side observation processes (e.g., Markov chains and periodic sequences) as well. Section 7 provides a summary table and a simple necessary condition concerning the extent of the benefit obtained from observing $\{X_\tau\}$. Section 8 concludes the paper.

2. Formulations

2.1. Side information

To characterize explicitly the correlation among $C_0 = (\theta_1, \theta_2)$, $\{X_\tau\}$, $\{Y_\tau^1\}$ and $\{Y_\tau^2\}$, the probability distribution of the two-armed bandit with side observations is modelled as follows. At times $t_1, \ldots, t_k$, the joint probability distribution of $(X_{t_i}, Y_{t_i}^1, Y_{t_i}^2)_{i=1,\ldots,k}$ is

$$G_{t_1,\ldots,t_k|C_0}(x_{t_1}, \ldots, x_{t_k}) \prod_{i=1}^{k} F_{\theta_1}\big(y_{t_i}^1 \big| x_{t_i}\big) F_{\theta_2}\big(y_{t_i}^2 \big| x_{t_i}\big), \tag{2.4}$$

where $G_{t_1,\ldots,t_k|C_0}(x_{t_1}, \ldots, x_{t_k})$ is the finite cylinder distribution of the side information $\{X_\tau\}$, which may or may not depend on C_0. Both families of distributions, $\{G_{\cdots|C_0}\}_{C_0}$ and

$\{F_\theta(\cdot|x)\}_\theta$, are known to the decision maker, and only the true value of C_0 is unknown. There is few restriction on $\mathbf{X}$ and $\boldsymbol{\Theta}$, the ranges of X_t and θ. To significantly simplify the notation, both $\mathbf{X}$ and $\boldsymbol{\Theta}$ are assumed to be subsets of $\mathbb{R}$.

Some useful notation is as follows.

$$M_{C_0}(x) := \arg\max\bigl(\mu_{\theta_1}(x), \mu_{\theta_2}(x)\bigr),$$

which denotes the index of the more rewarding arm (having higher conditional expected reward $\mu_{\theta_i}(x)$) given the side observation $X_t = x$. For any configuration pair $C_0 = (\theta_1, \theta_2)$, we may use $1(C_0) := \theta_1$ and $2(C_0) := \theta_2$ to denote the first and second coordinates of the configuration pair C_0. For example, $\mu_{2(C_0)}(x) = \mu_{\theta_2}(x)$ and $F_{1(C_0)}(\cdot|x) = F_{\theta_1}(\cdot|x)$.

Remark 1. For example, a bandit problem with i.i.d. side observation sequence means $G_{t_1,\ldots,t_k|C_0}(x_{t_1}, \ldots, x_{t_k}) = \prod_{i=1}^k G_{t_i|C_0}(x_{t_i}) = \prod_{i=1}^k G_{t_1|C_0}(x_{t_i})$.

Remark 2. 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 a sequence of independent rewards for $i = 1, 2$.

2.2. *Even distribution properties*

Our goal is to extract the essential "evenly distributed" properties of a side observation process that are beneficial to uniformly good rules. Three levels of evenly distributed properties will be formally defined in this subsection.

Suppose X_t takes values in a finite state set $\mathbf{X}$, and the relative frequency of x up to time t is denoted as $f_{\mathrm{r}}(x,t) = (\sum_{\tau=1}^t 1\{X_\tau = x\})/t$.

Definition 2.1 (*Evenly distributed in L^1*). $\{X_\tau\}$ is evenly distributed in L^1 if

$$\forall x \in \mathbf{X}, \quad \pi(x) := \liminf_{t\to\infty} \mathsf{E}\bigl\{f_{\mathrm{r}}(x,t)\bigr\} > 0.$$

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

$$\forall x \in \mathbf{X}, \quad \mathsf{E}\left\{\sum_{\tau=1}^{\infty} 1\bigl\{f_{\mathrm{r}}(x,\tau) < \pi(x)\bigr\}\right\} < \infty.$$

This property automatically implies that $\forall x$, $\liminf_{t\to\infty} f_{\mathrm{r}}(x,t) \geqslant \pi(x)$ 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

$$\forall T,\ \forall x \in \mathbf{X}, \quad \mathsf{E}\bigl\{H_T(x)|T\bigr\} \leqslant B,$$

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

It is easy to verify that these three properties hold for non-degenerate i.i.d. sequences, Markov chains, and fixed periodic sequences, which shows the generality of these classes of distributions.

Remark. It can be shown that each of Definitions 2.2 and 2.3 implies Definition 2.1, and Definition 2.2 does not imply Definition 2.3. Whether Definition 2.2 can be derived from Definition 2.3 remains an open problem.

The following four sections are devoted to determining even distribution properties that are sufficient for different levels of improvement.

3. Case 1: direct information from side observations

In this setting, the side observation X_t directly reveals information about $C_0 = (\theta_1, \theta_2)$. As a result, the dilemma between learning C_0 and control (sampling the superior arm) can be solved by learning C_0 from X_t and sampling the seemingly better arm, Y_t^1 or Y_t^2. The formal definition of this situation is given below and can be viewed as an identifiability condition.

Definition 3.1 (*Direct information*). If $C_0 \neq C_0'$, then $\exists t_1, \ldots, t_k$, such that $G_{t_1,\ldots,t_k|C_0} \neq G_{t_1,\ldots,t_k|C_0'}$.

3.1. 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 can be described as Algorithm 1, an algorithm executed at time[6] t.

Algorithm 1 (*ϕ_t, the decision at time t*)

```
1: Obtain an estimate $\hat{C}_t$ based on the side observations $X_1, \ldots, X_t$. 2: Set $\phi_t = M_{\hat{C}_t}(X_t)$.
```

To further bound the performance of this scheme, we need the following condition.

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

Example 4. Suppose $\mathbf{\Theta} = \mathbb{R}$ and $\mathbf{X} \subset \mathbb{R}$ is finite. If $\forall x \in \mathbf{X}$, $\mu_\theta(x)$ is continuous with respect to θ, then Condition 3.1 is satisfied.

Example 5. Suppose $\mathbf{\Theta}$ and $\mathbf{X}$ are arbitrary subsets of $\mathbb{R}$. If $F_\theta(\cdot|x) \sim \mathcal{N}(\theta x, 1)$, a standard Gaussian distribution with mean θx, then Condition 3.1 is satisfied.

[6] "At time t" means after observing X_t but before the decision ϕ_t is made. It is basically the moment when we are determining the value of ϕ_t.

Theorem 3.1. *Suppose both Definition* 3.1 *and Condition* 3.1 *are satisfied. For all C_0 and any sequence of estimates $\{\hat{C}_\tau\}$, there exists $\varepsilon > 0$ such that Algorithm* 1 *satisfies*

$$\lim_{t\to\infty} \frac{\mathsf{E}_{C_0}\{T_{\text{inf}}(t)\}}{\sum_{\tau=1}^{t} \mathsf{P}_{C_0}(|\hat{C}_\tau - C_0| > \varepsilon)} \leqslant 1.$$

A detailed proof is given in Appendix A.

The above theorem provides an upper bound on the best achievable expected inferior sampling time, and is illustrated in the following examples.

Example 6. Suppose $\{X_\tau\}$ is an i.i.d. sequence with marginal distribution G_{C_0} on $\mathbb{R}$, and the mapping from C_0 to G_{C_0} is one-to-one. Then by the large deviations theorem on $\mathbb{R}$, there exists $\{\hat{C}_\tau\}$ such that $\forall C_0$, $\varepsilon > 0$, $\lim_{t\to\infty} \sum_{\tau=1}^{t} \mathsf{P}_{C_0}(|\hat{C}_\tau - C_0| > \varepsilon) < \infty$. By Theorem 3.1, $\forall C_0$, we have $\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\text{inf}}(t)\} < \infty$, and thus the proposed scheme is uniformly good.

Example 7. Suppose $\{X_\tau\}$ 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 similar reasoning as in the i.i.d. case, there exists a uniformly good rule such that $\forall C_0, \lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\text{inf}}(t)\} < \infty$.

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 $\boldsymbol{\Theta}$ is finite, we can easily find $\{\hat{C}_\tau\}$ such that $\forall C_0$, $\varepsilon > 0$, $\lim_{t\to\infty} \sum_{\tau=1}^{t} \mathsf{P}_{C_0}(|\hat{C}_\tau - C_0| > \varepsilon) < \infty$. Hence $\forall C_0$, $\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\text{inf}}(t)\} < \infty$, and the proposed scheme is uniformly good.

4. Case 2: best arm as a function of X_t

In Sections 4 to 6, we turn to another formalism for the interaction of X_t with the bandits. In particular, here and in the following two sections, we consider the situation in which the distribution of $\{X_t\}$ is not a function of C_0, namely, $G_{t_1,\ldots,t_k|C_0} := G_{t_1,\ldots,t_k}$. For convenience, we will assume throughout these three sections that $\mathbf{X} \subseteq \mathbb{R}$. Three cases offering further refinements concerning the relationships between $M_C(x)$ and x will be discussed separately (one in each section).

4.1. Formulation

In this section, we assume that the side observation X_t is *always* able to change the preference order, formally defined as follows and illustrated in Fig. 1.

Definition 4.1 (*Best arm is a function of X_t*). $\forall C \in \boldsymbol{\Theta}^2$, there exist $x_1, x_2 \in \mathbf{X}$, such that $M_C(x_1) = 1$ and $M_C(x_2) = 2$.

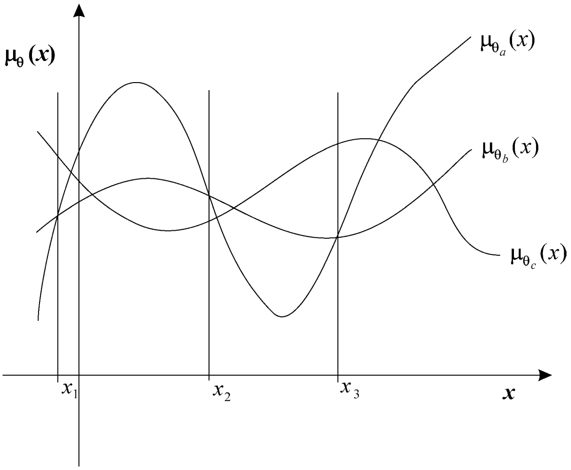

Fig. 1. The best arm at time t is a function of the side observation X_t. That is, for any possible pair $C = (\theta_1, \theta_2)$, the two curves, $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ (as a function of x) always intersect each other.

Three necessary regularity conditions are as follows.

R1: $\mathbf{X}$ is finite.

R2: $I(\theta_1, \theta_2|x)$ is finite and strictly positive for all possible θ_1, θ_2, and x, where the conditional K-L information number $I(\theta_1, \theta_2|x)$ is defined as the K-L information between the conditional distributions $F_{\theta_1}(\cdot|x)$ and $F_{\theta_2}(\cdot|x)$.

R3: $\boldsymbol{\Theta} \subseteq \mathbb{R}$, and $\forall x$, $\mu_\theta(x)$ is continuous as a function of θ.

An example that satisfies these regularity conditions is as follows:

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

Remark. 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., they have *non-identical* arms.

4.2. Scheme with bounded $\lim_t \mathsf{E}_{C_0}\{T_{\mathrm{inf}}(t)\}$

Although no information about C_0 is revealed through observing X_t, significant improvement, i.e., bounded $\lim_t \mathsf{E}_{C_0}\{T_{\mathrm{inf}}(t)\}$, can be obtained when the best arm is a function of X_t. This is seen from the following result.

Theorem 4.1. *Suppose the best arm is a function of X_t as in Definition* 4.1*, and regularity conditions* R1, R2*, and* R3 *are satisfied. If the side observation sequence $\{X_\tau\}$ is evenly distributed in probability series, then there exists a uniformly good rule $\{\phi_\tau\}$ such that $\forall C_0$, the expected inferior sampling time is bounded*:

$$\forall C_0,\quad \mathsf{E}_{C_0}\big\{T_{\inf}(t)\big\} \leqslant \lim_{t\to\infty} \mathsf{E}_{C_0}\big\{T_{\inf}(t)\big\} < \infty.$$

The $\log t$ *lower bound for traditional bandit problems is thus surpassed.*

Remark. Although the side observation X_t does not reveal any information about C_0, the even distribution of X_t on different values x results in the alternation of the best arm $M_{C_0}(X_t)$. With this alternation, it is then possible to always pull the seemingly better arm $M_{\hat{C}_{t-1}}(X_t)$, 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 this is where the major improvement ($\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\inf}(t)\} < \infty$) comes from.

Algorithm 2 (ϕ_{t+1}, *the decision at time* $t+1$)

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

$$T_i^x(t) := \sum_{\tau=1}^{t} 1\{X_\tau = x,\ \phi_\tau = i\},\quad \text{and}\quad x_i^\star := \arg\max_x \big\{T_i^x(t)\big\},\quad T_i^{x^\star}(t) := \max_x \big\{T_i^x(t)\big\}.$$

Construct $\mathbf{C}_t \subseteq \mathbf{\Theta}^2$ as follows:

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

where

$$\sigma(C,t) := \rho\big(F_{1(C)}\big(\cdot\,|x_1^\star\big), L_1^{x_1^\star}(t)\big) + \rho\big(F_{2(C)}\big(\cdot\,|x_2^\star\big), L_2^{x_2^\star}(t)\big),$$

and $L_i^x(t)$ is the current empirical measure of rewards sampled from arm i at those time instants when $X_\tau = x$. (Here, ρ denotes the Prohorov metric[7] over distributions on $\mathbb{R}$.) After constructing $\mathbf{C}_t$, arbitrarily choose $\hat{C}_t \in \mathbf{C}_t$.

Algorithm:

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

Note that lines 1 and 2 guarantee that in line 3, there is at most one i satisfying $T_i(t) < \sqrt{t+1}$.

An example scheme $\{\phi_\tau\}$ achieving $\lim_t \mathsf{E}_{C_0}\{T_{\inf}(t)\} < \infty$ is described in Algorithm 2.

The intuition behind Algorithm 2 is as follows. Since the forced sampling mechanism (in line 3) guarantees that each arm will be sampled often enough, at least $O(t^{1/2})$, the expected duration of $\{|\hat{C}_t - C_0| > \varepsilon\}$ is bounded. As a result, most of the time $\hat{C}_t$ and C_0 will have the same arm preference and $T_{\inf}(t)$ will be mostly contributed to by choices $\phi_{t+1} = i$ (line 4) instead of choices $\phi_{t+1} = M_{\hat{C}_t}(X_{t+1})$ (line 6). However, if we apply Algorithm 2 to traditional bandit problems, this forced sampling mechanism (line 3) inevitably results in $O(t^{1/2})$ inferior samplings, which is too often for a uniformly good rule. But when applied to a side-observation-aided bandit problem, the alternating nature of $M_{C_0}(x)$ in Definition 4.1 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. Both $T_1(t)$ and $T_2(t)$ will grow linearly with t, and the forced sampling mechanism will rarely be triggered. As a result, $\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{\inf}(t)\}$ is finite in Algorithm 2. A detailed analysis is provided in Appendix B.

5. Case 3: best arm is not a function of X_t

Following Section 4, we assume that X_t reveals no information about C_0, i.e., $G_{C_0} = G$. In this section, we consider the case in which $\forall C_0$, X_t *never* changes the preference order. This setting is illustrated in Fig. 2 and is formally defined as follows.

Definition 5.1 (*Best arm is not a function of* X_t). Given any $C = (\theta_1, \theta_2)$, 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 $M_C(x)$.

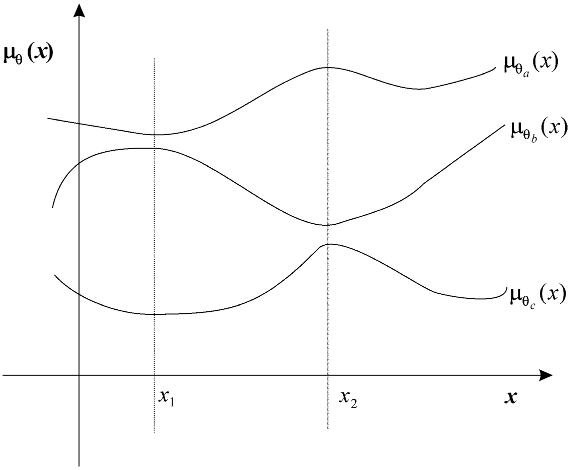

Fig. 2. The best arm at time t is not a function of the side observation X_t. That is, for any possible pair, (θ_1, θ_2), the two curves, $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$, do not intersect each other. In this case, we can postpone our sampling until the most informative time instants.

Within the three regularity conditions:

R1: $\mathbf{X}$ is finite,
R2: $\forall\theta_1, \theta_2, x$, the conditional K-L information number $I(\theta_1, \theta_2|x)$ is finite and strictly positive, and
R4: the parameter space $\mathbf{\Theta} \subseteq \mathbb{R}$ can be relabelled,[8] so that $\forall x$, the conditional expected reward $\mu_\theta(x)$ is strictly increasing with respect to θ,

we can still obtain improvements over the traditional bandit problems in this case.

An example that satisfies these regularity conditions is as follows.

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

5.1. $\log t$ *lower bound*

Unlike the situation in Section 4, the side observation X_t is not able to alter the preference arm $M_{C_0}(x)$, so the dilemma between learning and control still exists. For this situation, a $\log t$ lower bound with a new constant was proved in [23] for bandit problems with i.i.d. side observation $\{X_\tau\}$. Since the same proof applies to arbitrary random processes $\{X_\tau\}$, we restate the $\log t$ lower bound theorem [23, Theorem 5, p. 13] for general random processes $\{X_\tau\}$.

Theorem 5.1 [23, Theorem 5, p. 13]. *Suppose that for all possible C_0, the best arm $M_{C_0}(x)$ is constant for all x as in Definition* 5.1, *and the regularity conditions* R1, R2, *and* R4 *are satisfied. For any uniformly good rule $\{\phi_\tau\}$, $T_{\inf}(t)$ is lower bounded by*

$$\begin{aligned}
&\lim_{t\to\infty} \mathsf{P}_{C_0}\left(T_{\inf}(t) \geqslant \frac{(1-\varepsilon)\log t}{K_{C_0}}\right) = 1, \quad \forall \varepsilon > 0, \quad \textit{and} \\
&\liminf_{t\to\infty} \frac{\mathsf{E}_{C_0}\{T_{\inf}(t)\}}{\log t} \geqslant \frac{1}{K_{C_0}},
\end{aligned} \tag{5.1}$$

where K_{C_0} is a constant depending on C_0. If $M_{C_0} = 2$, then $T_{\inf}(t) = T_1(t)$. The constant K_{C_0} can be expressed[1] *as follows.*

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

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

Note that by the convexity of the Kullback–Leibler information, we have

$$\sup_x I(\theta_1, \theta|x) \geqslant \int I(\theta_1, \theta|x) G_{t,C_0}(x)\,\mathrm{d}x \geqslant I(\theta_1, \theta).$$

[8] 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). This condition is imposed simply for convenience.

As a result, the new constant $1/K_{C_0}$ in (5.2) is no larger than the old constant in (1.3) for traditional bandit problems. This shows that the additional side information X_t improves the decision in the bandit problem, which of course it must.

5.2. Scheme achieving the lower bound

To construct a tractable scheme achieving the $\log t$ lower bound (5.1), we first need the following assumptions.

A1: The parameter space $\boldsymbol{\Theta}$ is finite.
A2: The side observations $\{X_\tau\}$ are u.s.e. distributed in L^1.
A3: The value of the game,

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

exists.[9]

We then consider a specific subset of uniformly good rules for traditional bandit problems, which was introduced in [3] for the case of a finite parameter space $\boldsymbol{\Theta}$. This type of decision rule possesses the following three properties when being applied to traditional bandit problems.

1. After time t, an estimate $\hat{C}_t = (\hat{\theta}_1, \hat{\theta}_2)$ is constructed and is 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 finite,[10] namely,

$$\lim_{t\to\infty} \mathsf{E}_{C_0}\left\{\sum_{\tau=1}^{t} 1\bigl\{\hat{C}_\tau \neq C_0\bigr\}\right\} < \infty.$$

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

$$\lim_{t\to\infty} \frac{\mathsf{E}_{C_0}\bigl\{\sum_{\tau=1}^{t} 1\{\hat{C}_\tau = C_0, \phi_{\tau+1} \neq M_{C_0}\}\bigr\}}{\log t} \leqslant \frac{1}{K_{C_0}},$$

where K_{C_0} is defined[1] as $\inf_{\{\theta:\ \theta>\theta_2\}} I(\theta_1,\theta)$ if $M_{C_0} = 2$.

[9] A sufficient condition for the existence of the value of the game is that θ is the dominant factor (compared to x) in determining the conditional distributions $F_\theta(\cdot|x)$. In many cases of interest, the parameter plays a more critical role in determining the distribution than the side observation x. Therefore this condition on the value of the game is a reasonable assumption and is generally satisfied.

[10] In 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.

Definition 5.2 (*Tight decision rule ϕ_t*). A decision rule ϕ_t, for a traditional bandit problem (without side observations) with finite $\boldsymbol{\Theta}$, 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 $T_{\mathrm{inf}}(t)$ in Theorem 1.1. The detailed construction of a tight ϕ_t can be found in [3]. In this subsection, the tight ϕ_t's (for traditional bandit problems) will serve as constituent components in a composite decision rule Φ_t dealing with the side-observation-aided bandit problems.

Suppose $|\mathbf{X}| = k < \infty$. Using the values of X_t, we can partition the observed rewards Y_t^1 (or Y_t^2) into k sub-sequences, corresponding to different x's. Consider the sub-sequence obtained when $X_t = x_0$. At those time instants, the decision maker is facing $F_{\theta_1}(\cdot|x_0)$ and $F_{\theta_2}(\cdot|x_0)$, and thus this sub-sequence can be viewed as resulting from a traditional bandit problem with the family of possible distributions being $\{F_\theta(\cdot|x_0)\}_\theta$. For each x_0, we use B_{x_0} to denote the corresponding sub-bandit problem.

For example, if $X_1X_2X_3X_4\cdots = x_ax_bx_ax_c\cdots$, then after time $t = 4$, we have 2 samples in B_{x_a}, 1 sample in B_{x_b}, and 1 sample in B_{x_c}. One straightforward composite decision rule Φ_t is to apply a tight $\phi_{x,t}$ on each sub bandit B_x. The resulting composite rule is uniformly good but does not yield sharp results matching the new $\log t$ lower bound in Eq. (5.1).

Let $\hat{C}_{x,t}$ denote the corresponding estimates of the tight constituent $\phi_{x,t}$. A more sophisticated composite rule Φ_t for the side-observation-aided bandits is constructed as in Algorithm 3, and is asymptotically optimal.

Theorem 5.2 (Asymptotic optimality). *Suppose for all possible C_0, $M_{C_0}(x)$ does not vary with respect to x. With the regularity conditions* R1, R2, R4, *and assumptions* A1, A2,

Algorithm 3 (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 $\mathsf{B}_x$. 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
```

A tie-breaking mechanism is necessary while evaluating "arg max" in line 5, and a natural choice of a randomized tie-breaking mechanism is sufficient for rigorous analysis. However, to minimize the distraction of this minor point, we assume here that no tie exists during the execution of this algorithm.

A3, *the composite*[11] *rule* Φ_t *described in Algorithm* 3 *achieves the* $\log t$ *lower bound in Eq.* (5.1), *that is,*

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

This Φ_t *is thus asymptotically optimal.*

The intuition behind this result is that having different side information values x is like having several related bandit machines B_x's. Each B_x has its own reward distribution pair $(F_{\theta_1}(\cdot|x), F_{\theta_2}(\cdot|x))$, but all these B_x's share the same common, but unknown, configuration pair (θ_1, θ_2). The information obtained from one machine is thus applicable to the other machines. If arm 2 is always better than arm 1, we wish to sample arm 2 most of the time (the control part), and force sample arm 1 once in a while (the learning part). With the help of the side information X_t, we can postpone our forced sampling (learning) to the most informative machine $X_t = x^* = \arg\max_x \inf_{\{\theta:\ \theta > \hat{\theta}_2\}} I(\theta_1, \theta|x)$. With the assumption of the existence of the value of the game, this composite Φ_t thus achieves the new constant in the $\log t$ lower bound. A detailed analysis of this case is provided in Appendix C.

6. Mixed case

It is worth noting that the main difference between Sections 4 and 5 is that in one case, X_t *always* changes the preference order, while in the other case, X_t *never* changes the order. A much more general case is a mixture of these two cases, which will be discussed in this section and which leads to the main result of this paper.

Definition 6.1 (*Mixed condition*). As illustrated in Fig. 3, for some $C \in \mathbf{\Theta}^2$, $M_C(x)$ is not a function of x, i.e., $M_C(x) := M_C$. For the remaining C, there exist x_1 and x_2 such that $M_C(x_1) = 1$ and $M_C(x_2) = 2$. For future reference, if such x_1 and x_2 exist, we say the configuration pair C_0 is *implicitly revealing*.

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

[11] To perform a rigorous analysis, the constituent $\phi_{x,t}$ must be fully encapsulated in Algorithm 3. Namely, only those samples obtained from performing $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$ (lines 2 and 8) can be counted as valid samples for $\phi_{x,t}$. In other words, the time instants when we let $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ (line 6) must be excluded from the computation of $\hat{C}_{x,t}$ and $\phi_{x,t+1}$. Otherwise it may spoil the tightness of the original $\phi_{x,t+1}$. For example, suppose $X_1X_2X_3X_4\cdots = x_a x_b x_a x_c \cdots$. At time instants 1 and 2, we have executed $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$, while at time instants 3 and 4, $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ is executed. Then from the sub-bandit-problem point of view, we have only one sample in B_{x_a}, one sample in B_{x_b}, and no samples in B_{x_c}, and only those samples can be used to generate the corresponding value of $\hat{C}_{x,t}$ and $\phi_{x,t+1}$. Samples made at time instants 3 and 4 will be discarded.

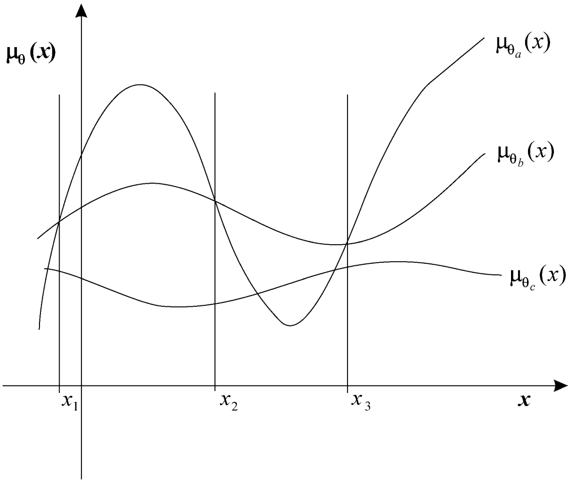

Fig. 3. For some (θ_1, θ_2) the best arm is a function of x, i.e., $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ intersect each other as in Section 4. For the remaining (θ_1, θ_2) the best arm is not a function of x, i.e., $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ do not intersect each other as first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

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 $\lim_{t\to\infty} \mathsf{E}\{T_{\inf}(t)\} < \infty$ 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 following two regularity conditions R1 and R2:

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

we can achieve this goal.

6.1. Lower bound

Similar to Theorem 5.2, a $\log t$ lower bound on $\mathsf{E}\{T_{\inf}(t)\}$ is obtained for uniformly good rules, and is formally stated as follows.

Theorem 6.1. *Suppose the side observation sequence $\{X_\tau\}$ is evenly distributed in L^1, and the mixed condition in Definition* 6.1, *and regularity conditions* R1 *and* R2 *are satisfied. For any uniformly good rule, if the authentic parameter pair C_0 is not implicitly revealing then $\mathsf{E}_{C_0}\{T_{\inf}(t)\}$ is $\log t$ lower bounded*:

$$\lim_{t\to\infty} \mathsf{P}_{C_0}\left(T_{\inf}(t) \geqslant \frac{(1-\varepsilon)\log t}{K_{C_0}}\right) = 1, \quad \forall \varepsilon > 0, \quad \textit{and}$$

$$\liminf_{t\to\infty} \frac{\mathsf{E}_{C_0}\{T_{\inf}(t)\}}{\log t} \geqslant \frac{1}{K_{C_0}}, \tag{6.1}$$

where K_{C_0} is a constant depending on C_0. If $M_{C_0} = 2$, we have $T_{\inf}(t) = T_1(t)$, and the constant K_{C_0} is

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

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

[23] proved a similar version of Theorem 6.1 for i.i.d. $\{X_\tau\}$, and one can easily modify the proof there by exploiting the assumption that $\{X_\tau\}$ is evenly distributed in L^1.

6.2. Scheme achieving the lower bound

With the following three assumptions:

A1: the parameter space $\boldsymbol{\Theta}$ is finite,
A2: the side observations $\{X_\tau\}$ are u.s.e. distributed in L^1,
A4: the value of the game,

$$\inf_{\{\theta:\ \exists x_0,\ \mu_\theta(x_0) > \mu_{\theta_2}(x_0)\}} \sup_{x \in \mathbf{X}} \big\{ I(\theta_1, \theta | x) \big\} = \sup_{x \in \mathbf{X}} \inf_{\{\theta:\ \mu_\theta(x) > \mu_{\theta_2}(x)\}} \big\{ I(\theta_1, \theta | x) \big\},$$

exists,

we are able to construct schemes achieving bounded $\lim_{t\to\infty} \mathsf{E}\{T_{\inf}(t)\} < \infty$ when being applied to implicitly revealing C_0 or otherwise achieving the $\log t$ lower bound in Theorem 6.1. One instance is the composite control scheme Φ_t described in Algorithm 4, the details of which are described in the following paragraphs.

The sub-bandit machines B_x, the corresponding tight decision rules ϕ_t, and the estimate $\hat{C}_{x,t}$ are as defined in Section 5.2, along with a number of newly-introduced counters (actually $|X| \cdot |\boldsymbol{\Theta}^2|^2$ counters). These new counters are named $\mathsf{ctr}(x, C', C'')$ and are initially set to zero. The $\ddot{C}_t$ used in Algorithm 4 is an estimate of C_0 generated from the sampling when $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ is active, namely, when line 10, 14, or 19 is executed. On the other hand, those samples when $\Phi_{t+1} \leftarrow \phi_{x,t+1}$ is active, namely, when line 2, 8, or 21 being executed, are used to generate $\hat{C}_{x,t}$ and $\phi_{x,t+1}$.

For example, suppose $X_1 X_2 X_3 X_4 \cdots = x_a x_b x_a x_c \cdots$ and at time instants 1 and 2, $\Phi_{t+1} \leftarrow \phi_{x,t+1}$ (lines 2, 8, 21), while at time instants 3 and 4, $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$ (lines 10, 14, 19). As a result, after four pulls of the bandit machine, we have one sample in B_{x_a} to generate $\hat{C}_{x_a,4}$, one sample in B_{x_b} for $\hat{C}_{x_b,4}$, and no samples in B_{x_c} for $\hat{C}_{x_c,4}$. At the same time, we have a total of one sample in B_{x_a}, no samples in B_{x_b} and one sample in B_{x_c} being used to generate $\ddot{C}_4$.

We will prove that with any "good" $\ddot{C}_t$, Algorithm 4 will result in a bound-achieving Φ_t. The definition of a "good" $\ddot{C}_t$ is as follows.

Definition 6.2 (*Good estimate $\ddot{C}_t$*). An estimate $\ddot{\theta}$ is good if there exist $a, b > 0$ such that the mis-detection probability $\mathsf{P}_\theta(\ddot{\theta} \neq \theta) \leqslant a \exp(-bN)$, where N is the number of samples

Algorithm 4 (Φ_{t+1}, *the decision at time* $t + 1$)

```
if not all $\hat{C}_{x,t}$ are identical, then
  $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$.
else
  Denote $\hat{C}_t = (\hat{\theta}_1, \hat{\theta}_2)$ as the common estimate for all $\mathsf{B}_x$.
  if $\hat{C}_t$ is implicitly revealing, then
    if $\ddot{C}_t \neq \hat{C}_t$ (including the cases that $\ddot{C}_t$ does not exist), then
      if ctr$(X_{t+1}, \hat{C}_t, \ddot{C}_t)$ is even, then
        $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$.
      else
        $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$.
      end if
      ctr$(X_{t+1}, \hat{C}_t, \ddot{C}_t) \leftarrow$ ctr$(X_{t+1}, \hat{C}_t, \ddot{C}_t) + 1$.
    else
      $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$.
    end if
  else
    Without loss of generality, we may assume $M_{\hat{C}_t} = 2$. The case in which $M_{\hat{C}_t} = 1$
    can be obtained by symmetry.
    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
      $\Phi_{t+1} \leftarrow M_{\hat{C}_t}(X_{t+1})$.
    else
      $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t+1}$.
    end if
  end if
end if
```

that $\ddot{\theta}$ is based upon. An estimate pair $\ddot{C}_t = (\ddot{\theta}_1, \ddot{\theta}_2)$ is *good* if $\ddot{\theta}_1$ and $\ddot{\theta}_2$ are good estimates for θ_1 and θ_2 respectively.[12]

Theorem 6.2 (Asymptotic optimality). *Suppose the mixed condition in Definition* 6.1*, the regularity conditions* R1 *and* R2*, and the assumptions* A1, A2*, and* A4 *are satisfied. With the tight constituent* $\phi_{x,t}$*, and a good estimate* $\ddot{C}_t$*, the* Φ_t *described in Algorithm* 4 *either has bounded inferior sampling time, or achieves the* $\log t$ *lower bound in Theorem* 6.1*, depending on whether or not the unknown underlying configuration pair* C_0 *is implicitly revealing.*

The intuition behind Theorem 6.2 is exactly the mixture of our previous discussions on the pure cases. When the unknown C_0 is implicitly revealing, the evenly distributed side information X_t will direct the player to sample both arms often enough, which leads

[12] By the large deviations principle and the regularity condition R2, a good estimate $\ddot{C}_t$ generally exists.

to bounded $\mathsf{E}\{T_{\inf}(t)\}$. If the underlying C_0 is not implicitly revealing, then postponing the forced sampling will reduce the constant $1/K_{C_0}$ in front of the $\log t$ lower bound. A detailed proof is given in Appendix D.

7. A note on the necessity of the even distribution properties

In Sections 4 through 6, we have discussed the benefits of having side observations under various situations. The main results are summarized in Table 1. Since all of the given conditions are sufficient, the question naturally arises as to whether these evenly distributed properties are necessary for the various levels of improvement.

Table 1
Summary of the relationships between X_t and Y_t^i

Characterization	Regularity conditions	Even distribution conditions	Results for all $C_0 \in \mathbf{\Theta}^2$
$\forall C_1 \neq C_2$, $G_{C_1} \neq G_{C_2}$	As $\hat{C}_t \to C_0$, $\forall x$, $M_{\hat{C}_t}(x) = M_{C_0}(x)$		$\exists\{\phi_\tau\}$ such that $\lim \mathsf{E}_{C_0}\{T_{\inf}(t)\}/\sum \mathsf{P}(\lvert\hat{C}_\tau - C_0\rvert > \varepsilon) \leqslant 1$
All $C_0 \in \mathbf{\Theta}^2$ have $G_{C_0} = G$, and are implicitly revealing	(i) $\mathbf{X}$ is finite, (ii) $\forall \theta_1 \neq \theta_2, x$, $I(\theta_1, \theta_2 \mid x) > 0$, (iii) $\forall x$, $\mu_\theta(x)$ is continuous w.r.t. θ	$\{X_\tau\}$ is evenly distributed in probability series	$\exists\{\phi_\tau\}$ such that $\lim \mathsf{E}_{C_0}\{T_{\inf}(t)\} < \infty$
$\forall C_0 \in \mathbf{\Theta}^2$, $G_{C_0} = G$; no $C_0 \in \mathbf{\Theta}^2$ is implicitly revealing	(i) $\mathbf{X}$ is finite, (ii) $\forall \theta_1 \neq \theta_2, x$, $I(\theta_1, \theta_2 \mid x) > 0$		The performance of any uniformly good $\{\phi_\tau\}$ is lower bounded by $\lim \mathsf{E}_{C_0}\{T_{\inf}(t)\}/\log t \geqslant 1/K_{C_0}$, where $K_{C_0} \triangleq \inf_\theta \sup_x I(\theta_1, \theta \mid x)$
	(i), (ii), and (iii) $\mathbf{\Theta}$ 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 $\lim \mathsf{E}_{C_0}\{T_{\inf}(t)\}/\log t \leqslant 1/K_{C_0}$, namely $\{\phi_\tau\}$ achieves the lower bound
$\forall C_0 \in \mathbf{\Theta}^2$, $G_{C_0} = G$; in $\mathbf{\Theta}^2$, some C_0 are implicitly revealing and some are not	(i) $\mathbf{X}$ is finite, (ii) $\forall \theta_1 \neq \theta_2, x$, $I(\theta_1, \theta_2 \mid x) > 0$	$\{X_\tau\}$ is evenly distributed in L^1	If C_0 is not implicitly revealing, the performance of any uniformly good $\{\phi_\tau\}$ is lower bounded by $\lim \mathsf{E}_{C_0}\{T_{\inf}(t)\}/\log t \geqslant 1/K_{C_0}$, where $K_{C_0} \triangleq \inf_\theta \sup_x I(\theta_1, \theta \mid x)$
	(i), (ii), and (iii) $\mathbf{\Theta}$ 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 implicitly revealing (i.r.), $\lim \mathsf{E}_{C_0}\{T_{\inf}(t)\} < \infty$; if C_0 is not i.r., $\{\phi_\tau\}$ achieves the lower bound: $\lim \mathsf{E}_{C_0}\{T_{\inf}(t)\}/\log t \leqslant 1/K_{C_0}$

Suppose $\{X_\tau\}$ reveals no information about C_0, as in Sections 4 to 6. We need some minimal amount of even distributedness to guarantee the benefit of a side observation sequence, as the following result states.

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*

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

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

If there exists x_0 such that $\mathsf{P}(\exists \tau,\ \text{s.t. } X_\tau = x_0) = 0$, then the range of the side observation can be reduced to the positive support of X_t. The benefit of the characterization properties (helpful structure between X_t, Y_t^i) may degenerate to another case with new support $\mathbf{X}' = \mathbf{X} \backslash \{x_0\}$, which severely affects the attainable results. Take Example 1 in Section 1.3 for example, the implicitly revealing $C_0 = (\theta_1, \theta_2) = (1, 2)$ is no longer implicitly revealing if the support $\mathbf{X} = \{1, 2, 3\}$ is reduced to $\{1, 2\}$. The achievable $\mathsf{E}\{T_{\text{inf}}(t)\}$ thus becomes $\mathcal{O}(\log t)$ lower bounded, rather than $\mathcal{O}(\text{constant})$. Theorem 7.1 shows that the benefit of side observations indeed comes from the even distribution properties.

8. Conclusions

It has been shown in [23] that observing additional i.i.d. side information can 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 processes and proved their efficacy for bandit problems. When the side observation X_t reveals information about C_0, 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 $\lim_{t\to\infty} \mathsf{E}\{T_{\text{inf}}(t)\} < \infty$ for many types of $\{X_\tau\}$.

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 [23] that for i.i.d. $\{X_\tau\}$, case (1) leads to bounded expected inferior sampling time, case (2) leads to asymptotically sharp $\log t$ lower bound, and case (3) leads to $\log t$ lower bound for some C_0, and bounded expected inferior sampling time for other C_0. And in this paper (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, including Markov chains, and all 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 the Kullback–Leibler divergence.

Finally, we have also provided a simple necessary condition, namely $\forall x$, $\mathsf{P}(\exists\tau$, 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.

Appendix A. Proof of Theorem 3.1

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

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

Let $\overline{\mathbf{C}}_e$ denote the closure of $\mathbf{C}_e$. By Condition 3.1, we have that C_0 is not in $\overline{\mathbf{C}}_e$ and there exists $\varepsilon > 0$ such that $\overline{\mathbf{C}}_e \subseteq \{C\colon |C - C_0| > \varepsilon\}$. For any $t \geqslant 1$,

$$\begin{aligned}
\mathsf{P}_{C_0}\bigl(\phi_t \neq M_{C_0}(X_t)\bigr) &= \mathsf{P}_{C_0}\bigl(M_{\hat{C}_t}(X_t) \neq M_{C_0}(X_t)\bigr)\\
&\leqslant \mathsf{P}_{C_0}\bigl(\exists x, M_{\hat{C}_t}(x) \neq M_{C_0}(x)\bigr)\\
&= \mathsf{P}_{C_0}\bigl(\hat{C}_t \in \mathbf{C}_e\bigr)\\
&\leqslant \mathsf{P}_{C_0}\bigl(\hat{C}_t \in \overline{\mathbf{C}}_e\bigr)\\
&\leqslant \mathsf{P}_{C_0}\bigl(\bigl|\hat{C}_t - C_0\bigr| > \varepsilon\bigr),
\end{aligned}$$

and

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

This completes the proof.

Appendix B. Proof of Theorem 4.1

We define $\mathbf{C}_e$ similarly to (A.1). The necessary result [23, Lemma 1, p. 23] is quoted as follows.

Lemma B.1 [23, Lemma 1, p. 23]. *With the regularity conditions specified in Section* 4, $\exists a_1, a_2 > 0$ *such that*

$$\mathsf{P}_{C_0}\bigl(\hat{C}_t \in \mathbf{C}_e\bigr) \leqslant a_1 \exp\bigl(-a_2 \min\bigl\{T_1^{x^\star}(t), T_2^{x^\star}(t)\bigr\}\bigr).$$

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

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

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

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

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

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:

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

From the definition, we have $T_1(t) = N_1 - N_{1\to 2} + N_{2\to 1}$. 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,

$$\begin{aligned} N_{1\to2}+N_{2\to1} &= \sum_{s\in[1,\sqrt{t}]} 1\{\phi_s \neq M_{C_0}(X_s)\} + \sum_{s\in[\sqrt{t}+1,t]} 1\{\phi_s \neq M_{C_0}(X_s)\} \\ &\leqslant \sqrt{t} + \sum_{s\in[\sqrt{t}+1,t]} 1\{\phi_s \neq M_{\hat{C}_{s-1}}(X_s)\} \\ &\leqslant 2\sqrt{t}+1. \end{aligned} \tag{B.3}$$

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 $\forall i$, $T_i(t) \geqslant \sqrt{t}$, it follows that $T_1(t) = N_1 - N_{1\to2} + N_{2\to1} = \sqrt{t}$. Combining the result in (B.3), we conclude that

$$\begin{aligned} B^{1.1} &\subseteq \{N_1 \leqslant 3\sqrt{t}+1\} \\ &= \left\{ \sum_{s\in[1,t]} 1\{M_{C_0}(X_s)=1\} \leqslant 3\sqrt{t}+1 \right\}. \end{aligned} \tag{B.4}$$

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

$$\begin{aligned} \mathsf{P}(B^1_{t+1}) &\leqslant \left(\sum_{s\in[\sqrt{t},t-1]} \mathsf{P}(\hat{C}_s \in \mathbf{C}_e) \right) + \mathsf{P}(B^{1.1}) \\ &\leqslant \sum_{s\in[\sqrt{t},t-1]} a_1 e^{-a_2\sqrt{s}} + \mathsf{P}\left(\frac{\sum_{s\in[1,t]} 1\{X_s \in \mathbf{X}^1_{C_0}\}}{t} \leqslant \frac{3\sqrt{t}+1}{t} \right), \end{aligned} \tag{B.5}$$

where the second inequality follows from the application of Lemma B.1 to the first term, while the second term follows from (B.4). By simple algebra, we have

$$\sum_{t+1=7}^{\infty} \sum_{s\in[\sqrt{t},t-1]} a_1 e^{-a_2\sqrt{s}} < \infty. \tag{B.6}$$

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

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

From (B.5), (B.6), and (B.7), we conclude

$$\sum_{t+1=7}^{\infty} \mathsf{P}(B_{t+1}) \leqslant \sum_{t+1=7}^{\infty} \left(\mathsf{P}\left(B_{t+1}^1\right) + \mathsf{P}\left(B_{t+1}^2\right)\right) < \infty,$$

and by (B.1),

$$\lim_{t\to\infty} \mathsf{E}\left\{T_{\mathrm{inf}}(t)\right\} \leqslant 6 + \sum_{t+1=7}^{\infty} \left(\mathsf{P}(A_{t+1}) + \mathsf{P}(B_{t+1})\right) < \infty,$$

which completes the analysis.

Appendix C. Proof of Theorem 5.2

We need the following lemma for the later proof.

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

$$\mathsf{sum} := \sum_{j=1}^{\infty} (T_j - S_j + 1) \quad \textit{and} \quad U := \sup\{j \in \mathbb{N}\colon\ S_j < \infty\}.$$

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

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

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:

$$\begin{aligned}
&\mathsf{sum} = \sum_{j=1}^{\infty} 1\{S_j < \infty\}(T_j - S_j + 1) \\
\Rightarrow\quad &\mathsf{E}\{\mathsf{sum}\} = \sum_{j=1}^{\infty} \mathsf{E}\left\{1\{S_j < \infty\} \cdot \mathsf{E}\{T_j - S_j + 1 | S_j\}\right\} \\
&\leqslant \sum_{j=1}^{\infty} B \cdot \mathsf{E}\left\{1\{S_j < \infty\}\right\} \\
&= B \sum_{j=1}^{\infty} \mathsf{P}(U \geqslant j) = B \cdot K. \qquad \square
\end{aligned}$$

With the help of Lemma C.1, we prove Theorem 5.2 by making the following arguments.

Argument 1. *The expected duration over which $\hat{C}_t$ does not exist is finite, i.e.,*

$$\lim_{t\to\infty} \mathsf{E}\left\{\sum_{\tau=1}^{t} 1\left\{\hat{C}_\tau \textit{ does not exists}\right\}\right\} < \infty.$$

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

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

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

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 $\log t / K_{C_0}$, i.e.,*

$$\lim_{t\to\infty} \frac{\mathsf{E}\{\sum_{\tau=1}^{t} 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\}\}}{\log t} \leqslant \frac{1}{K_{C_0}},$$

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

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

and

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

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 disjoint[13] intervals, with x specifying the value of the side observation X_t at the leading time instant $S_{x,j}$. We then have

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

Since

$$T_{x,j} \leqslant \inf\Biggl\{t>S_{x,j}\colon \mathbf{X}=\bigcup_{s\in(S_{x,j},t]}\{X_s\}\Biggr\},$$

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

$$\forall x,\quad \mathsf{E}\bigl\{\sup\{j\in\mathbb{N}\colon\ S_{x,j}<\infty\}\bigr\}<\infty, \tag{C.1}$$

then by Lemma C.1, we have

$$\mathsf{E}\Biggl\{\sum_{t=1}^{\infty} 1\bigl\{1\bigl\{\hat{C}_t\bigr\}=0\bigr\}\Biggr\}<\infty.$$

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.

- $\hat{C}_{x,t}\neq C_0$:
  - 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 $\sup\{j\in\mathbb{N}\colon\ S_{x,j}<\infty\}\leqslant\sup\{j\in\mathbb{N}\colon\ T_{x,j}<\infty\}+1$, 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 $\hat{C}_{x,T_{x,j}}\neq\hat{C}_{x,t}=C_0$.

[13] In some cases, the intervals may overlap with each other, but the overlap can only happen at the end points, which does not affect the validity of the proof.

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

$$\sup\{j \in \mathbb{N}\colon\ T_{x,j} < \infty\} \leqslant \sum_{\tau=1}^{\infty} 1\bigl\{X_\tau = x,\ \Phi_\tau \leftarrow \phi_{x,\tau},\ \hat{C}_{x,\tau} \neq C_0\bigr\} \\ + \sum_{x'\colon x' \neq x} \sum_{\tau=1}^{\infty} 1\bigl\{X_{\tau+1} = x',\ \Phi_{\tau+1} \leftarrow \phi_{x',\tau+1},\ \hat{C}_{x',\tau} \neq C_0\bigr\}.$$

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

From the previous discussions, we have proved $\mathsf{E}\{\sup\{j \in \mathbb{N}\colon\ S_{x,j} < \infty\}\} < \infty$ and Argument 1. □

Proof of Argument 2. Consider a fixed $C' := (\theta'_1, \theta'_2) \neq C_0$ and set

$$x^* := \arg\max_x \inf_{\{\theta\colon\ \theta > \theta'_2\}} I\bigl(\theta'_1, \theta|x\bigr).$$

We then iteratively define the stopping time pairs $S_{C',j}$ and $T_{C',j}$ as follows.

$$S_{C',j} := \inf\bigl\{t > S_{C',j-1}\colon\ \hat{C}_t = C',\ \text{ and either } 1\bigl\{\hat{C}_{t-1}\bigr\} = 0, \\ \text{or } \hat{C}_{t-1} \neq C',\ \text{ or } X_t = x^*\bigr\},$$

and

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

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

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

Since

$$T_{C',j} \leqslant \inf\bigl\{t > S_{C',j}\colon\ X_{t+1} = x^*\bigr\},$$

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

$$\forall x, \quad \mathsf{E}\bigl\{\sup\{j \in \mathbb{N}\colon\ S_{C',j} < \infty\}\bigr\} < \infty,$$

then by Lemma C.1, we have $\mathsf{E}\{\sum_{t=1}^{\infty} 1\{\hat{C}_t \neq C_0\}\} < \infty$.

By observing $\sup\{j \in \mathbb{N}\colon S_{C',j} < \infty\} \leqslant \sup\{j \in \mathbb{N}\colon T_{C',j} < \infty\} + 1$, 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.

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

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

$$\sup\{j \in \mathbb{N}\colon T_{C',j} < \infty\} \leqslant \sum_{t=1}^{\infty} 1\bigl\{X_{t+1} = x^*,\ \Phi_{t+1} \leftarrow \phi_{x^*,t+1},\ \hat{C}_{x^*,t} \neq C_0\bigr\}.$$

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 $x^* = \arg\max_x \inf_{\{\theta\colon \theta > \theta_2\}} I(\theta_1, \theta|x)$. We then have

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

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

$$\lim_{t\to\infty} \frac{\mathsf{E}\bigl\{\sum_{\tau=1}^{t} 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\}\bigr\}}{\log t} \leqslant \frac{1}{K_{C_0}},$$

where

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

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

Appendix 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,

$$\lim_{t\to\infty} \mathsf{E}\left\{\sum_{\tau=1}^{t} 1\{\hat{C}_\tau \text{ does not exists}\}\right\} < \infty.$$

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,

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

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,

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

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 $\log t/K_{C_0}$, namely,

$$\lim_{t\to\infty} \frac{\mathsf{E}\{\sum_{\tau=1}^{t} 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\}\}}{\log t} \leqslant \frac{1}{K_{C_0}},$$

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 4 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 $\hat{C}_t = \ddot{C}_t$, we have

$$\sum_{t=1}^{\infty} 1\bigl\{\hat{C}_t = C' \neq C_0\bigr\} = \sum_{t=1}^{\infty} 1\bigl\{\hat{C}_t = C' \neq C_0,\ \ddot{C}_t \neq \hat{C}_t\bigr\} + \sum_{t=1}^{\infty} 1\bigl\{\hat{C}_t = C' \neq C_0,\ \ddot{C}_t = \hat{C}_t\bigr\}.$$

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.

$$S_{x,C',C'',j} := \inf\bigl\{t > S_{x,C',C'',j-1}\colon X_{t+1} = x,\ \hat{C}_t = C',\ \ddot{C}_t = C'',\ \text{and either } \hat{C}_{t-1} \neq C',\ \text{or } \ddot{C}_{t-1} \neq C'',\ \text{or } X_t \neq x\bigr\},$$

and

$$T_{x,C',C'',j} := \inf\bigl\{t > S_{x,C',C'',j}\colon \text{either } \hat{C}_t \neq C',\ \text{or } \ddot{C}_t \neq C'',\ \text{or } X_{t+1} = x\bigr\},$$

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 $\{\hat{C}_t = C', \ddot{C}_t = C''\}$ 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

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

Since

$$T_{x,C',C'',j} \leqslant \inf\bigl\{t > S_{x,C',C'',j}\colon X_{t+1} = x\bigr\},$$

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

$$\forall x, C'',\ \exists K, \quad \mathsf{E}\bigl\{\sup\{j \in \mathbb{N}\colon S_{x,C',C'',j}\}\bigr\} < K,$$

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

Let $t := S_{x,C',C'',j}$. By the definition of Algorithm 4, for odd j the decision rule results in $\Phi_{t+1} \leftarrow \phi_{X_{t+1},t}$ (since at time t, $\operatorname{ctr}(x, C', C'') = j - 1$). Thus we have

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

$$\leqslant 2\sum_{\tau=1}^{\infty} 1\bigl\{X_{t+1} = x,\ \hat{C}_t = C' \neq C_0,\ \ddot{C} = C'',\ \Phi_{t+1} \leftarrow \phi_{x,t+1}\bigr\}.$$

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.

$$S_{x,C',j} := \inf\Biggl\{t > S_{x,C',j-1}\colon\ X_{t+1} = x,\ \hat{C}_t = \ddot{C}_t = C',\ \text{and either } \hat{C}_{t-1} \neq C',$$

$$\text{or } \ddot{C}_{t-1} \neq C',\ \text{or } \{1,2\} = \bigcup_{s \in (S_{x,C',j-1},t]} \bigl\{M_{C'}(X_s)\bigr\}\Biggr\},$$

and

$$T_{x,C',j} := \inf\Biggl\{t > S_{x,C',j}\colon\ \text{either } \hat{C}_t \neq C',\ \text{or } \ddot{C}_t \neq C',$$

$$\text{or } \{1,2\} = \bigcup_{s \in (S_{x,C',j},t]} \bigl\{M_{C'}(X_s)\bigr\}\Biggr\},$$

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

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

Since

$$T_{x,C',j} \leqslant \inf\Biggl\{t > S_{x,C',j}\colon\ \mathbf{X} = \bigcup_{s \in (S_{x,C',j},t]} \{X_s\}\Biggr\},$$

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

$$\forall x \in \mathbf{X},\quad \mathsf{E}\bigl\{\sup\{j \in \mathbb{N}\colon\ S_{x,C',j} < \infty\}\bigr\} < \infty, \tag{D.1}$$

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

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'. Recalling that $1(C)$ denotes the first coordinate of the configuration pair C, we 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

$$\begin{aligned}&\sup\{j \in \mathbb{N}\colon\ S_{x,C',j} < \infty\}\\&\quad = \sum_{j=1}^{\infty} 1\{S_{x,C',j} < \infty\}\\&\quad \leqslant \sum_{\tau=1}^{\infty} 1\bigl\{X_{t+1} = x,\ 1\bigl(\ddot{C}_t\bigr) = 1\bigl(C'\bigr) \neq 1(C_0),\ \Phi_{t+1} \leftarrow M_{C'}(x) = 1\bigr\} \triangleq D_1.\end{aligned}$$

  Since every time the event $\{X_{t+1} = x, 1(\ddot{C}_t) = 1(C') \neq 1(C_0), \Phi_{t+1} \leftarrow M_{C'}(x) = 1\}$ occurs, the effective sample size of arm 1 (used to generate $\ddot{C}_t$) increases by one. Because $\ddot{C}_t$ 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 condition implies that $2(C') \neq 2(C_0)$. By noting that $\sup\{j \in \mathbb{N}\colon\ S_{x,C',j} < \infty\} \leqslant \sup\{j \in \mathbb{N}\colon\ T_{x,C',j} < \infty\} + 1$, we prove that the latter can have at most finite expectation. If $T_{x,C',j} < \infty$, it follows that we have either $M_{C'}(X_{T_{x,C',j}}) = 2$ or $1(\ddot{C}_{T_{x,C',j}}) \neq 1(C_0)$. As a result,

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

  Since the estimate $\ddot{C}_t$ 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 $\hat{C}_t \neq \ddot{C}_t$, and $\hat{C}_t = \ddot{C}_t$ 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 $\ddot{C}_t \neq \hat{C}_t = C_0$. For any fixed $C' \neq C_0$, construct the stopping time pairs as follows.

$$S_{C',j} := \inf\Biggl\{ t > S_{C',j-1}\colon\ \hat{C}_t = C_0,\ \ddot{C}_t = C',\ \text{and either } \hat{C}_{t-1} \neq C_0,\ \text{or } \ddot{C}_{t-1} \neq C',\\ \text{or } \{1,2\} = \bigcup_{s\in\mathbf{S}_{j-1,t-1}} \bigl\{M_{C_0}(X_s)\bigr\}\Biggr\},$$

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

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

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

$$T_{C',j} := \inf\Biggl\{t > S_{C',j}\colon\ \text{either } \hat{C}_t \neq C_0, \text{or } \ddot{C}_t \neq C', \text{or } \{1,2\} = \bigcup_{s\in\mathbf{S}_{j,t}} \bigl\{M_{C_0}(X_s)\bigr\}\Biggr\}.$$

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

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

By line 7 in Algorithm 4, for any $X_{t+1} = x$, $\hat{C}_t = C_0$, $\ddot{C}_t = C'$, the decision rule Φ_{t+1} is alternating between $\phi_{x,t}$ and $M_{C_0}(x)$. As a result, we have

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

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

$$\forall x \in \mathbf{X}, C', \quad \mathsf{E}\bigl\{\sup\{j \in \mathbb{N}\colon\ S_{C',j} < \infty\}\bigr\} < \infty,$$

then by Lemma C.1, we have $\mathsf{E}\{\sum_{t=1}^{\infty} 1\{\hat{C}_t = C_0 \neq \ddot{C}_t\}\} < \infty$.

Since $\sup\{j\colon\ S_{C',j} < \infty\} \leqslant \sup\{j\colon\ T_{C',j} < \infty\} + 1$, equivalently, we can focus on proving $\mathsf{E}\{\sup\{j \in \mathbb{N}\colon\ T_{C',j} < \infty\}\} < \infty$. 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 $\ddot{C}_t$ must have increased by 1. Since $\ddot{C}_t$ is a good estimate, this portion contributes at most bounded expectation.

- $\ddot{C}_{t'} \neq C'$: 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 $\ddot{C}_t$ must increase by 1 during the interval $[S_{C',j}, T_{C',j}]$. By the assumption that $\ddot{C}_t$ 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 $\ddot{C}_t$) increases by at least one. We also note that $\ddot{C}_{t'} \neq \ddot{C}_{t'-1} = C'$ and $1(\ddot{C}_{t'}) \neq 1(C_0)$. Combining the above observations and the assumption that $\ddot{C}_t$ is good, this portion can contribute at most bounded expectation.

From the above discussions, we have

$$\mathsf{E}\left\{\sum_{t=1}^{\infty} 1\{\hat{C}_t = C_0 \neq \ddot{C}_t\}\right\} < \infty. \qquad \square$$

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

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

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

$$\lim_{t\to\infty} \frac{\mathsf{E}\{\sum_{\tau=1}^{t} 1\{\hat{C}_\tau = C_0, \Phi_{\tau+1} \neq M_{C_0}(X_{\tau+1})\}\}}{\log t} \leqslant \frac{1}{K_{C_0}},$$

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. $\square$

References

[1] K. Adam, Learning while searching for the best alternative, J. Economic Theory 101 (2001) 252–280.

[2] R. Agrawal, M.V. Hegde, D. Teneketzis, Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost, IEEE Trans. Automat. Control 33 (10) (1988) 899–906.

[3] R. Agrawal, D. Teneketzis, V. Anantharam, Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: finite parameter space, IEEE Trans. Automat. Control 34 (3) (1989) 258–267.

[4] R. Agrawal, D. Teneketzis, V. Anantharam, Asymptotically efficient adaptive allocation schemes for controlled Markov chains: finite parameter space, IEEE Trans. Automat. Control 34 (12) (1989) 1249–1259.

[5] V. Anantharam, P. Varaiya, J. Walrand, Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—part I: i.i.d. rewards, IEEE Trans. Automat. Control 32 (11) (1987) 968–976.

[6] V. Anantharam, P. Varaiya, J. Walrand, Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—part II: Markovian rewards, IEEE Trans. Automat. Control 32 (11) (1987) 977–982.

[7] D.A. Berry, A Bernoulli two-armed bandit, Ann. Math. Stat. 43 (3) (1972) 871–897.

[8] J.A. Bucklew, Large Deviation Techniques in Decision, Simulation, and Estimation, Wiley, New York, 1990.

[9] H. Chernoff, Sequential Analysis and Optimal Design, SIAM, Philadelphia, PA, 1972.

[10] M.K. Clayton, Covariate models for Bernoulli bandits, Sequential Anal. 8 (4) (1989) 405–426.

[11] A. Dembo, O. Zeitouni, Large Deviation Techniques and Applications, Springer, New York, 1998.

[12] J.C. Gittins, Bandit processes and dynamic allocation indices, J. Roy. Statist. Soc. Ser. B (Methodological) 41 (2) (1979) 148–177.

[13] J.C. Gittins, A dynamic allocation index for the discounted multiarmed bandit problem, Biometrika 66 (3) (1979) 561–565.

[14] B.K. Ghosh, P.K. Sen, Handbook of Sequential Analysis, Dekker, New York, 1991.

[15] S.R. Kulkarni, G. Lugosi, Finite-time lower bounds for the two-armed bandit problem, IEEE Trans. Automat. Control 45 (4) (2000) 711–714.

[16] M.N. Katehakis, H. Robbins, Sequential choice from several populations, Proc. Natl. Acad. Sci. USA 92 (1995) 8584–8585.

[17] S.R. Kulkarni, On bandit problems with side observations and learnability, Proc. 31st Allerton Conf. Commun. Contr. Comp., September 1993, pp. 83–92.

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

[19] T.L. Lai, H. Robbins, Asymptotically efficient allocation rules, Adv. in Appl. Math. 6 (1) (1985) 4–22.

[20] T.L. Lai, S. Yakowitz, Machine learning and nonparametric bandit theory, IEEE Trans. Automat. Control 40 (7) (1995) 1199–1209.

[21] H. Robbins, Some aspects of the sequential design of experiments, Bull. Amer. Math. Soc. 58 (1952) 527–535.

[22] J. Sarkar, One-armed bandit problems with covariates, Ann. Statist. 19 (4) (1991) 1978–2002.

[23] C.C. Wang, S.R. Kulkarni, H.V. Poor, Bandit problems with side observations, IEEE Trans. Automat. Control 50 (5) (2005).

[24] M. Woodroofe, A one-armed bandit problem with a concomitant variable, J. Amer. Statist. Assoc. 74 (368) (1979) 799–806.

[25] T. Zoubeidi, Optimal allocations in sequential tests involving two populations with covariates, Comm. Statist. Theory Methods 23 (4) (1994) 1215–1225.