Proceedings of the 41st IEEE
Conference on Decision and Control
Las Vegas, Nevada USA, December 2002

Bandit Problems with Side Observations[1]

Chih-Chun Wang, Sanjeev R. Kulkarni, H. Vincent Poor[2]

Abstract

An extension of the traditional two-armed bandit problem is considered, in which the decision maker has access to some side information before deciding which arm to pull. At each time t, before making a selection, the decision maker is able to observe a random variable, X_t, that provides some information on the rewards to be obtained. The focus is on finding uniformly good rules (that minimize the growth rate of the regret) and on quantifying how much the additional information helps. Various settings are considered and asymptotically tight lower bounds on the achievable regret are provided.

Keywords: two-armed bandit, side information, regret, allocation rule, asymptotic, efficient, adaptive

1 Introduction

Since the publication of [12], bandit problems have attracted much attention in various areas of statistics, control, learning, and economics (e.g., see [5, 10, 11]). In the classical two-armed bandit problem, at each time a player selects one of two arms and receives a reward from a distribution associated with the arm selected. The essence of the bandit problem is that the reward distributions are unknown, and so there is a fundamental tradeoff between gathering information about the unknown reward distributions and choosing the arm we currently think is best. A rich set of problems arises in trying to find an optimal/reasonable balance between these conflicting objectives (also referred to as learning versus control, or exploration versus exploitation).

In the traditional setting, the underlying distribution of the arms is expressed by a pair of parameters, (θ_1, θ_2). The sequence of rewards from the arms, denoted as $\{Y_t^i\}_{t\in\mathbf{N},i=1,2}$, are assumed to be independent and identically distributed (i.i.d.). The goal is to maximize the sum of the rewards or minimize the "regret", the shortfall of a rule compared with assuming complete knowledge of the reward distributions. The basic problem and a number of variations and extensions can be found in [11] and [1, 2, 7, 9].

In this paper, we consider an extension of the classical two-armed bandit where we have access to side information before making our decision about which arm to pull. At time t, in addition to the history of previous decisions, outcomes, and observations, we have access to a side observation, X_t, to help us make our current decision. The extent to which this side observation can help depends on the relationship of X_t to the reward distributions, namely the configuration (θ_1, θ_2) and the sequence of rewards Y_t^i.

Previous work on bandit problems with side observations includes [6, 8, 13, 14, 15]. Woodroofe [14] considered a one-armed bandit in a Bayesian setting. A simple criterion for asymptotically optimal rules was constructed. Sarkar [13] extended the side information model of [14] to the exponential family. In [8], Kulkarni considered classes of reward distributions and their effects on performance using results from learning theory. In [6], with the help of the "link" function, an index-based rule is provided. In [15], side information was used as the index of different categories.

In contrast with this previous work, we explore various general settings and focus on finding asymptotically tight lower bounds. Our work is very much along the lines of [10] and subsequent work such as [1, 2]. The settings are described as follows.

1. **Direct Information:** In this case, X_t provides information directly about the underlying configuration, $C_0 = (\theta_1, \theta_2)$, which allows a type of separation between learning and control. This has a dramatic effect on the achievable regret. Specifically, estimating (θ_1, θ_2) by observing $\{X_t\}$, and using the estimate $(\hat{\theta}_1, \hat{\theta}_2)$ to make the decision, results in bounded expected regret.

In the case that $\{X_t\}$ is independent of C_0, we are not able to learn C_0 through $\{X_t\}$. However, the rewards of the two arms may vary depending on the value of the side observation. Thus, by observing X_t in advance, we can hope to do better than without any side observation. We can also view this situation as having many related two-armed bandit machines – one for each value of X_t, so that observing X_t tells us which bandit machine we are playing at time t. The connec-

[1]This work is supported in part by the National Science Foundation under Grants ECS-9873451 and ECS-9811095, the Army Research Office under Contract DAAD19-00-1-0466, and the New Jersey Center for Pervasive Information Technologies.

[2]The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544. Email: {chihw, kulkarni, poor}@princeton.edu.

0-7803-7516-5/02/$17.00 ©2002 IEEE

tion between different machines is that they have the same common configuration pair (θ_1, θ_2), so that the rewards observed from one machine, provide information on *all* of the others (different values of X_t). This is the key aspect that makes the setup distinct from simply having many independent bandit problems.

We consider the following three cases of further refinements within this setting.

2. **Best Arm Depends on X_t:** In this case, for *all* configurations (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. Surprisingly, we exhibit an algorithm that achieves *bounded* expected regret.

3. **Best Arm Does Not Depend on X_t:** In this case, for *all* configurations (θ_1, θ_2), one of the arms is always preferred regardless of the value of X_t. However, since the rewards obtained will depend on X_t, the intuition is that we can postpone our learning until it is the most advantageous to us. We show that, asymptotically, our performance will be governed by the most "informative" bandit (among different values of X_t). Hence, we pay no penalty in terms of the constant in the $\log t$ growth rate of regret for having to learn which is the most advantageous bandit.

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

Our paper is organized as follows. In Section 2, we introduce the general formulation. In Sections 3 through 6, we consider the above four cases in turn.

2 General Formulation

The following discussions are all based on the probability space $(\Omega, \mathcal{F}, P_{C_0})$[3], where P_{C_0} is an element of a family of probability measures, $\{P_C\}_{C \in \Theta^2}$. The whole family is known to the player, but the true value of the corresponding index, $C_0 = (\theta_1, \theta_2)$, has to be learned through experiments.

[3] $\mathcal{F} = \bigcup_{t \in \mathbf{N}} \mathcal{F}_t$, where $\mathcal{F}_t$ is defined in Table 1.

Consider the two-armed bandit problem defined as follows. Suppose we have two random processes, $\{Y_t^i\}$, $Y_t^i : \Omega \mapsto \mathbf{R}$, $\forall i = 1, 2$, and an i.i.d. side observation sequence $\{X_t\}$, $X_t : \Omega \mapsto \mathbf{X}$. The relationship between $\{Y_t^i\}, \forall i = 1, 2$, and $\{X_t\}$ is as follows.

- Conditioned on $\{X_t\}$, $\{Y_t^i\}$ 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.

Necessary notations and several quantities of interest are defined in Table 1. We also assume that all the needed expectations exist and are finite. The goal is to find an adaptive allocation rule, $\{\phi_t\}$, to maximize the expected reward, $E\{W_\phi(t)\}$. Instead of maximizing the expected reward, it is equivalent to minimize the expected regret, $E\{R_\phi(t)\}$. With the assumption that all expectations are finite, the regret function is strongly related to the inferior sampling time, $T_{inf}(t)$.

We define a uniformly good rule as follows.

Definition 2.1 *An allocation rule is uniformly good if for all* $C_0 = (\theta_1, \theta_2)$, $E_{C_0}\{T_{inf}(t)\} = o(t^\alpha)$, $\forall \alpha > 0$.

In what follows, we only consider uniformly good rules and regard other rules as uninteresting.

3 Direct Information

3.1 Formulation

In this setting, the side observation, X_t, directly reveals information about $C_0 = (\theta_1, \theta_2)$ in the following way. Let G_C denote the marginal distribution of X_t under configuration C.

Dependence: $G_C = G_{C'}$ iff $C = C'$.

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

3.2 Scheme of bounded regret

Condition 1 *For any fixed* C_0, *and any sequence,* $\{\hat{C}_\tau\}$, $\hat{C}_\tau \to C_0$, $\exists t_0$ *such that* $(\mu_{1(\hat{C}_\tau)}(x) - \mu_{2(\hat{C}_\tau)}(x))(\mu_{1(C_0)}(x) - \mu_{2(C_0)}(x)) \geq 0, \forall x \in \mathbf{X}, \forall \tau > t_0$.

- *Condition 1* says that whenever our estimation $\{\hat{C}_\tau\}$ is close enough to our authentic pair, C_0, $\forall x$, the preference order of $(\mu_{1(\hat{C}_\tau)}(x), \mu_{2(\hat{C}_\tau)}(x))$ is the same as $(\mu_{1(C_0)}(x), \mu_{2(C_0)}(x))$.

Table 1: Glossary

Not'n	Description
Θ, Θ^2	Θ is the set of all possible θ; Θ^2 is the space of parameter pairs.
$\mathbf{X} \subset \mathbf{R}$	The support of X_t.
C_0	The authentic configuration parameter pair (θ_1, θ_2), which is the unknown index of the given probability distribution family, $\{P_C\}_{C \in \Theta^2}$.
$1(C_0)$, $2(C_0)$	$1(C_0) = \theta_1$, $2(C_0) = \theta_2$
$M_C(x)$	$\text{argmax}\{\mu_{1(C)}(x), \mu_{2(C)}(x)\}$
G_C	The marginal distribution of X_t, $\forall t$, under configuration C.
$H_{\theta_i}(\cdot\|x)$	The conditional marginal distribution of arm i, Y_t^i.
$\mathcal{F}_t$	The filtration until current time t, i.e. $\mathcal{F}_t = \sigma(X_t, X_\tau, 1_{\{\phi_\tau=1\}}Y_\tau^1 + 1_{\{\phi_\tau=2\}}Y_\tau^2, 1 \le \tau < t)$
ϕ_t	The decision rule, $\phi_t : \Omega \mapsto \{1, 2\}$, which is measurable with respect to $\mathcal{F}_t$.
$\mu_\theta(x)$	The conditional expectation of the reward, $\mu_\theta(x) = E_\theta(Y\|x)$.
$W_\phi(t)$	The reward function, $W_\phi(t) = \sum_{\tau=1}^t \left(1_{\{\phi_\tau=1\}}Y_\tau^1 + 1_{\{\phi_\tau=2\}}Y_\tau^2\right)$.
$R_\phi(t)$	The regret function, $R_\phi(t) = \sum_{\tau=1}^t 1_{\{\phi_\tau = m_{C_0}(X_\tau)\}} * \|\mu_{\theta_1}(X_\tau) - \mu_{\theta_2}(X_\tau)\|$
$T_i(t)$	The total number of samples on arm i up to time t. $T_i(t) = \sum_{\tau=1}^t 1_{\{\phi_\tau=i\}}$.
$T_{inf}(t)$	The total number of samples on the inferior arm up to time t. $T_{inf}(t) = \sum_{\tau=1}^t 1_{\{\phi_\tau = m_{C_0}(X_\tau)\}}$.
Y_a^{ih}	The value of arm Y^i at the a-th pull of the instants $X_t = x^h$.
Y_a^i	The value of arm Y^i at the a-th pull of all time instants.
$T_i^h(t)$	The total times arm 1 has been pulled up to time t, when $X_t = x^h$.
$I(F, G)$	The Kullbeck-Leibler (K-L) information number, $I(F,G) = E_F\left\{\log\left(\frac{dF}{dG}\right)\right\}$
$I(\theta_1, \theta_2\|x)$	The conditional K-L information number, $I(\theta_1, \theta_2\|x) = I(H_{\theta_1}(\cdot\|x), H_{\theta_2}(\cdot\|x))$

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

Theorem 3.1 *If* Condition 1 *is satisfied, then* $\exists\{\phi_t\}$ *such that* $\lim_{t\to\infty} E\{T_{inf}(t)\} < \infty$, *and* $\lim_{t\to\infty} T_{inf}(t) < \infty$ *a.s.*

We prove Theorem 3.1 by providing a bounded-regret scheme as follows.

Step 1: After time t, construct

$$\mathbf{C}_t = \left\{C \in \Theta^2 : \rho(G_C, L_t) \le \inf_{C \in \Theta^2} \rho(G_C, L_t) + \frac{1}{t}\right\}, \quad (1)$$

where L_t is the empirical measure of the side observations, $\{X_1, \cdots, X_t\}$, and ρ is the Prohorov metric[4] for probability measures on $\mathbf{R}$.

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

By the large deviation theorem and *Condition 1*, the estimate $\hat{C}_t$ approaches C_0 in an exponentially fast way, which implies the expected duration of $\rho(\hat{C}_t, C_0) \ge \epsilon$ is finite, and thus the expectation of $T_{inf}(t)$ is bounded, too.

[4]Discussion of the Prohorov metric can be found in [4].

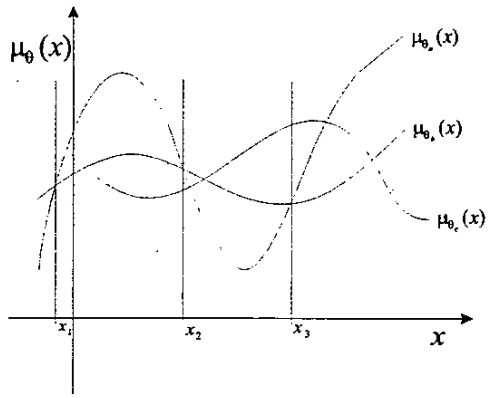

Figure 1: The best arm at time t *always* depends on the side observation X_t. That is, for any possible pair, (θ_1, θ_2), the two curves, $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$, (w.r.t. x) always intersect each other. For example, in the case $(\theta_1, \theta_2) = (\theta_a, \theta_b)$, if $X_t \in (-\infty, x_1) \cup (x_2, x_3)$, arm 2 is better. Otherwise, arm 1 is better.

4 Best Arm Depends on X_t

4.1 Formulation

Henceforth, we consider the case in which observing X_t will not reveal any information about C_0, but only reveals information about the upcoming reward, Y_t^i. In this section, we assume that the side observation, X_t, is *always* able to change the preference order, as in Figure 1. Here are the formal statements (with titles), and the needed regularity conditions (enumerated).

Independence: $G_{C_0} = G$ does not depend on C_0.

Best arm as a function of X_t: $\forall C \in \Theta^2$, we have both $P(\mu_{1(C)}(X_t) > \mu_{2(C)}(X_t)) > 0$ and $P(\mu_{1(C)}(X_t) < \mu_{2(C)}(X_t)) > 0$.

1. $\mathbf{X}$ is a finite space, denoted by $\mathbf{X} = \{x^1, \ldots, x^k\}$
2. $\forall \theta_1, \theta_2, x$, $I(\theta_1, \theta_2|x)$ is strictly larger than 0, and is finite.

The first condition embodies the idea of treating X_t as the index of several different bandit problems, which also simplifies our proof. The second condition is to make sure all of these different bandit problems are non-trivial, with *non-identical* arms. To facilitate our proof, we also assume the parameter space is a subspace of the reals, $\Theta \subset \mathbf{R}$, and $\forall x$, $\mu_\theta(x)$ is continuous w.r.t. θ.

4.2 Scheme of bounded regret

Theorem 4.1 *If the above conditions are satisfied, then there exists an allocation rule, $\{\phi_t\}$, such that*

$$\lim_{t\to\infty} E\{T_{inf}(t)\} < \infty. \quad (2)$$

Such a rule is obviously uniformly good.

We construct a bounded-regret allocation rule as follows.

Step 1: Since $\mathbf{X}$ is finite, $\mathbf{X} = \{x^1, x^2, \cdots, x^k\}$, we can define $n^i(t) = \max_{h\in\{1,\ldots,k\}}\{T_i^h(t)\}$, $i = 1, 2$. $h^{i*}(t)$ are the indices of the corresponding $n^i(t)$.

Step 2: $\phi_1 = 1$, $\phi_2 = 2$, $\phi_3 = 1$, $\phi_4 = 2$, $\phi_5 = 1$, $\phi_6 = 2$.

Step 3: After time t, construct the following set.

$$\mathbf{C}_t = \Big\{C = (\theta_1, \theta_2) \in \Theta^2 : \sigma(C, t) \le \inf\{\sigma(C, t) : C \in \Theta^2\} + \frac{1}{t}\Big\}, \quad (3)$$

with

$$\sigma(C,t) \triangleq \rho(H_{1(C)}(\cdot|x^{h^{1*}(t)}), L_1^{h^{1*}(t)}(t)) + \rho(H_{2(C)}(\cdot|x^{h^{2*}(t)}), L_2^{h^{2*}(t)}(t)) \quad (4)$$

where $L_i^h(t)$ is the empirical measure of sampling in group $\{Y_a^{ih}\}$. As before, $\rho(P, Q)$ is the Prohorov metric on $\mathbf{R}$. Then arbitrarily choose $\hat{C}_t \in \mathbf{C}_t$.

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

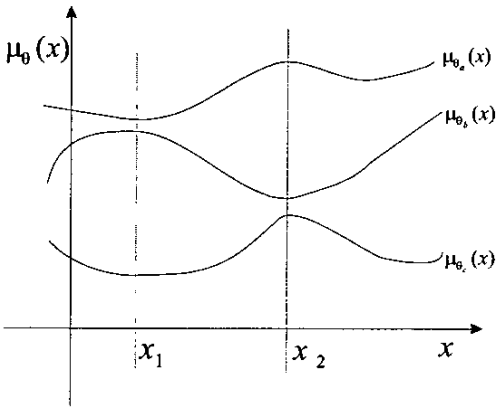

Figure 2: The best arm at time t *never* depends on the side observation X_t. That is, for any possible pair, (θ_1, θ_2), the two curves, $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$, do not intersect each other. However, in this case, we can postpone our sampling until the most informative time instants. For example, if $(\theta_1, \theta_2) = (\theta_a, \theta_b)$, we only perform our forced sampling on arm 2 when $X_t = x_2$, where x_2 has the largest information distance $I(\theta_b, \theta_a|x)$.

In this scheme, we have introduced an initial forced sampling mechanism which guarantees that we pull both arms a number of times with order greater than $O(t^{\frac{1}{2}})$. Since the error between the estimates $\hat{C}_t$ and C_0 goes to 0 exponentially fast w.r.t. the sample size, $O(t^{\frac{1}{2}})$, the expected duration of $\rho(\hat{C}_t, C_0) > \epsilon$ is bounded. When the forced sampling mechanism is not activated, we need only sample the seemingly better arm (depending on X_t), and the regular appearances of all $X_t = x$ ensure that we sample both arms often enough (with the order $O(t)$). As a result, the forced sampling mechanism will terminate quickly (bounded expectation). Since incorrect sampling results either from incorrect estimation, or from forced sampling, the scheme provided has $E\{T_{inf}(t)\} < \infty$.

5 Best Arm Does Not Depend on X_t

5.1 Formulation

Following Section 4, we assume $\{X_t\}$ is independent of C_0. But now, $\forall C_0$, $X_t = x$ will *not* change the preference order of $(\mu_{\theta_1}(x), \mu_{\theta_2}(x))$, as in the following statements and in Figure 2.

Independence: $G_{C_0} = G$ does not depend on C_0.

Best arm as a function of X_t: $\forall C = (\theta_1, \theta_2)$, $\theta_1 \neq \theta_2$, we have either $P(\mu_{1(C)}(X_t) < \mu_{2(C)}(X_t)) = 1$ or $P(\mu_{1(C)}(X_t) > \mu_{2(C)}(X_t)) = 1$.

Within the same two regularity conditions as in Section 4.1, we have improvements over the traditional bandit problems.

5.2 Lower bound

Theorem 5.1 *Under the above assumptions, for any uniformly good rule, suppose* $\mu_{\theta_1}(x) < \mu_{\theta_2}(x), \forall x$, *(i.e., arm 1 is always the inferior arm), then* $T_{inf}(t) = T_1(t)$ *satisfies*

$$\lim_{t\to\infty} P_{C_0}\left(T_1(t) \geq \frac{\log t}{K_{C_0}}\right) = 1, \tag{5}$$

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

$$\liminf_{t\to\infty} \frac{E_{C_0}\{T_1(t)\}}{\log t} \geq \frac{1}{K_{C_0}}. \tag{6}$$

This situation is like having several related bandit machines, whose reward distributions are all determined by the common configuration pair, (θ_1, θ_2). The information obtained from one machine is also applicable to the other machines. If arm 2 is always better than arm 1, we wish to sample arm 2 most of the time (the control part), and force-sample arm 1 once in a while (the learning part). With the help of the side information X_t, we can pull the seemingly better arm most of the time, and postpone our forced sampling (learning) to the most informative $X_t = x$. As a result, the constant in the $\log t$ lower bound in [10] has been further reduced to K_{C_0}.

With the additional assumption that Θ is finite, we have the following tightness theorem. Inspired by [1], the rather involved proof is omitted in this summary.

Theorem 5.2 (Asymptotic Tightness) *There exists a scheme achieving the lower bound (6). Accordingly, (6) is an asymptotically tight lower bound.*

6 Mixed Case

6.1 Formulation

In Sections 4 and 5, we dealt with the cases in which the distribution of X_t is independent of C_0. The main difference between Sections 4 and 5 is that in one case, X_t *always* changes the preference, in the other, X_t *never* changes the order. A much more general case is a mixture of these previous two cases. In this section, we consider this mixed case, which leads to the main result of this paper. The formal statements are as follows.

Independence: $G_{C_0} = G$ does not depend on C_0.

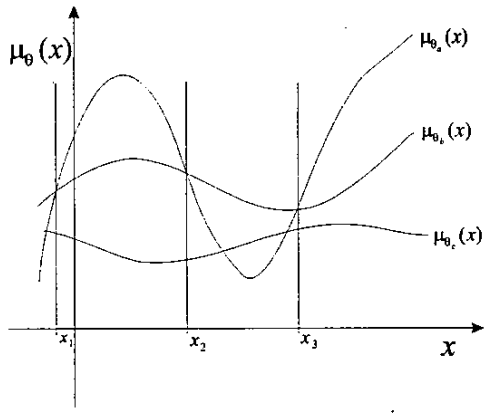

Figure 3: If $(\theta_1, \theta_2) = (\theta_a, \theta_b)$, the best arm depends on x, i.e. $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ intersect with each other as in Section 4. If $(\theta_1, \theta_2) = (\theta_b, \theta_c)$, the best arm does not depend on x, i.e. $\mu_{\theta_1}(x)$ and $\mu_{\theta_2}(x)$ do not intersect with each other as in Section 5.

Best arm as a function of X_t: As in Figure 3, for some $C_0 \in \Theta^2 \subset \mathbf{R}^2$, *Condition 2* is satisfied, for some other C_0, *Condition 2* is not satisfied, where

Condition 2 (θ_1, θ_2) *satisfies both the following inequalities:*

$$P(\mu_{\theta_1}(X_t) \geq \mu_{\theta_2}(X_t)) > 0$$
$$\text{and } P(\mu_{\theta_2}(X_t) \geq \mu_{\theta_1}(X_t)) > 0.$$

Whenever C_0 satisfies *Condition 2*, the side observation, X_t, changes the order. It can also be the case that C_0 does not satisfy *Condition 2*, such that the preference order remains the same no matter what X_t is. However, without knowledge of the authentic underlying configuration C_0, we do not know which case is active. In view of the results of Sections 4 and 5, we would like to find a scheme that has bounded regret when being applied to an unknown bandit with parameters satisfying *Condition 2*, and achieves the $\log t$ lower bound when being applied to bandits not satisfying *Condition 2*. Within the same two regularity conditions as in Section 4.1, we can obtain good results in this context.

6.2 Lower bound

Theorem 6.1 *For any uniformly good rule, if* $C_0 = (\theta_1, \theta_2)$ *does not satisfy* Condition 2, *we have*

$$\lim_{t\to\infty} P_{C_0}\left(T_{inf}(t) \geq \frac{\log t}{K_{C_0}}\right) = 1. \tag{7}$$

Furthermore, by Markov's inequality we also have

$$\liminf_{t\to\infty} \frac{E_{C_0}\{T_{inf}(t)\}}{\log t} \geq \frac{1}{K_{C_0}}. \tag{8}$$

If it is the case that $\mu_{\theta_1}(x) < \mu_{\theta_2}(x), \forall x$, *then* K_{C_0} *is given by*

$$K_{C_0} = \inf_{\{\theta: P(\mu_\theta(X_t) > \mu_{\theta_2}(X_t)) > 0\}} \sup_x \{I(\theta_1, \theta | x)\}. \quad (9)$$

The only difference between the lower bounds (6) and (8) is that, in (8), K_{C_0} has been changed from taking the infimum over $\{\mu_\theta(x) > \mu_{\theta_2}(x)\}$ to a larger set, $\{\theta : P(\mu_\theta(X_t) > \mu_{\theta_2}(X_t)) > 0\}$. The reason for this is that we need to consider the case in which $C' = (\theta, \theta_2)$, satisfying $P(\mu_\theta(X_t) > \mu_{\theta_2}(X_t)) > 0$, is mis-detected as C_0. To avoid having this type of mis-detection, which introduces a linear order of incorrect sampling, we need to perform the forced sampling more often. The constant in front of $\log t$ increases as a result of the relaxed conditions on the parameter space. The following theorem shows tightness of the result above.

Theorem 6.2 (Asymptotic Tightness) *If* Θ *is finite, then there exists a scheme that has a bounded regret if the underlying configuration pair* C_0 *satisfies* Condition 2, *and achieves the* $\log t$ *lower bound if* C_0 *does not satisfy* Condition 2.

7 Conclusion

We have shown that observing additional side information can definitely improve sequential decisions in bandit problems. If the side observation itself directly provides information about the underlying configuration, then it resolves the dilemma of forced sampling and optimal control. The expected regret will be finite as t tends to infinity, as described in Section 3. If the side observation does not provide information of the underlying configuration (θ_1, θ_2), but *always* affects the preference order, then the myopic approach of sampling the seemingly-best arm will automatically sample both arms often enough. The expected regret is bounded, as described in Section 4. If the side observation *does not* affect the preference order at all, the dilemma still exists. However, by postponing our forced sampling to the most informative time instants, we can reduce the constant in the $\log t$ lower bound, as described in Section 5. In Section 6, we combine the settings of Sections 4 and 5, and obtain a general result. When the underlying configuration C_0 is good (such that X_t will change the preference order), we obtain bounded expected regret as in Section 4. Even if C_0 is not good (in that X_t does not change the preference order), the new $\log t$ lower bound can be achieved as in Section 5.

References

[1] 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.

[2] 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. AC-32, no. 11, pp. 968-976, Nov. 1987.

[3] D. A. Berry and B. Fristedt, *Bandit Problems, Sequential Allocation of Experiments*, London: Chapman and Hall, 1985.

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

[5] H. Chernoff, *SIAM*, Philadelphia: Society for Industrial and Applied Mathematics, 1972.

[6] M. K. Clayton, "Covariate Models For Bernoulli Bandits," *Sequential Analysis*, vol. 8, no. 4, pp. 405-426, 1989.

[7] M. N. Katehakis and H. Robbins, "Sequential Choice from Several Populations," in *Proc. Nat. Acad. Sci. USA*, vol. 92, pp. 8584-8585, Sept. 1995.

[8] S. R. Kulkarni, "On Bandit Problems With Side Observations and Learnability," in *Proc. 31st Allerton Conf. Commun. Contr. Comp.*, pp. 83-92, Sept. 1993.

[9] 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.

[10] T. L. Lai and H. Robbins, "Asymptotically Optimal Allocation of Treatments in Sequential Experiments," *in Design of Experiments : Ranking and Selection : Essays in Honor of Robert E. Bechhofer*, Thomas J. Santner, Ajit C. Tamhane, New York: Dekker, 1984.

[11] T. L. Lai and H. Robbins, "Asymptotically Efficient Allocation Rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4-22, 1985.

[12] H. Robbins, "Some Aspects of the Sequential Design of Experiments," *Bull. Am. Math. Soc.*, vol. 58, pp. 527-535, 1952.

[13] J. Sarkar, "One-Armed Bandit Problems with Covariates," *Ann. Statist.*, vol. 19, no. 4, pp. 1978-2002, 1991.

[14] M. Woodroofe, "A One-Armed Bandit Problem With a Concomitant Variable," *J. Amer. Stat. Assoc.*, vol. 74, no. 368, Theory and Methods Section, pp. 799-806, Dec. 1979.

[15] 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.