# Multi armed bandit problem: some insights

July 14, 2011

## Introduction

Multi Armed Bandit problems have been widely studied in the context of sequential analysis. The application areas include clinical trials, adaptive filtering, online advertising etc. The study is also characterized as a policy selection which maximizes a gambler's reward when there are multiple slot machines that are generating them. It is under this framework, that we describe the model and develop subsequent results. Lai and Robbins [8] studied this problem under statistical settings and developed asymptotically efficient adaptive allocation rules.

In this current work, we provide an alternate (and simpler) technique to derive the lower bound result in Lai and Robbins [8]. We use this bounding technique to demonstrate the complexity of bandit problems where an extra initial information about the parameter(s) in the arms is available. Finally, we add insights on the fundamental complexity of these sampling problems by establishing necessary and sufficient conditions for optimal policies. Our proof techniques rely on martingale techniques and information theory arguments and hence substantially differ from the traditional approach of Lai and Robbins [8] which hinges upon the change of measure argument.

## Model Formulation and Related Literature

In the typical two armed bandit setting, there are two statistical populations characterized by univariate density functions $f_{\theta i}(x), i = 1, 2$ where $\theta_i$ being the unknown parameter assumed to belong to a family set $\Theta$. The set of policies that are considered are adaptive i.e. they depend only on the past actions and observations. Such a policy $\pi$ at each time instant samples from one of the populations. The reward obtained from sampling from arm $i$ is $Y_t = Y_t^{(i)}$ if sampled from $i^{th}$ arm at time $t$. The total reward obtained by a policy $\pi$ until stage $n$ is

$$Reward_n(\pi, \theta) = E_\theta^{\pi,n} \sum_{i=1}^{n} Y_t^{(i)}$$

where $E_\theta^{\pi,n}$ denotes expected value w.r.t to joint distribution of $P_\theta^{\pi,n}$ of observations collected until stage $n$. We also define the *oracle* policy $\pi^*$ as the policy that ex-ante knows the parameter

governing the arms. Such a policy $\pi^*$ always selects the arm which has the highest mean reward i.e. arm for which $\mu(\theta)$ is the highest. The efficiency of a policy $\pi$ is determined by comparing it with respect to the oracle policy. The *regret* of a policy is defined as

$$R_n(\pi, \theta) := Reward_n(\pi^*, \theta) - Reward_n(\pi, \theta)$$

$R_n(\pi, \theta)$ can also be simplified as,

$$R_n(\pi, \theta) := |\mu(\theta_1) - \mu(\theta_2)| E_\theta^{\pi,n}[T_{inf}(n)]$$

The objective is to construct policies that would minimize the regret $R(n)$ as $n \to \infty$. Lai and Robbins[] establish the complexity of any policy satisfying a set of assumptions (given below), has to make at least $o(log(n))$ sub optimal pulls asymptotically. We develop here a similar bound in the worst case scenario. The set of all policies considered in this study, is much bigger. The proof technique used here is predicated by information theoretic arguments, much simpler, in establishing the fundamental complexity of multi armed bandit problems.

***Assumptions*** (in the Lai and Robbins Formulation):
In the following, $I(\theta, \lambda)$ is the Kullback-Liebler divergence between probability distribution functions $f(x;\theta)$ and $f(x;\lambda)$ and $\mu(r)$ is the mean of the density function $f(x;r)$, $\Theta$ be the set from which the parameters governing the arms are chosen from.

- $\forall \epsilon > 0$ and $\forall \theta, \lambda$ such that $\mu(\lambda) > \mu(\theta), \exists \Delta = \Delta(\epsilon, \theta, \lambda) > 0$ for which $|I(\theta, \lambda) - I(\theta, \lambda^{'}) | < \epsilon$ whenever $\mu(\lambda) \leq \mu(\lambda') \leq \mu(\lambda) + \Delta$
- $\forall \lambda \in \Theta$ and $\forall \Delta > 0, \exists \lambda' \in \Theta$ such that $\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \Delta$
- Let $\theta = (\theta_1, \theta_2, ....\theta_k)$ and $\Theta_j^*$ defined as $\theta : \mu(\theta_j) > max_{i \neq j} \mu(\theta_i)$. The set of feasible rules satisfy for every $\theta \in \Theta_j^*$, as $n \to \infty$ $\Sigma_{i \neq j} E_\theta T_n(i) = o(n^a)$ for every $a > 0$ where $T_n(i)$ is the number of times an inferior arm $i$ is pulled until stage $n$

**Theorem 1** *For every such feasible rule satisfying the above set of assumptions,*
$\lim_{n\to\infty} P_\theta\{T_n(j) \geq (1-\epsilon)(\log(n))/I(\theta_j, \theta^*)\} = 1$ *and* $liminf_{n\to\infty} E_\theta T_n(j) \geq \frac{\log(n)}{I(\theta_j, \theta^*)}$ *for every* $\theta \in \Theta_j$ *and every* $\epsilon > 0$ *where,* $\Theta_j = \{\theta : \mu(\theta_j) < max_{i \neq j} \mu(\theta_i)\}$

*Proof* : Refer to Theorem 2, Lai and Robbins [8]

# Main Results

## Lower bounds for bandit problem

We present here an alternate characterization of the same problem. Our proof approach does not rely on the strict set of assumptions as needed in the Lai - Robbins [8] formulation of the bandit

problem. Instead, we develop bounds on the worst case performance of any adaptive policy. The proof here is for the case of a bandit with 2 arms.

**Lemma 2** *The lower bound on the sum of error probabilities in hypothesis testing in terms of the Kullback - Liebler number is given by,*

$$P^{\pi,t}_{H_{(0)}}(\psi_t = 1) + P^{\pi,t}_{H_{(1)}}(\psi_t = 0) \geq \frac{1}{2} exp\{-I(P^{\pi,t}_{H_{(0)}}, P^{\pi,t}_{H_{(1)}})\}$$

*Proof* : Refer to Tysabakov [13], 2009, Theorem 2.2

**Lemma 3** *Let $\Theta$ be a parameter set in $R^2$ such that if $(\theta_a, \theta_b) \in \Theta$ then $|\mu(\theta_a) - \mu(\theta_b)| \geq \epsilon$ for some $\epsilon > 0$, Then for any arbitrary policy $\pi$, configurations $\theta_{(0)}$, $\theta_{(1)} \in \Theta$ and all $n$, we have*

$$S_n(\pi, \Theta) = \sup_{\theta \in \Theta^2} E^{\pi,n}_{\theta}[T_{inf(n)}] \quad \geq \frac{1}{4} \sum_{t=1}^{n} exp\{-I((P^{\pi,t}_{\theta_{(0)}}, P^{\pi,t}_{\theta_{(1)}}))\}$$

*Proof* : The proof is based on mapping the above problem to a hypothesis testing problem and using lemma 1. Without loss of generality, let us assume arm 2 to be the best. i.e. $\mu(\theta_2) > \mu(\theta_1)$. Thus, the *oracle* policy, $\pi^*$ plays arm 2 at every decision time.

Now, for any 2 configurations $\theta_{(0)}, \theta_{(1)} \in \Theta$ and any arbitrary policy $\pi$ we have

$$\begin{aligned} S_n(\pi, \Theta) &= \sup_{\theta \in \Theta^2} E^{\pi,n}_{\theta}[T_{inf(n)}] \\ &= \sup_{\theta \in \Theta^2} \Sigma_{t=1}^{n} P^{\pi,t}_{\theta}(\pi_t \neq \pi^*_t) \\ &\geq \sum_{t=1}^{n} (P^{\pi,t}_{\theta_0}(\pi_t \neq \pi^*_t) + P^{\pi,t}_{\theta_1}(\pi_t \neq \pi^*_t))/2 \end{aligned} \tag{1}$$

Choose $\theta_3$ such that $\mu(\theta_2) > \mu(\theta_3) > \mu(\theta_1)$ and consider the following configurations $\theta_{(0)}$ : $(\theta_1, \theta_3)$ and $\theta_1 : (\theta_2, \theta_3)$. The existence of $\theta_3$ is clear from the assumption on the parameter family.

Fix the time horizon, $t = 1, 2.....n$ and consider the following hypothesis testing problem,
Hypothesis $H_{(0)}$: $\theta_{(0)}$ is the parameter set defining the arms, versus
Hypothesis $H_{(1)}$: $\theta_{(1)}$ is the parameter set defining the arms

$\pi^*_t$, the *oracle* policy plays arm 2 under $H_{(0)}$ and plays arm 1 under $H_{(1)}$. Let $p_t$ denote a particular policy that plays arm 1 always and define $\psi_t = 1(\pi_t = p_t)$

$$\{\psi_t = 1\} = \{\pi_t = p_t\} = \{\pi_t \neq \pi^*_t\} under H_{(0)}$$
$$\{\psi_t = 0\} = \{\pi_t \neq p_t\} = \{\pi_t \neq \pi^*_t\} under H_{(1)}$$

Hence, $P^{\pi,t}_{\theta_{(0)}}(\psi_t = 1) + P^{\pi,t}_{\theta_{(1)}}(\psi_t = 0) = P^{\pi,t}_{\theta_0}(\pi_t \neq \pi^*_t) + P^{\pi,t}_{\theta_1}(\pi_t \neq \pi^*_t))$

From Lemma 1,

$$P^{\pi,t}_{\theta_0}(\pi_t \neq \pi^*_t) + P^{\pi,t}_{\theta_1}(\pi_t \neq \pi^*_t)) \geq \frac{1}{2}\exp\{-I(P_{\theta_{(0)}}, P_{\theta_{(1)}})\}.$$

which implies,

$$S_n(\pi, \Theta) \geq \frac{1}{4}\sum_{t=1}^{n} exp\{-I((P^{\pi,t}_{\theta_{(0)}}, P^{\pi,t}_{\theta_{(1)}}))\}$$

**Theorem 4** *Under the assumptions on the parameter family, the performance of any adaptive policy i.e. the worst case regret on the inferior sampling rate is* $o(\log(n))$

$$S_n(\pi, \Theta) \geq C\log(n)$$

*where* $C$ *is a constant. The explicit value of the constant can be found in the proof below.*

*Proof* : For a particular policy $\pi$, let $J^\pi_i(t)$ be the subset of indices from 1,2,....t when the policy $\pi$ selects $i^{th}$ arm. With the same configurations described earlier as the parameters defining the arms, we apply lemma 2. Calculating $I(.,.)$, the KL distance explicitly, we have

$$\begin{aligned}
I((P^{\pi,t}_{\theta_{(0)}}, P^{\pi,t}_{\theta_{(1)}})) &= E_{\theta_{(0)}}\left(\log\left(\frac{f(X1,\theta_1)f(X2,\theta_1).....f(X_{J^\pi_1},\theta_1)f(Y1,\theta_3)f(Y2,\theta_3).....f(Y_{J^\pi_2},\theta_3)}{f(X1,\theta_2)f(X2,\theta_2).....f(X_{J^\pi_1},\theta_2)f(Y1,\theta_3)f(Y2,\theta_3).....f(Y_{J^\pi_2},\theta_3)}\right)\right) \\
&= E_{\theta_{(0)}}\left(\log\left(\frac{f(X1,\theta_1)f(X2,\theta_1).....f(X_{J^\pi_1},\theta_1)}{f(X1,\theta_2)f(X2,\theta_2).....f(X_{J^\pi_1},\theta_2)}\right)\right) \\
&= E_{\theta_{(0)}}\left(\Sigma^{J^\pi_1}_{t=1}\log(\frac{f(X1,\theta_1)}{f(X1,\theta_2)})\right) \\
&= I(\theta_1,\theta_2)E_{\theta_{(0)}}\left(J^\pi_1\right), \\
&= I(\theta_1,\theta_2)E_{\theta_{(0)}}[T_{inf}(n)],
\end{aligned}$$

where $I(\theta_1, \theta_2)$ is the KL distance between the density functions $f(x;\theta_1)$ and $f(x;\theta_2)$

From Lemma (2),

$$\begin{aligned}
S_n(\pi,\Theta) &\geq \sum_{t=1}^{n}\exp\{(-I(\theta_1,\theta_2)E_{\theta_{(0)}[T_{inf}(n)]})\} \\
&\geq \sum_{t=1}^{n}\exp\{(-I(\theta_1,\theta_2)\sup_{\theta\in\Theta^2}E^{\pi,n}_\theta[T_{inf(n)}])\} \\
&= \frac{1}{4}\sum_{t=1}^{n}\exp\{-I(\theta_1,\theta_2)S_t\} \\
&\geq \frac{n}{4}\exp\{(-I(\theta_1,\theta_2)S_n)\}
\end{aligned} \tag{2}$$

where for brevity we write $S_t := S_t(\pi, \Theta)$ and the last step follows since $S_t$ is a non decreasing sequence. The above result is true for any arbitrary adaptive policy and all $n$.
Taking logarithm on both sides of the last statement, we get

$$S_n(I(\theta_1, \theta_2) + \frac{\log(S_n)}{S_n}) \geq \log(n) - \log(4)$$

The last statement is true only if for any $\epsilon$, the numerical sequence $S_n$ satisfies for all $n$ large enough $S_n \geq (\frac{1}{I(\theta_1,\theta_2)+\epsilon} + o(1)) \log(n)$, since $\frac{\log(S_n)}{S_n} \leq \epsilon$ for every $\epsilon > 0$ for $n$ large enough.
If the horizon of the play is decided to be finite (i.e. fixed $n$) , then, consider the solution,

$$S_{n'} = argmin_{S_n \uparrow}(I(\theta_1, \theta_2)S_n + \log(S_n) \geq \log(n) - \log(4)$$

Fixing $C = \frac{S'_n}{\log(n)}$ completes the statement of the theorem.

### Sufficient conditions: General MAB problem

The discussion so far characterized an alternate way of deriving the Lai and Robbins [8] lower bound on optimal sampling. Lai and Robbins [8] also formulate a policy which attains the lower bound (i.e. asymptotically optimal). Stemming from that work, is a rich stream of literature on various optimal policies for bandit problems and its variants. To name a few, [**?**] characterized set of asymptotically optimal policies which at time depended only on the observed sample mean. The more recent work of [2] provides a simple policy for the traditional bandit problem with the reward having a bounded support. The following proofs in the paper was in some sense, motivated by the ideas in Auer et al. [2]. To wit, the proof techniques hinge upon martingale techniques rather than measure theoretic arguments as in [8]. Instead of a detailed survey, we refer the reader to the works of Woodroofe [14], Agarwal [**?**] etc. and the more recent additions of Goldenshluger and Zeevi [7, 6] Rusmevichientong and Tsitsiklis [**?**] develop optimal policies for variants of the traditional bandit problems.

The underlying intuition in all the aforementioned papers, is to develop a policy that captures the ideal interplay between the growth of information and sampling the best arm. In the theorem below, we mathematically establish the characteristics that are "hidden" or "exploited" in all these works. Thus explaining the intuitive and quantitative connect that models the bandit type problems.

Let the $i^{th}$ observation from arm $j$ be denoted by $X_j^i$. Let $X_1, X_2.....X_t$ denote the vector of observations till time $t$. Clearly, this vector depends on the policy and observations at each time step. Then the probability density function (p.d.f.) of this vector is given by

$$f(X_1, X_2.....X_{t+1}) = f(X_1^1)f(X_1^2)....f(X_1^{J_1^\pi(t+1)})f(X_2^1)f(X_2^2)....f(X_2^{J_2^\pi(t+1)}) \tag{3}$$

Equation (3) follows from the adaptive nature of the policy i.e. the decision step at time $t+1$ is adapted w.r.t. to the filtration $\mathcal{F}_t = \sigma(X_1, Y_1, X_2, Y_2...X_t, Y_t)$. For the sake of completion, we give a proof in the appendix.

We define empirical fisher information of the parameter in arm $i$ in the following way:

$$F_t^i = -\frac{\partial^2 log(f(X_t))}{\partial \theta_i^2} = -\sum_{j=1}^{t} \frac{\partial^2 log(f(X_i^j))}{\partial \theta_i^2} I\{Y_j = i\} \tag{4}$$

which follows from (1) and that the 2 arms are independent of each other.Reference to the above characterization of fisher information can be found in [12]

The expected fisher information in this problem has the following expression,

$$EF_t^i := -E[\frac{\partial^2 log(f(X_t))}{\partial \theta_i^2}|\theta_i] = E[J_i^\pi(t)]F(\theta_i) \tag{5}$$

where the last equality follows from (1) and Wald's equation type argument ($F(\theta_i = -E[\frac{\partial^2 log(f(X_i))}{\partial \theta_i^2}|\theta_i]$ i.e. Fisher information collected from arm $i$ from one pull of that arm )

Few observations and assumptions before we proceed.
For the exponential family random variables, the empirical and expected fisher information are the same. In particular, for normal random variable every time it is sampled, the fisher information increases at the rate 1. With that in mind, and the following regularity assumption, $-\frac{\partial^2 log(f(X_i))}{\partial \theta_i^2}] \leq M(\theta_i)$ for some $M$ for each arm $i$. This kind of condition is satisfied by most of the common distributions e.g. the exponential family.
In all further discussions, myopic policy is denoted as one which at each time step chooses the arm with highest sample mean thus far.

$$Y_t = 1I(\hat{\mu}_t(\theta_1) \geq \hat{\mu}(\theta_2)) + 2I(\hat{\mu}_t(\theta_2) \geq \hat{\mu}(\theta_1)) \tag{6}$$

Now we develop a sufficiency result for the bandit problem using the empirical fisher information characterization and establish the connect between the growth of empirical fisher information and the expected fisher information. The key ideas underlying the proof and the statement of the theorem can be articulated as:

- **Fair exploration** Any policy should not be dictatorial in the choice of the arms as such a policy can always be vanquished by swapping the arm parameters. Hence, any policy should "optimally" explore all the arms and learn "on the fly" the best arm. This condition

ensures the policy samples atleast as much indicated by lower bound which is the required information needed for any policy to distinguish the parameter family.

- **Proximity to myopic policy** Expected number of times an optimal policy differs from the corresponding myopic policy is not significantly high. As the number of samples from the arms increases, the estimate of the parameters is proximal to the true parameters. Therefore, the myopic policy would be able to perform significantly better and closer to the optimal policy asymptotically. Hence, any proposed policy should try to mimic the actions of the corresponding myopic policy except for "small" exploration phases. We prove the expected difference of less than $O(log(T))$ is sufficient.

**Theorem 5** *The sufficient conditions for a policy $Y_t$, to have regret $O(log(T))$ in the traditional multi armed bandit problem is*
*1)* $\sum_{t=1}^{\infty} P(F_t^i \leq C_1 log(t)) < \infty$ *for each arm* $i$
*2)* $\sum_{t=1}^{T} E[Y_t \neq Y_t^m] \leq C_2 log(T)$

where $F_t^i$ is the empirical fisher information till time $t$ observed from arm $i$. $Y_t^m$ represents the decision of myopic policy at time $t$. Also, w.l.o.g lets assume $\mu(\theta_2) \geq \mu(\theta_1)$ i.e. arm 2 is the superior arm. Define, $Z(t) = |\bar{X}_2(t-1) - \mu(\theta_2)|$ where $\bar{X}_2(t-1)$ is the sample mean of arm 2 till time $t-1$

**Proof :**

$$
\begin{aligned}
E[T\_inf(T)] &= \sum_{t=1}^{T} E[Y_t = 1] \\
&= \sum_{t=1}^{T} (E[Y_t = 1, Y_t^m = 2] + E[Y_t = 1, Y_t^m = 1]) \\
&\leq_{(a)} C_2 log(T) + \sum_{t=1}^{T} E[Y_t = 1, Y_t^m = 1] \\
&\leq C_2 log(T) + \sum_{t=1}^{T} E[Y_t^m = 1] \\
&=_{(b)} C_2 log(T) + \sum_{t=1}^{T} (\bar{X}_1(t-1) \geq \bar{X}_2(t-1)) \\
&= C_2 log(T) + \sum_{t=1}^{T} P(\bar{X}_1(t-1) \geq \bar{X}_2(t-1), Z(t) > x) + \sum_{t=1}^{T} P(\bar{X}_1(t-1) \geq \bar{X}_2(t-1), Z(t) < x) \\
&= C_2 log(T) + A(T) + B(T)
\end{aligned}
\tag{7}
$$

where (a) follows from condition (2) and (b) follows from the definition of myopic policy and

$$A(T) := \sum_{t=1}^{T} P(\bar{X}_1(t-1) \geq \bar{X}_2(t-1), Z(t) > x), B(T) := \sum_{t=1}^{T} P(\bar{X}_1(t-1) \geq \bar{X}_2(t-1), Z(t) < x) \tag{8}$$

We bound $A(T)$ an $B(T)$ separately with the following observation, If $F_t^i > C_i log(t)$ then $M(\theta_i)J_1^\pi(t) > C_i log(t)$ which implies $J_1^\pi(t) > \frac{C_i}{M(\theta_i)} log(t)$
i.e. if for a particular policy the empirical fisher information is increasing at a particular rate then the number of pulls of that arm should also increase with the same rate. (here, same in in $O(.)$ sense and it is pathwise) s

$$\begin{aligned} A(T) \leq \sum_{t=1}^{T} P(Z(t) > x) &= \sum_{t=1}^{T} P(Z(t) > x, F_t^2 > C_1 log(t)) + \sum_{t=1}^{T} P(Z(t) > x, F_t^2 < C_1 log(t)) \\ &\leq_a \sum_{t=1}^{T} 2exp\{-\frac{x^2}{4\sigma^2} C_1 log(i)\} + \sum_{t=1}^{T} P(F_t^2 < C_1 log(t)) \end{aligned} \tag{9}$$

where (a) follows from Chernoff bound (see Appendix), the above observation and sufficient condition (1).

$$\begin{aligned} B(T) &\leq P(\bar{X}_1(t-1) \geq \bar{X}_2(t-1), \theta_2 - x < \bar{X}_2(t-1) < \theta_2 + x) \\ &\leq P(\bar{X}_1(t-1) \geq \bar{X}_2(t-1), \theta_2 - x < \bar{X}_2(t-1)) \\ &\leq P(\bar{X}_1(t-1) > \theta_2 - x) \\ &= P(\bar{X}_1(t-1) - \theta_1 > \theta_2 - \theta_1 - x, F_t^1 > C_1 log(t)) + P(\bar{X}_1(t-1) - \theta_1 > \theta_2 - \theta_1 - x, F_t^1 < C_1 log(t)) \\ &\leq exp\{-\frac{(\theta_2 - \theta_1 - x)^2}{4\sigma^2} C_1 log(t)\} + P(F_t^1 < C_1 log(t))) \end{aligned} \tag{10}$$

(by chernoff bound again and similar conditions as above)
By suitably choosing constants $0 < x < \theta_2 - \theta_1$ and $C_1 \geq max\{8\frac{\sigma^2}{x^2}, 8\frac{\sigma^2}{(\theta_2 - \theta_1 - x)^2}\}$ and using condition (1) both the above terms are summable giving as $O(log(T))$ regret.

■

Empirical fisher information is a pathwise quantity, the sufficient conditions ensure the growth of the empirical fisher information, below we refer to the martingale theory to relate it to the growth of expected fisher information.

**Corollary 1** *If the empirical fisher information increases at the rate of condition (1), then the expected fisher information increases at $O(log(t))$ asymptotically.*

**Proof :**

$$\sum_{i=1}^{\infty} P(|F_t^i - J_i^\pi F(\theta_i)| > J_i^\pi \epsilon) = \sum_{i=1}^{\infty} P(|F_t^i - J_i^\pi F(\theta_i)| > J_i^\pi \epsilon, F_t^i > C_i log(t)) + \sum_{i=1}^{\infty} P(|F_t^i - J_i^\pi F(\theta_i)| > J_i^\pi \epsilon, F_t^i < C_i lo$$
$$\leq \sum 2exp\{-C_i \frac{J_i^{\pi 2} \epsilon^2}{4\sigma^2}\} + P(F_t^i < C_i log(t)) \tag{11}$$

where the first term is bounded by Azuma - Hoeffeding inequality (See Lemma in Appendix) and the second term is bounded by sufficient condition (1). Essentially, the LHS is summable which implies,

$$\lim_{t->\infty} P(|F_t^i - J_i^\pi F(\theta_i)| < J_i^\pi \epsilon) = 1 \text{ (by B-C lemma)} \tag{12}$$

Now, observe
$P(\frac{F_t^i - O(log(t))}{J_i^\pi} > \epsilon) \leq P(\frac{F_t^i - J_i^\pi F(\theta_i)}{J_i^\pi} > \epsilon/2) + P(\frac{J_i^\pi F(\theta_i) - O(log(t))}{J_i^\pi} > \epsilon/2) \forall \epsilon$

For some suitable $\epsilon$, we consider the limiting behavior and using $\lim_{t->\infty} P(|F_t^i - J_i^\pi F(\theta_i)| > J_i^\pi \epsilon) = 0$ and sufficient condition (1) we have

$1 = \lim_{t->\infty} P(F_t^i - O(log(t)) > \epsilon) \leq 0 + \lim_{t->\infty} P(J_i^\pi F(\theta_i) - O(log(t)) > \epsilon/2)$

hence, $\lim_{t->\infty} P(J_i^\pi F(\theta_i) - O(log(t)) > \epsilon/2) = 1$

Now by Markov's inequality, $(O(log(t)) + \epsilon) P(J_i^\pi F(\theta_i) - O(log(t)) > \epsilon/2) \leq E[J_i^\pi F(\theta_i)]$
Now as $t$ tends to $\infty$, the expected fisher information $E[J_i^\pi F(\theta_i)]$ also grows as O(log(t))

**Simple examples of optimal policies**

(a) Forced sampling policy
Consider the following policy
(b) Finite time analysis policy.

We would like to highlight before the end of this section that similar sufficient conditions can characterize also other variants of aforementioned bandit problems. In some sense, these conditions can form the basis and drive in necessary intuition for the design of optimal algorithms for new experiments.

### Necessary condition of optimality

**Theorem 6** *Given a policy $\pi^*$, the growth of empirical fisher information of that policy (for each arm) should be such that, its deviation from the expected fisher information should decay exponentially (in the number of pulls of the arm) for the adversarial configuration.*

Where adversarial configuration is the worst possible configuration as chosen by the adversary for that particular policy. The intuition being, if it does not decay exponentially then there is a smaller order growth in pathwise inferior sampling rate which would contradict the lower bounds obtained earlier. If it indeed decays exponentially quicker, its a good sign towards optimality, still some set of sufficiency conditions has to be met.

**Proof :** We prove that if the claimed policy $\pi^*$ is indeed optimal then it should satisfy the above statement. Since, the policy is optimal, the mix-max lower bound obtained is $O(log(t))$.

In particular the $E[T\_inf(t)] \geq Clog(t)$ for the worst case configuration. It follows that $E[I(Y_1 = 1) + I(Y_2 = 1) + I(Y_3 = 1) + .....I(Y_t = 1)] \geq Clog(t)$ which clearly implies atleast $Clog(t)$ indicators have to be $1's$ hence, $J_1^\pi(t) > Clog(t)$

We follow the similar ideas as before to form the Fisher information martingale and use Azuma-Hoeffeding inequality to conclude the result.

$$P(|F_t^i - J_i^\pi F(\theta_i)| > J_i^\pi \epsilon) = P(|F_t^i - J_i^\pi F(\theta_i)| > J_i^\pi \epsilon, J_i^\pi(t) > C_i log(t)) \tag{13}$$

Note: We cannot use the strong law of large numbers because of lack of independence in the subsequent observations. ■

### Insights when we have additional initial information

Earlier, we discussed the complexity of the bandit learning problem by establishing a $O(log(n))$ lower bound. In this section, we explore certain minimal conditions on the structure of the parameter space, where optimal learning is possible in finite time ($S_n \leq M$ for some $M, \forall n$). The proofs below are for bandits for 2 arms and can be extended for arbitrary number of arms. In the following cases, we refer to $(\theta_1, \theta_2)$ as the true parameter pair in the arms. Define $X_i^j$ as the reward obtained by $j^{th}$ pull of arm $i$ and $S_i(n) = \Sigma_{j=1}^{n-1} X_i^j$

#### Information on the convex combination of the parameters

Suppose, $\gamma = \lambda\mu(\theta_1) + (1 - \lambda)\mu(\theta_2)$ (where $0 \leq \lambda \leq 1$) is an input to the bandit problem, then the regret obtained by strategy *opt_conv* is finite.
*opt_conv* strategy: Pull each arm once, Pull the arm with higher realization (say $r$) till $t = \tau$,

Pull the other arm for all $t \geq \tau + 1$ where $\tau = inf\{n \geq 3 : S_r(n) \leq (n-1)\gamma\}$

*Proof* : The proof technique here maps the learning problem to that of that of hitting times of random walks. We then use finiteness of hitting time moments to prove optimality of the above strategy. Without loss in generality, lets say $\mu(\theta_1) \geq \mu(\theta_2)$, then

$$E[T\_inf(t)] = E[\pi_t \neq \pi_t^*] = 1 + P(X_2^1 \geq X_1^1)E[min(\tau, t)] \leq 1 + P(X_2^1 \geq X_1^1)E[\tau]$$

E[$\tau$] is finite since $S(n) - n\gamma$ is a random walk martingale with negative mean $(\mu(\theta_2) - \gamma)$ [4] To be more specific, there exists $M$ such that E[$\tau | X_2^1 \geq X_1^1, \gamma] \leq M$. Hence,

$$E[\tau] \leq \int_{\gamma}^{\infty} M dF_2(x) + \int_{-\infty}^{\gamma} [F_2(\gamma) + \int_{\gamma}^{\infty} M dF_2(x)] dF_2(x) = M(1 - F_2(\gamma)^2) + F_2(\gamma)^2)$$

*rand_opt_conv* strategy: Fix a large positive number $U$. Let $\mu_j(t)$ be the estimated mean of arm $j$ at time $t$. Define
$p(t) = 1, \mu_2(t) \geq U$
$p(t) = \frac{\mu_2(t)}{U}, \gamma < \mu_2(t) < U$
$p(t) = 0, \mu_2(t) \leq \gamma$

Pull each arm once, set $p = 1$ in favor of the higher realization. Update the probability $p(t)$ as defined above after every pull then on. Pull arm 2 with probability $p(t)$ and arm 1 with probability $1 - p(t)$ till the end of the horizon. Following the analysis above, we get

$$E[T\_inf(t)] \leq 1 + P(X_1^1 \geq X_2^1)\Sigma_{j=1}^{t} E[\frac{\mu_1(j)}{U}]$$

A lower bound for the above problem,
For the case of normal populations with $\lambda = 1/2(\gamma = (\theta_1 + \theta_2)/2)$ the following strategy gives a much tighter bound. Let $\delta = (\theta_1 - \theta_2)/2$
Strategy: Define $L_n = \Sigma_{i=1}^{J_1^{\pi}}(X_i^1 - \gamma) + \Sigma_{i=1}^{J_2^{\pi}}(X_i^2 - \gamma)$. Pull an arm uniformly at random, If $L_n \geq 0$, then pull arm 1 at stage $n$, otherwise pull arm 2. If arm 1 has greater mean, then

$$\begin{aligned}
E[regret] &= \frac{1}{2} + \Sigma_{i=1}^{n} P(L_i < 0) \\
&= \frac{1}{2} + \Sigma_{i=1}^{n} F(-\delta\sqrt{(}i)) \\
&= \frac{1}{2} + \Sigma_{i=1}^{n} \overline{F}(\delta\sqrt{(}i)) \\
&\leq \frac{1}{2} + \Sigma_{i=1}^{\infty} \frac{1}{\delta\sqrt{(}i)} exp - \frac{\delta^2 i}{2} \\
&\leq \frac{1}{2} + \frac{2}{(k-k')\gamma} \frac{1}{exp((k-k')^2\gamma^2)/8 - 1}
\end{aligned}$$

Consider the configurations $(H_0 : (\theta_1, \theta_2), H_1 : (\theta_2, \theta_1)$, the KL divergence between the arms is $(\theta_2 - \theta_1)^2$
Using the inequality () above,

$$\begin{aligned} S_n(\Pi, \Theta) &\geq \Sigma_{t=1}^n E[J_1^\pi](\theta_1 - \theta_2)^2 + E[J_2^\pi](\theta_1 - \theta_2)^2 \\ &= \Sigma_{t=1}^n E[J_1^\pi](\theta_1 - \theta_2)^2 + (t - E[J_1^\pi])(\theta_1 - \theta_2)^2 \\ &= \Sigma_{t=1}^n exp(-t(\theta_1 - \theta_2)^2) \\ &= \frac{1}{exp((k-k')^2\gamma^2) - 1} \end{aligned}$$

where $\theta_2 = k\gamma, \theta_1 = k'\gamma \ k > 1 and k' < 1$
For any other general distribution
Proceeding similarly as before.

$$\begin{aligned} S_n(\Pi, \Theta) &\geq \Sigma_{t=1}^n E[J_1^\pi](I_1(\theta_1, \theta_2) - I_2(\theta_2, \theta_!)) + tI_2(\theta_2, \theta_1) \quad = \Sigma_{t=1}^n exp(-(I_1 - I_2)E[J_1^\pi] - tI_2) \\ &= \Sigma_{t=1}^n exp(-(I_1 - I_2)S_t - tI_2) \\ &= \Sigma_{t=1}^n exp(-(I_1 - I_2)S_n - tI_2) \\ &\geq \frac{1}{exp(I_2) - 1} \end{aligned}$$

normal population
also plots

**Information on the parameter family - set of discrete parameters**

**Information on the first highest arm, second highest**

follow lai and second highest finite analysis.

## Lower bounds in an adversarial setup

In the following version of the traditional bandit problem, once the horizon of the problem is specified, an adversary is allowed to choose the parameter configuration governing the rewards from the arms. In such a situation, an adversary can choose the parameters as close to each other not allowing the decision maker to learn the parameters easily in the exploration phase.
Lets say that the adversary picks the parameter with a $O(\frac{1}{\sqrt{(n)}})$ separation. Lets say, the parameters chosen in the arms be of the form $(\theta_1 - \frac{1}{2\sqrt{(n)}}, \theta_1 + \frac{1}{2\sqrt{(n)}})$

## References


[1] R. Agrawal, M.V. Hedge, and D. Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10):899–906, 1988.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[3] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge Univ Pr, 2006.

[4] R. Durrett and R. Durrett. *Probability: theory and examples*. Cambridge Univ Pr, 2010.

[5] J.C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

[6] A. Goldenshluger and A. Zeevi. Performance limitations in bandit problems with side observations. *Unpublished manuscript*, 2008.

[7] Alexander Goldenshluger and Assaf Zeevi. Woodroofes one-armed bandit problem revisited. *The Annals of Applied Probability*, 19(4):1603–1633, August 2009.

[8] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[9] Tze Leung Lai. SEQUENTIAL ANALYSIS : SOME CLASSICAL PROBLEMS. *Sequential Analysis*, 11:303–408, 2001.

[10] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[11] H. Robbins and D.O. Siegmund. Sequential tests involving two populations. *Journal of the American Statistical Association*, 69(345):132–139, 1974.

[12] J. Takeuchi and A.R. Barron. Robustly minimax codes for universal data compression. *The*, 21:2–5, 1998.

[13] A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer Verlag, 2009.

[14] M. Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

## Appendix

The probability density function of the vector of observations $X_1, X_2.....X_t$ is given by

$$\begin{aligned} f(X_1, X_2.....X_{t+1}) &= f(X_{t+1}|\mathcal{F}_t) f(X_1, X_2, ...X_t) \\ &= f(X_1^{J_1^\pi(t)+1} I\{Y_{t+1}=1\} + X_2^{J_2^\pi(t)+1} I\{Y_{t+1}=2\}|\mathcal{F}_t) f(X_1, X_2, ...X_t) \quad (14) \\ &= f(X_1^1) f(X_1^2)....f(X_1^{J_1^\pi(t+1)}) f(X_2^1) f(X_2^2)....f(X_2^{J_2^\pi(t+1)}) \end{aligned}$$

since, $Y_{t+1}$ is $\mathcal{F}_t$ measurable, the above follows. (where $J_k^\pi(t)$ denotes the number of pulls of arm $k$ till time $t$.

Azuma hoeffeding lemma
forced sampling proof.
finite time analysis proof