Proceedings of the 42nd IEEE
Conference on Decision and Control
Maui, Hawaii USA, December 2003

Bandit Problems with Arbitrary Side Observations

Chih-Chun Wang, Sanjeev R. Kulkarni, H. Vincent Poor
Dept. of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA

***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, the essential properties of the side observations that allow achievability results with respect to the minimal inferior sampling time are extracted and formulated. The sufficient conditions for good side information obtained here contain various kinds of random processes as special cases, including i.i.d sequences, Markov chains, periodic sequences, etc. A necessary condition is also provided, giving more insight into the nature of bandit problems with side observations. A game-theoretic approach simplifies the analysis and justifies the viewpoint that the side observation serves as an index of different sub-bandit machines.**

I. INTRODUCTION

The classical two-armed bandit problem can be described in the context of finding the optimal choice between two slot machines, of which the reward distributions are unknown. At each time t, a player must balance the tradeoff between (1) learning which of the two machines gives better rewards, and (2) playing the better one. To accumulate enough experimental data to *learn* which arm is better, we inevitably are forced to sample both machines sufficiently often, which conflicts with the goal of sampling the best arm as many times as possible. Various optimal strategies for this problem have been found under different settings [1], [2], [3], [4], [5].

Most of the approaches of bandit problems are based on parametric models, where the underlying configurations/distributions of the arms are represented by a pair of parameters, $C_0 := (\theta_1, \theta_2)$. The sequences of rewards from the arms, denoted as $\{Y_\tau^i\}_{\tau\in\mathbf{N}}$, $i = 1, 2$, are assumed to be independent and identically distributed (i.i.d.) with marginals f_{θ_i} having unknown but fixed parameter θ_i. As the number of plays t tends to infinity, asymptotic analysis shows that appropriate decision rules are able to perform as well as those assuming complete knowledge of the unknown distributions. The convergence speed of a decision rule can be analyzed by estimating the growth rate of the "inferior sampling time". Generally, $\log t$ is the lowest order that can be uniformly achieved for every possible (θ_1, θ_2) with a fixed adaptive decision rule as discussed in [4]. A number of variations and extensions of this basic problem are investigated in [6], [7], [8].

This work is supported by the Office of Naval Research under Grant No. N00014-00-1-0141, Army Research Office under Contract No. DAAD19-00-1-0466, and the New Jersey Center for Pervasive Information Technologies.

Woodroofe [9] introduced the notion of side observations into classical bandit problems. In this situation, in addition to the history of previous decisions and outcomes, the player has access to side information before making a decision. Woodroofe proved that a myopic approach for i.i.d. side observations $\{X_\tau\}$ with simple relationships to $\{Y_\tau^i\}$ is asymptotically optimal. Sarkar [10] extended Woodroofe's results to exponential families. In [11], by using the side observations $\{X_\tau\}$ as the index of different sub-bandits with common configuration pair C_0, different levels of asymptotic efficiency improvements have been found for several types of relationships between $\{X_\tau\}$ and $\{Y_\tau^i\}$. Other approaches to bandit problems with side observations can be found in [12], [13].

The results in [9], [10], [11] suggest that the benefits of side observations in bandit problems is not from the *random* appearances, but rather is from the *evenly* distributed appearances of all values taken on by the i.i.d. $\{X_\tau\}$. In this paper, we extract the essential properties to be "evenly distributed" and investigate their effects on the attainable results. This paper is very much along the lines of [11] with a more general class of side observation sequences being considered, including i.i.d. sequences, Markov chains, and periodic sequences as special cases.

The extent to which side observations can help depends on the relationship of $\{X_\tau\}$ to the distributions of the rewards $\{Y_\tau^i\}$. Four different cases suggested in [11] are considered here, and we summarize these previous results for i.i.d. $\{X_\tau\}$ as follows. (See [11] for details.)

1) **Direct Information:** $\{X_\tau\}$ provides information directly about the underlying configuration $C_0 = (\theta_1, \theta_2)$, which allows a type of separation between learning and control. Under this setting, bounded expected inferior sampling time can be uniformly achieved.
2) **Best Arm Depends on X_t:** $\{X_\tau\}$ provides no information about C_0. For *every* configuration (θ_1, θ_2), arm 1 is preferred for some values of X_t, while arm 2 is preferred for other values of X_t. In this case, bounded expected inferior sampling time can again be uniformly achieved.
3) **Best Arm Does Not Depend on X_t:** $\{X_\tau\}$ provides no information about C_0, and for *every* configuration C_0, one of the arms is always preferred regardless of the value of X_t. An asymptotically tight $\log t$ lower bound still exists but the constant in front of $\log t$

can be improved by exploiting the notion of the most informative sub-bandits.

4) **Mixed Case:** This general case combines the previous two. For some configurations, one arm is always preferred (for all X_t), while for other possible configurations, the best arm depends on X_t. The best of the two individual cases can be achieved. I.e. bounded expected inferior sampling time can be uniformly achieved for those configurations in which the best arm depends on X_t as in Case 2. For those C_0 satisfying Case 3, asymptotically tight $\log t$ lower bounds can be uniformly achieved.

This paper is organized as follows. In Section II, we introduce the formulation of bandit problems with arbitrary side observations. Section III provides formal definitions of several "evenly distributed" properties used in the general theorems provided. In Sections IV through VII, we provide results for each of the four cases above, replacing the assumption of i.i.d. $\{X_\tau\}$ by "evenly distributed" properties. A more general framework is presented and thus other side observation processes (e.g. Markov chains and deterministic periodic sequences) in addition to i.i.d. $\{X_\tau\}$ can be addressed.

II. General Formulation

Consider a two-armed bandit problem defined as follows. Suppose we have two sequences of real-valued random variables $\{Y_\tau^i\}$, $i = 1, 2$, and a side observation sequence $\{X_\tau\}$ taking values in $\mathbf{X}$. The distribution of $\{X_\tau\}$ is described by the probabilities of finite cylinders, denoted by $G_{t_1, t_2, \cdots, t_k | C_0}$. The relationship between $\{Y_\tau^i\}$, $i = 1, 2$ and $\{X_\tau\}$ is as follows.

- Conditioned on the entire side observation $\{X_\tau\}$, $\{Y_\tau^i\}$, $i = 1, 2$, are independently distributed sequences.
- For any specific t, conditioned on X_t, the distribution of Y_t^i depends only on θ_i and nothing else.

The joint distribution of X_t and Y_t^i is $G_{t|C_0}(dx) H_{\theta_i}(dy|x)$. The entire families $\{G_C\}_{C \in \Theta^2}$ and $\{H_\theta\}_{\theta \in \Theta}$ are known to the decision maker and only the underlying configuration C_0 is unknown.

Necessary notation and several quantities of interest are defined in Table I. It is assumed throughout that all the necessary expectations exist and are finite.

Our goal is to find an adaptive allocation rule $\{\phi_\tau\}$ to maximize the growth rate of the expected reward $\mathsf{E}\{W_\phi(t)\}$, where

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

Instead of maximizing the growth rate of $\mathsf{E}\{W_\phi(t)\}$, it is equivalent to minimize the growth rate of the expected inferior sampling time, $\mathsf{E}\{T_{inf}(t)\}$, where

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

Therefore, we define a uniformly good rule as follows.

Definition 2.1 (Uniformly Good Rules): An allocation rule is uniformly good if for all $C_0 = (\theta_1, \theta_2)$, $\mathsf{E}_{C_0}\{T_{inf}(t)\} = o(t^\alpha)$, $\forall \alpha > 0$.
In what follows, we consider only uniformly good rules and regard other rules as uninteresting.

In the following development, we will make use of three different levels of required conditions, which are named as follows.

- *Ch1, Ch2,...*: "Characterization conditions" specify to which category the bandit problem belongs.
- *R1, R2, ...*: "Regularity conditions" are general enough to be satisfied for most cases, and may be removed by adding more complexity in the proof/analysis.
- *A1, A2, ...*: "Assumptions" are the conditions required in the proof/analysis, which are not stringent but may not be as general as the regularity conditions.

TABLE I
GLOSSARY

Not'n	Description
$1(C_0)$, $2(C_0)$	$1(C_0) = \theta_1$, $2(C_0) = \theta_2$.
$M_C(x)$	$M_C(x) := \arg\max_{i=1,2}\{\mu_{i(C)}(x)\}$.
$\mu_\theta(x)$	The conditional expectation of the reward, $\mu_\theta(x) := \mathsf{E}_{\theta_i=\theta}\{Y_t^i \mid X_t = x\}$.
$T_i(t)$	The total number of samples on arm i up to time t. $T_i(t) := \sum_{\tau=1}^{t} 1_{\{\phi_\tau = i\}}$.
$I(F,G)$	The Kullbeck-Leibler (K-L) information number, $I(F,G) := \mathsf{E}_F\left\{\log\left(\frac{dF}{dG}\right)\right\}$.
$I(\theta_1,\theta_2 \mid x)$	The conditional K-L information number, $I(\theta_1,\theta_2 \mid x) := I(H_{\theta_1}(\cdot \mid x), H_{\theta_2}(\cdot \mid x))$.

III. Essential Properties

Our goal is to extract the essential evenly-distributed properties of a side observation process that are helpful to the improvement of uniformly good rules.

To define "evenly distributed" among all $x \in \mathbf{X}$, we first assume $\mathbf{X}$ is finite. The relative frequency of x up to time t is denoted as $f_r(x,t) = \left(\sum_{\tau=1}^{t} 1\{X_\tau = x\}\right)/t$.

Definition 3.1 (Evenly Distributed in L^1): $\{X_\tau\}$ is evenly distributed in L^1 if

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

Definition 3.2 (Evenly Distributed in Probability Series): $\{X_\tau\}$ is evenly distributed "in probability series" if there exists a strictly positive mapping $\pi(x)$, such that the expected duration of the event $\{f_r(x,t) < \pi(x)\}$ is finite. That is,

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

Definition 3.3: (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 hitting time of x after T has a global upper bound, i.e. $\exists d > 0$ such that

$$\mathsf{E}\{H_T(x)|T\} \leq d < \infty, \forall T, x.$$

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

The following examples demonstrate that the above properties are quite general and include many interesting random processes.

- *Example 1:* If $\{X_\tau\}$ is an i.i.d. random process with strictly positive probability on each x, then by large deviations results, $\{X_\tau\}$ is evenly distributed in L^1, evenly distributed in probability series, and u.s.e. distributed in L^1.
- *Example 2:* If $\{X_\tau\}$ is a finite Markov chain with strictly positive entries in its transition matrix, then by similar reasonings, $\{X_\tau\}$ satisfies the above three conditions.
- *Example 3:* If we redefine $\mathbf{X}$ to be the set of values taken on during one period, any deterministic periodic sequence $\{X_\tau\}$ satisfies the above three conditions.

Note that both u.s.e. distributed in L^1 and evenly distributed in probability series imply evenly distributed in L^1.

IV. Direct Information

A. Formulation

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

- Dependence (*Ch1*): If $C \neq C'$, $\exists t_1, \cdots, t_k$, such that $G_{t_1,\cdots,t_k|C} \neq G_{t_1,\cdots,t_k|C'}$.

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

B. Scheme of Separating Learning and Control

Since we are able to obtain information about C_0 from $\{X_\tau\}$, one simple scheme is to sample only the seemingly better arm and to leave the learning task to $\{X_\tau\}$:

- Step 1: After time t, obtain an estimate $\hat{C}_t$ from the past side observations $X_1, \cdots, X_t$.
- Step 2: At time $t+1$, we set $\phi_{t+1} = M_{\hat{C}_t}(X_{t+1})$.

To find a bound of the performance, we use the following condition.

Condition 4.1 (A1): $\exists \epsilon > 0$ such that if $||\hat{C} - C_0|| < \epsilon$, $M_{\hat{C}}(x) = M_{C_0}(x)$, $\forall x \in \mathbf{X}$.

- *Example 4:* If (1) $\mathbf{X}$ is finite, and (2) $\forall x \in \mathbf{X}$, $\mu_\theta(x)$ is continuous with respect to θ, then *A1* is satisfied.
- *Example 5:* If $H_\theta(\cdot|x) \sim \mathcal{N}(\theta x, 1)$, then *A1* is satisfied.

Let E_{C_0} and P_{C_0} denote the expectation and the probability when the underlying configuration is C_0. We have

Theorem 4.1: Suppose both *Ch1* and *A1* hold. Then for any sequence of estimates $\{\hat{C}_\tau\}$, $\exists \epsilon > 0$ such that the inferior sampling time $T_{inf}(t)$ of the above algorithm satisfies

$$\lim_{t\to\infty} \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{1 + \sum_{\tau=1}^{t} \mathsf{P}_{C_0}(||\hat{C}_\tau - C_0|| > \epsilon)} \leq 1, \quad \text{for some } \epsilon > 0.$$

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

Corollary 4.1: If $\exists\{\hat{C}_\tau\}$ such that for all C_0 and any $\epsilon > 0$, $\lim_{t\to\infty} \sum_{\tau=1}^{t} \mathsf{P}_{C_0}(||\hat{C}_\tau - C_0|| > \epsilon)$ is finite, then $\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{inf}(t)\}$ is finite for all C_0.

- *Example 6:* If $\{X_\tau\}$ is an i.i.d. sequence with marginal distribution G_{C_0} and no two C's have the same G_C, then there exists $\{\phi_\tau\}$ such that $\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{inf}(t)\} < \infty$ for all C_0, and the constructed scheme is uniformly good.
- *Example 7:* If $\{X_\tau\}$ is a Markov chain with transition matrix A_{C_0}, and the mapping from C_0 to A_{C_0} is one-to-one, then there exists a uniformly good scheme $\{\phi_\tau\}$ such that $\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{inf}(t)\} < \infty$.
- *Example 8:* Consider the case in which $\{X_\tau\}$ is a deterministic sequence denoted by $\{x_\tau\}_{C_0}$. If the mapping from C_0 to $\{x_\tau\}_{C_0}$ is one-to-one, and Θ is finite, then there exists $\{\phi_\tau\}$ such that $\lim_{t\to\infty} \mathsf{E}_{C_0}\{T_{inf}(t)\} < \infty$ for all C_0.

From the above examples, we see that when the side observations reveal information about the underlying configuration C_0 in a fast enough fashion, by separating the learning and control by learning from observing $\{X_\tau\}$ and letting $\phi_{t+1} = M_{\hat{C}_t}(X_{t+1})$, we can achieve bounded expected inferior sampling time.

V. Best Arm Depends on X_t

A. Formulation

Henceforth, we consider the case in which observing X_t will not reveal any information about C_0, but reveals information only about the upcoming reward Y_t^i. In this section, we assume that the side observation X_t is *always* able to change the preference order as in Fig. 1. The characterization conditions are as follows.

- Independence (*Ch2*): $G_{t_1,t_2,\ldots,t_k|C_0} = G_{t_1,t_2,\ldots,t_k}$ does not depend on C_0.
- Best arm is a function of X_t (*Ch3*): $\forall C \in \Theta^2$, $\exists x_1, x_2 \in \mathbf{X}$, such that $M_C(x_1) = 1$ and $M_C(x_2) = 2$.

And the regularity conditions are

- *R1:* $\mathbf{X}$ is a finite space.
- *R2:* $\forall \theta_1, \theta_2, x$, $I(\theta_1, \theta_2|x)$ is strictly positive, and is finite.
- *R3:* $\Theta \subset \mathbf{R}$, and $\forall x$, $\mu_\theta(x)$ is continuous with respect to θ.

R1 embodies the idea of regarding X_t as the index of several different sub-bandit problems. *R2* ensures all these different bandit problems are non-trivial, i.e. with *non-identical* arms. *R3* facilitates our proof.

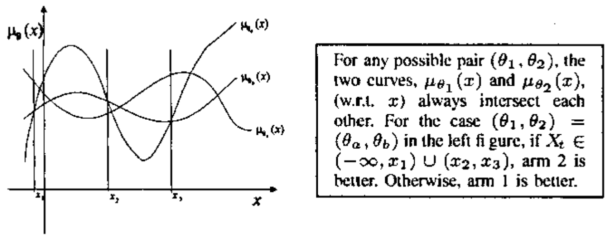

Fig. 1. The best arm at time t *always* depends on the side observation X_t.

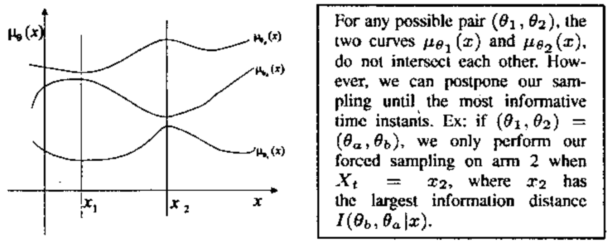

Fig. 2. The best arm at time t *never* depends on the side observation X_t.

B. Bounded Inferior Sampling Time

Theorem 5.1: Suppose *Ch2, Ch3, R1, R2,* and *R3* are satisfied. If the side observation random process $\{X_\tau\}$ is evenly distributed in probability series, then $\exists\{\phi_\tau\}$, such that

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

In the case that the side observation $\{X_\tau\}$ does not reveal any information about C_0, we consider the myopic rule, which samples only the seemingly better arm, i.e. $\phi_{t+1} = M_{\hat{C}_t}(X_{t+1})$. Since the even appearances of all x will direct the myopic rule to sample both arms often enough, the dilemma between learning and control is solved implicitly. As expected, we can surpass the $\log t$ lower bound and achieve bounded expected inferior sampling time, as long as the $\{X_\tau\}$ is evenly distributed in probability series.

VI. Best Arm Does Not Depend on X_t

A. Formulation

Following Section V, we assume that $\{X_\tau\}$ is independent of C_0. But now, $\forall C_0$, X_t *never* changes the preference order as illustrated in Fig. 2. Formal statements are as follows.

- Independence (*Ch2*): as in Section V.
- Best arm as a function of X_t (*Ch4*): $\forall C = (\theta_1, \theta_2)$, $\theta_1 \neq \theta_2$, we either have $M_C(x) = 1, \forall x$ or have $M_C(x) = 2, \forall x$.

With the two regularity conditions *R1* and *R2* as in Section V, we have improvements over the traditional bandit problems.

B. Lower Bound

Theorem 6.1: Under *Ch2, Ch4, R1,* and *R2*, for any uniformly good rule, suppose $M_{C_0}(x) = 2$, $\forall x$, (i.e. arm 2 is always better). Then $T_{inf}(t) = T_1(t)$ satisfies

$$\lim_{t\to\infty} \mathsf{P}_{C_0}\left(T_1(t) \geq \frac{\log t}{K_{C_0}}\right) = 1,$$

where $K_{C_0} = \inf_{\{\theta:\mu_\theta(x)>\mu_{\theta_2}(x),\forall x\}} \sup_{x\in\mathbf{X}}\{I(\theta_1, \theta|x)\}$. Furthermore, by Markov's inequality we have

$$\liminf_{t\to\infty} \frac{\mathsf{E}_{C_0}\{T_1(t)\}}{\log t} \geq \frac{1}{K_{C_0}}.$$

C. Asymptotic Tightness

To prove the asymptotic tightness of the lower bound in *Theorem 6.1*, we need additional conditions.

- Parameter space (*A2*): Θ is finite.
- Side observations (*A3*): $\{X_\tau\}$ is u.s.e. distributed in L^1.
- The existence of the value of the game (*A4*):

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

  - *Example 9:* Consider the case that $\Theta = \{1, 2, 3\}$, $\mathbf{X} = \{1, 2\}$, and $\{a_{\theta,x}\}$ is an arbitrary matrix with all entries in $[0, 0.1]$. If $H_\theta(\cdot|x) \sim \mathcal{N}(\theta + a_{\theta,x}, 1)$, the value of the game exists.

Theorem 6.2 (Asymptotic Tightness): With additional assumptions *A2, A3,* and *A4*, the $\log t$ lower bound of $T_{inf}(t)$ in *Theorem 6.1* is asymptotically tight.

The intuition behind this result is that when we are facing different *sub-bandit* machines with reward distribution pairs $\{(H_{\theta_1}(x), H_{\theta_2}(x))\}$, $\forall x \in \mathbf{X}$, we are able to uniformly minimize our forced sampling time (for learning purposes) by postponing it until facing the most *informative* sub-bandits. This idea is reflected in the expression of K_{C_0}. It can also be viewed as a two-player-zero-sum game such that nature wants to maximize the forced sampling time by selecting a good θ, while the decision maker has a strategy to wait until the most favorable chances (sub-bandits). If $\{X_\tau\}$ is even enough, i.e. the decision maker does not pay too much penalty for waiting, and the value of the game exists, we are able to reach the equilibrium with a carefully designed decision rule (the strategy of the decision maker).

VII. Mixed Case

A. Formulation

The main difference between Sections V and VI is that in one case, X_t *always* changes the preference order, and in the other case, X_t *never* changes the order. A much more general case is a mixture of these previous two cases, which will yield the main result of this paper.

- Independence (*Ch2*) as in Sections V and VI.
- Best arm as a function of X_t (*Ch5*): As in Fig. 3, for some C_0, $M_{C_0}(x)$ is independent of x, while for other C_0, $M_{C_0}(x)$ varies with respect to x.

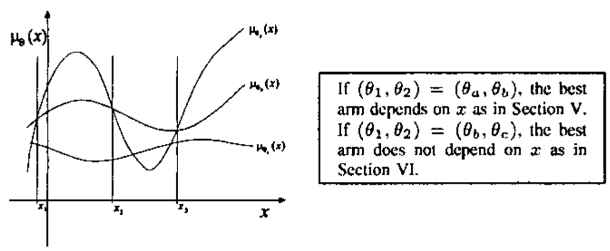

Fig. 3. Mixed case: The best arm at time t *may* or *may not* depends on the side observation X_t.

However, without knowledge of the authentic underlying configuration C_0, we do not know whether $M_{C_0}(x)$ remains the same for all x, or it changes with respect to x. In view of the results of Sections V and VI, we would like to find a scheme that has bounded inferior sampling time when being applied to an unknown configuration where X_t changes the preference order, and achieves the $\log t$ lower bound when being applied to configurations with a constant $M_{C_0}(x)$ for all x. With *R1* and *R2* as in Sections V and VI, we can obtain good results in this context.

B. Lower Bound

Suppose our side observation $\{X_\tau\}$ is evenly distributed in L^1. We have the following $\log t$ lower bound.

Theorem 7.1: Under *Ch2*, *Ch5*, *R1*, and *R2*, for any uniformly good rule, if the side observation $\{X_\tau\}$ is evenly distributed in L^1 and $M_{C_0}(x)$ remains the same for all x, we have

$$\lim_{t\to\infty} \mathsf{P}_{C_0}\left(T_{inf}(t) \geq \frac{\log t}{K_{C_0}}\right) = 1.$$

Furthermore, by Markov's inequality we have

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

If $M_{C_0}(x) = 2$ for all x, then K_{C_0} is given by

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

C. Asymptotic Tightness

We need the following assumptions to construct the desired scheme.

- Parameter space (*A2*): Θ is finite as in Section VI.
- Side observations (*A3*): $\{X_\tau\}$ is u.s.e. distributed in L^1.
- The existence of the value of the game (*A5*):

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

Theorem 7.2 (Asymptotic Tightness): With *Ch2*, *Ch5*, *R1*, *R2*, *A2*, *A3*, and *A5*, there exists a scheme either having bounded inferior sampling time, or achieving the $\log t$ lower bound of *Theorem 7.1*, depending on whether x is able to change the preference order $M_{C_0}(x)$.

The above theorem shows that for any C_0, depending on whether the dilemma between learning and control can be solved, we are able to achieve $\mathsf{E}T_{inf}(t) < \infty$ or the $\log t$ lower bound with constants derived from the game theoretic point of view, provided the side observation $\{X_\tau\}$ is evenly distributed among all x.

VIII. Necessary Conditions

The sufficient conditions for good side observations in Sections V through VII are summarized in *Table II.* It is useful to provide a necessary condition as well.

Theorem 8.1 (Common Necessary Condition): Suppose the conclusions of *Theorems 5.1, 6.2, and 7.2* hold for all distribution families $\{\mathsf{P}_C\}$ satisfying the characterization and regularity conditions. Then the following condition must hold as well:

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

The intuition behind *Theorem 8.1* is that if $\exists x_0$ such that $\mathsf{P}(\exists\tau, X_\tau = x_0) = 0$ (i.e. $\mathsf{P}(\forall\tau, X_\tau \neq x_0) = 1$), the benefit of the characterization properties (helpful structure between X_t, Y_t^i) may degenerate to another case with new support $\mathbf{X}' = \mathbf{X}\backslash\{x_0\}$, which significantly affects the attainable results.

IX. Conclusions

It has been shown in [11] that observing additional side information can improve sequential decisions in bandit problems. To further explore the origin of the improvement, in this paper we have extracted evenly distributed properties of the side observations and proved their efficacy for bandit problems. If $\{X_\tau\}$ provides information about the configuration C_0, with a scheme separating the learning and control, by observing $\{X_\tau\}$ for learning and playing arm $M_{\hat{C}_t}(X_{t+1})$ for control, the order of growth rate of the inferior sampling time is the same as the order of $\sum_{\tau=1}^{t} \mathsf{P}(||\hat{C}_\tau - C_0|| > \epsilon)$, which leads to bounded $\mathsf{E}\{T_{inf}(t)\} < \infty$, for i.i.d. sequences, Markov chains and deterministic periodic sequence $\{X_\tau\}$, among others.

If $\{X_\tau\}$ does not provide information about the configuration C_0, three cases have been considered: (1) the best arm depends on X_t, as in Section V, (2) the best arm does not depend on X_t, as in Section VI, and (3) the mixed case as in Section VII. We have proved that several sufficient conditions for the regular/even appearance of all $x \in \mathbf{X}$ can accomplish either bounded $\mathsf{E}\{T_{inf}(t)\}$ or the asymptotic $\log t$ lower bound.

Consequently, a much more general class of side observation sequences, which includes Markov chains, and fixed arbitrary sequences, has the same impact on bandit problems as those of i.i.d. sequences. A common necessary condition

TABLE II

SUMMARY OF THE RELATIONSHIP BETWEEN X_t AND Y_t^i.

Characterization	Regularity Conditions	Essential Properties	Results
$G_{t_1,\cdots,t_k\|C_1} \neq G_{t_1,\cdots,t_k\|C_2}$		As $\hat{C}_t \to C_0$, $\forall x$, $M_{\hat{C}_t}(x) = M_{C_0}(x)$.	$\exists\{\phi_\tau\}$ such that $\lim_{t\to\infty} \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{1+\sum \mathsf{P}(\|\|\hat{C}_\tau - C_0\|\|>\epsilon)} \leq 1$, $\forall C_0 \in \mathbf{\Theta}^2$.
$G_{t_1,\cdots,t_k\|C} = G_{t_1,\cdots,t_k}$, $\forall C$, $\exists x_1, x_2$, $M_C(x_1) = 1$, $M_C(x_2) = 2$.	$\mathbf{X}$ is finite. $\forall\theta_1 \neq \theta_2, x$, $I(\theta_1, \theta_2\|x) > 0$.	$\{X_\tau\}$ is evenly distributed in probability series, i.e. $\mathsf{E}\left\{\sum 1_{\{f_\tau(x,t)<\pi(x)\}}\right\} < \infty$.	$\exists\{\phi_\tau\}$ such that $\mathsf{E}_{C_0}\{T_{inf}(t)\} < \infty$, $\forall C_0 \in \mathbf{\Theta}^2$.
$G_{t_1,\cdots,t_k\|C} = G_{t_1,\cdots,t_k}$, $\forall C$, $M_C(x)$ only depends on C, not on x.	$\mathbf{X}$ is finite. $\forall\theta_1 \neq \theta_2, x$, $I(\theta_1, \theta_2\|x) > 0$.		For any uniformly good $\{\phi_\tau\}$, we have $\lim \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}}$, $K_{C_0} \triangleq \inf_\theta \sup_x I(\theta_1, \theta\|x)$.
	In addition to the above two, we need (1) $\mathbf{\Theta}$ is finite. (2) Value of the game, i.e. $\inf\sup I(\theta_1, \theta\|x) = \sup\inf I(\theta_1, \theta\|x)$.	$\{X_\tau\}$ is u.s.e. distributed in L^1, i.e. $\forall T$, $H_T(x)$, $\mathsf{E}\{H_T(x)\|T\} \leq d < \infty$.	$\exists\{\phi_\tau\}$ such that $\lim \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{\log t} \leq \frac{1}{K_{C_0}}$, $K_{C_0} \triangleq \inf_\theta \sup_x I(\theta_1, \theta\|x)$.
$G_{t_1,\cdots,t_k\|C} = G_{t_1,\cdots,t_k}$, $\exists\mathbf{C}_a \subset \mathbf{\Theta}^2$, $\forall C \in \mathbf{C}_a$, $\exists x_1, x_2$, such that $M_C(x_1) = 1$, $M_C(x_2) = 2$. $\forall C \in \mathbf{C}'_a$, $M_C(x)$ only depends on C, not on x.	$\mathbf{X}$ is finite. $\forall\theta_1 \neq \theta_2, x$, $I(\theta_1, \theta_2\|x) > 0$.	$\{X_\tau\}$ is evenly distributed in L^1, i.e. $\liminf \mathsf{E}\{f_\tau(x_0, t)\} > 0$.	For any uniformly good $\{\phi_\tau\}$, if $C \in \mathbf{C}'_a$, we have $\lim \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}}$, $K_{C_0} \triangleq \inf_\theta \sup_x I(\theta_1, \theta\|x)$.
	In addition to the above two, we need (1) $\mathbf{\Theta}$ is finite. (2) Value of the game, i.e. $\inf\sup I(\theta_1, \theta\|x) = \sup\inf I(\theta_1, \theta\|x)$.	$\{X_\tau\}$ is u.s.e. distributed in L^1, i.e. $\forall T$, $H_T(x)$, $\mathsf{E}\{H_T(x)\|T\} \leq d < \infty$.	$\exists\{\phi_\tau\}$ such that (1) if $C \in \mathbf{C}_a$, $\mathsf{E}_{C_0}\{T_{inf}(t)\} < \infty$, and (2) if $C \in \mathbf{C}'_a$, we have $\lim \frac{\mathsf{E}_{C_0}\{T_{inf}(t)\}}{\log t} \leq \frac{1}{K_{C_0}}$, $K_{C_0} \triangleq \inf_\theta \sup_x I(\theta_1, \theta\|x)$.

is also provided. From this paper, it is clear that the benefit of side observations lies mainly in the interactive structure of X_t and Y_t^i, and the evenly distributed appearances of all x.

X. REFERENCES

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

[2] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 148–177, 1979.

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

[4] T. L. Lai and H. Robbins, "Asymptotically efficient allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

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

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

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

[8] S. R. Kulkarni and G. Lugosi, "Finite-time lower bounds for the two-armed bandit problem," *IEEE Trans. Automat. Contr.*, vol. 45, no. 4, pp. 711–714, Apr. 2000.

[9] M. Woodroofe, "A one-armed bandit problem with a concomitant variable," *J. Amer. Stat. Assoc.*, vol. 74, no. 368, pp. 799–806, Dec 1979.

[10] J. Sarkar, "One-armed bandit problems with covariates," *Ann. Statist.*, vol. 19, no. 4, pp. 1978–2002, 1991.

[11] C. C. Wang, S. R. Kulkarni, and H. V. Poor, "Bandit problems with side observations," *IEEE Trans. Automat. Contr.*, under review.

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

[13] T. Zoubeidi, "Optimal allocations in sequential tests involving two populations with covariates," *Commun. Statist.: Theory and Methods*, vol. 23, no. 4, pp. 1215–1225, 1994.