Asymptotically Efficient Adaptive Allocation Rules

Can Chen, Bo Lu, Jiaxin Liang, Henry Oskar Singer

December 9, 2017

1 Introduction

1.1 Multi-armed Bandit Problem

We consider a general multi-armed bandit (MAB) problem. There are k independent arms. At every (discrete) time step $n = 1, 2, 3, ...$ we pull one out of the k arms. Let $I_n \in \{1, ..., k\}$ be the arm pulled at the n time step. After pulling an arm I_n, we observe a reward x_n. The reward could be a random variable x_n with density function $f(x; \theta_{I_n})$. For infinite time horizon, our goal is to maximize the total reward $S_n := \sum_{i=1}^{\infty} x_n$ by choosing the optimal strategy to select arm at each time n.

1.2 Exploration-Exploitation Trade-off

The MAB problem refers to a situation in which a gambler has access to k different slot machines and faces the problem of deciding which slot machine to play in each period so as to maximize the expected payoff. MAB problems as a whole are an important abstraction for decision problems that incorporate an *exploration-exploitation* trade-off. This trade-off can be seen clearly in the gambling scenario described above. After the gambler has discovered a slot machine whose average payoff is fairly good, there is a tension between continuing playing this slot machine (exploitation) versus trying other alternatives that have never been tested or that have only been tested infrequently (exploration). Because this type of trade-off is pervasive in the world of on-line decision problems, there are many applications of multi-armed bandit algorithms: some of these applications are described in **??**

1.3 Objectives and Roadmap

There are two well-discussed approaches in the MAB literature: the Gittens index and regret analysis. In this term project, we give a critical review of regret analysis in MAB problems. Based on the fundamental paper [Lai and Robbins, 1985], we first introduce the basic setting of the MAB problem with highlights on key concepts and assumptions. Then we introduce the constructions of asymptotically efficient allocation rules based on regret analysis. From an overview of the relationship between regret and confidence level we show the insights behind the construction of an *upper confidence bound* (UCB) and illustrate how the assumptions help with the the allocation rule design.

From Section ?? to section ??, we introduce fourfoundational regret analysis papers. We first discuss [Lai and Robbins, 1985], which provides the key assumptions, definitions, and algorithms on which the follow-up works build. [Anantharam et al., 1987a, Anantharam et al., 1987b] relax one of [Lai and Robbins, 1985]'s assumptions on the allocation rule by allowing the number of arms pulled at each time n to be any arbitrary $m \in \{1, \dots, k\}$. [Agrawal et al., 1988] changes the regret structure by introducing switching cost. With [Lai and Robbins, 1985] as the benchmark, for each of the other three papers, we highlight the difference of basic assumptions and analyze its impact on optimal allocation rule design. Furthermore, we reveal the structural similarity of the algorithms.

2 Model Formulation

This section reviews the basic k-Armed Bandit Problem including formulation of the problem and some important technical assumptions. The problem can be summarized as follows:

- Let Π_j be statistical populations specified by univariate density functions $f(x;\theta_j)$ with respect to some measure $\nu(x)$, where the density functions are known and the parameter $\theta_j \in \Theta$ are unknown for some set Θ for all $j = 1, 2, \ldots, k$. We sample $x_1, x_2, \ldots$ sequentially from the k populations.
- An adaptive allocation rule ϕ is a sequence of random variables $\phi_1, \phi_2, \ldots$ taking values from the set $\{1, 2, \ldots, k\}$ such that the event $\{\phi_n = j\}$ belongs to the $\sigma-$ field F_{n-1} generated by previous ϕ_i and x_i for $i = 1, 2, \ldots, n-1$.
- Assume that the rewards are integrable $\mu(\theta) = \int_{-\infty}^{\infty} |x| f(x;\theta) d\nu(x) < \infty$ and the expectation $\mu(\theta) = \int_{-\infty}^{\infty} x f(x;\theta) d\nu(x)$ is strictly monotone increasing function in $\theta \in \Theta$ and define $S_n = x_1 + \ldots + x_n$. Then by Wald's lemma

$$\mathbb{E}S_n = \sum_{j=1}^{k} \mu(\theta_j) \mathbb{E}T_n(j) \tag{2.1}$$

  where $T_n(j) = \sum_{i=1}^{n} \mathbf{1}[\phi_i = j]$ denotes the number of time that the rule ϕ samples from Π_j up to time n.
- Let $\mu^* = \max\{\mu(\theta_1), \ldots, \mu(\theta_k))\} = \mu(\theta^*)$ for some $\theta^* \in C = \{\theta_1, \ldots, \theta_k\}$. Define the sample regret

$$R_n(\theta_1, \ldots, \theta_k) = n\mu^* - \mathbb{E}S_n = \sum_{j:\mu(\theta_j)<\mu^*} [\mu^* - \mu(\theta_j)] \mathbb{E}T_n(j) \tag{2.2}$$

The objective of the problem is to maximize the expected value of the sum S_n, which is equivalent to minimizing the regret R_n. Some definitions in the above formulation may vary under settings of multiply plays [Anantharam et al., 1987a, Anantharam et al., 1987b] and switching cost [Agrawal et al., 1988].

Before stating the technical assumptions, we need to introduce the Kulback-Leibler number which is used to measure the distance between two distributions. The Kulback-Leibler number is defined by

$$I(\theta, \lambda) = \int_{-\infty}^{\infty} \log \left[\frac{f(x;\theta)}{f(x;\lambda)} \right] f(x;\theta) d\nu(x) \tag{2.3}$$

The primary results of [Lai and Robbins, 1985] are derived under the following technical assumptions:

- **A1** $0 < I(\theta, \lambda) < \infty$ whenever $\mu(\lambda) > \mu(\theta)$.
- **A2** For every $\epsilon > 0$ and $\theta, \lambda \in \Theta$ such that $\mu(\lambda) > \mu(\theta)$, there exists $\delta > 0$, such that

$$|I(\theta, \lambda) - I(\theta, \lambda')| < \epsilon \tag{2.4}$$

  whenever $\mu(\lambda) \leq \mu(\lambda') \leq \mu(\lambda) + \delta$.
- **A3** Θ is such that for any $\lambda \in \Theta$ and any $\delta > 0$, there exits λ' such that $\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta$.
- **A4** The parameter configuration $C = (\theta_1, \ldots, , \theta_k)$ is such that $\mu(\theta_j) < \mu^* = \mu(\theta^*)$ for all $\theta_j \neq \theta^*$.

When $\mu(\lambda) > \mu(\theta)$, $I(\theta, \lambda) > 0$ is trivially satisfied, and the condition $I(\theta, \lambda) < \infty$ suggests that the distribution of the samples with parameter θ is absolutely continuous with respect to the distribution of the samples with parameter λ. Most parametric families of distributions with mutually absolute continuity are expected to satisfy this condition. **A2** shows that $I(\theta, \lambda)$ is continuous in λ whenever $\mu(\lambda) > \mu(\theta)$ for fixed θ. **A3** is the denseness condition on the space Θ. **A2** and **A3** are required to draw the lower bound on the total regret. The last condition **A4** just implies that there is only one best population, which is crucial in finding the upper bound on the total regret. Some of the assumptions may vary in the settings of multiple plays and stitching cost. We will discuss it in detail later in the following sections.

3 Regret Analysis

3.1 Basic MAB Problem

In this section, we review the work by [Lai and Robbins, 1985]. This is the first paper that introduced the concept of "regret", and provided the mathematical framework for analysis of regret. The goal of the paper is to construct an asymptotically efficient rule, which is defined as that for any fixed values $\theta_1, \ldots, \theta_k$ such that the $\mu(\theta_j)$ are not all equal,

$$R_n(\theta_1, \ldots, \theta_k) \sim \left[\sum_{j:\mu(\theta_j)<\mu^*} \frac{\mu^* - \mu(\theta_j)}{I(\theta_j, \theta^*)} \right] \log n \qquad \text{as } n \to \infty. \tag{3.1.1}$$

Before digging into the asymptotically efficient rule, we would like to introduce what a "good" rule looks like. As mentioned in the introduction, to maximize our total rewards, we need to balance the trade-off between exploration and exploitation. Since, $\theta = (\theta_1, \ldots, \theta_k)$ are unknown parameters to us, we can never be certain about their true means. However, by law of large number, the more we sample from a population, the more we can approximate its true mean. Therefore, to make sure the inferior population is indeed bad, we need to sample infinite amount of times from that population as n goes to infinity. This is the nature of exploration. On the other hand, if we sample too often from an inferior population, our total rewards will definitely be hurt. Hence, we want to sample from the best population as frequently as possible while we still keep sound estimations of means of the bad populations. This is the nature of exploitation. It is intuitive to infer that a "good" rule should at least satisfy that its regret becomes infinite while its average regret per unit time becomes zero as time goes to infinity. Unfortunately, we cannot emphasize on both exploration both and exploration at the same time. If one is emphasized more, then the other one would be undermined to some degree. The goal of this paper is to find the allocation rule that perfectly balance between exploration and exploitation so as to maximize our total rewards as n goes to infinity.

To further narrow down the scope of rules from where we start to search for the best rule, [Lai and Robbins, 1985] restricted attention to the class of uniformly good rules, whose regret satisfies that for each fixed parameter configuration $\theta = (\theta_1, \ldots, \theta_k)$, as $n \to \infty$,

$$R_n(\theta) = o(n^a) \qquad \text{for every } a > 0 \tag{3.1.2}$$

Remark Notice that $\lim_{n\to\infty} o(n^a) = \infty$, and that $\lim_{n\to\infty} \frac{o(n^a)}{n} = 0$ for any $a > 0$. Since a can be arbitrarily small, we can infer that the class of uniformly good rules contains the best rules from the broader class of "good" rules.

However, it is difficult to infer a rule from the regret in terms of rewards. To make it easier, [Lai and Robbins, 1985] bridged between the regret as defined and the numbers of times the inferior populations are sampled by the following equation:

$$R_n(\theta) = \sum_{j:\mu(\theta_j)<\mu^*} (\mu^* - \mu(\theta_j)) \mathbb{E}_\theta T_n(j) \tag{3.1.3}$$

It comes naturally that the definition of uniformly good rules can be therefore expressed in terms of expected sample size from inferior population: for any $\theta \in \Theta$ such that θ_j is the "unique best" i.e. $\mu(\theta_j) > \max_{i \neq j} \mu(\theta_i)$, as $n \to \infty$,

$$\sum_{i \neq j} \mathbb{E}_\theta T_n(i) = o(n^a) \qquad \text{for every } a > 0 \tag{3.1.4}$$

Next, we will present that the asymptotically efficient rule defined in (3.1.1) is the "best" among the class of uniformly good rule. In the following theorem 1, [Lai and Robbins, 1985] derived a lower bound for the expected sample size from inferior populations.

Theorem 1. *Assume that $I(\theta, \lambda)$ satisfies A1 and A2 and that Θ satisfies A3. Then for any*

$$\liminf_{n \to \infty} \frac{\mathbb{E}_\theta T_n(j)}{\log n} \geq \frac{1}{I(\theta_j, \theta^*)} \qquad \text{for all } j \text{ such that } \mu_j < \mu^* \tag{3.1.5}$$

Furthermore, the lower bound of regret for uniformly good rules can also be derived.

Theorem 2. *Assume that $I(\theta, \lambda)$ satisfies A1 and A2 and that Θ satisfies A3. Then for any uniformly good rule,*

$$\liminf_{n \to \infty} \frac{\mathbb{E}_\theta R_n(\theta)}{\log n} \geq \frac{1}{I(\theta_j, \theta^*)} \sum_{j: \mu(\theta_j) < \mu^*} (\mu^* - \mu(\theta_j)). \tag{3.1.6}$$

Remark Theorem 2 can be easily obtained from Theorem 1 and (3.1.3). The implication of Theorem 2 is that the asymptotically efficient rules we are looking for can be found in the class of uniformly good rules. Actually, if the lower bound from Theorem 2 is attained, then the rule is asymptotically efficient. In other words, the asymptotically efficient rule is indeed the best among uniformly efficient rules. Equivalently, the rule is asymptotically efficient as long as the expected sample size from inferior populations achieves the lower bound from Theorem 1.

Now, we know that if we can find a rule in the class of uniformly good rule such that the regret achieves the lower bound in theorem 2, then the rule is asymptotically efficient. [Lai and Robbins, 1985] in the paper proposed an uniformly good rule that would be bounded above and that the gap between the upper bound and the lower bound can be ignored. The difficulty arises is that how to make a rule bounded above as we want.

[Lai and Robbins, 1985] introduced two significant statistics that keep updating at each stage. The first one is called "upper confidence bounds" for the true mean $\mu(\theta)$, which is defined by Borel functions as:

$$g_{ni} : R^i \to R \quad (n = 1, 2, \ldots; i = 1, \ldots, n) \tag{3.1.7}$$

such that for every $\theta \in \Theta$, g_{ni} satisfies the following properties:

$$P_\theta\{r \leq g_{ni}(Y_1, ..., Y_i) \text{ for all } i \leq n\} = 1 - o(n^{-1}) \qquad \text{for every } r < \mu(\theta) \tag{3.1.8}$$

$$\lim_{\epsilon \to 0} \left(\limsup_{n \to \infty} \frac{\sum_{i=1}^n P_\theta\{(g_{ni}(Y_1, ..., Y_i) \geq \mu(\lambda) - \epsilon\}}{\log n} \right) \leq \frac{1}{I(\theta, \lambda)} \quad \text{whenever } \mu(\lambda) > \mu(\theta) \tag{3.1.9}$$

$$g_{ni} \text{ is non-decreasing in } n \leq i \text{ for every fixed } i = 1, 2, \ldots \tag{3.1.10}$$

Remark (3.1.8)-(3.1.10) play an important role in establishing the upper bound on regret. (3.1.8) makes sure that the upper confidence bound rarely falls below the true mean. Actually, as time goes on, (3.1.8) ensures that it falling below the true mean becomes less and less frequent. Notice that (3.1.9) puts an upper bound on the sum of probabilities of the upper confidence bound of an inferior population being greater than the true mean of the superior population. Because the rule which will be introduced shortly decides

to sample from the inferior population only if the upper bound of the inferior population is greater than the estimated mean of the best population we believe at current stage, this upper bound forces the inferior population to be sampled less frequently as time goes on, and the more inferior the population is the less frequently it will be sampled. (3.1.10) makes sure the inferior population is not to be forever ignored once its upper confidence bound falls below the estimated mean of the best population we believe at current stage, so that it still gets chance to be sampled in the future.

Another statistic [Lai and Robbins, 1985] introduced is $h_i(Y_1, \ldots, Y_i)$, the point estimate of $\mu(\theta)$, where $h_i : R^i \to R$ $(i = 1, 2, ...)$ are Borel functions such that

$$h_i \leq g_{ni} \qquad \text{for all } n \geq i, \tag{3.1.11}$$

and for every $\theta \in \Theta$, h_i satisfies:

$$P_\theta\Big\{ \max_{\delta n \leq i \leq n} |h_i(Y_1, \ldots, Y_i) - \mu(\theta)| \geq \epsilon \Big\} = o(n^{-1}) \qquad \text{for all } \epsilon > 0 \text{ and } 0 < \delta < 1 \tag{3.1.12}$$

Remark h helps us get the sense of which population has the best chance to be the true best population at each stage, and thus we can exploit it. Also, note that law of large number suggests that a good candidate for h would be the sample mean.

With the help of the two statistics, [Lai and Robbins, 1985] constructed a rule as follows.

Define:

$$\hat{\mu}_n(j) = h_{T_n(j)}(Y_{j,1}, \ldots, Y_{j,T_n(j)})$$
$$U_n(j) = g_{n,T_n(j)}(Y_{j,1}, \ldots, Y_{j,T_n(j)})$$

Algorithm 1: Asymptotically Efficient Rule ϕ^* for Basic k-Armed Bandit Problem

1. Choose $0 < \delta < 1/k$
2. Define the leader θ_{j_n} at stage n: $\hat{u}_n(j_n) = \max_{j \in \{1,\ldots,k\}} \{\hat{\mu}_n(j) : T_n(j) \geq \delta n\}$

for $n = 1, \ldots, k$ **do**
  take a sample from $\prod_n$
end
for $n = k+1, k+2, \ldots$ **do**
  Set $j = n \mod k$
  if $\hat{u}_n(j_n) \leq U_n(j)$ **then**
    take a sample from $\prod_j$
  else
    take a sample from $\prod_{j_n}$
  end
end

Finally, theorem 3 proves that the rule just constructed is uniformly good, and it has an desirable upper bound. Most importantly, it is an asymptotically efficient rule.

Theorem 3. *Assume that $I(\theta, \lambda)$ satisfies A1 and A2 and that the functions g_{ni} and h_i satisfy (3.1.7)-(3.1.12). For $j = 1, ..., k$, let $T_n(j)$ be the number of times that the rule ϕ^* samples from Π up to stage n.*

(i) For every $\theta = (\theta_1, \ldots, \theta_k)$ and every j such that $\mu(\theta_j) < \mu(\theta^*)$,

$$\mathbb{E}_\theta T_n(j) \leq \left(\frac{1}{I(\theta_j, \theta^*)} + o(1) \right) \log n. \tag{3.1.13}$$

(ii) Assume also that Θ satisfies A3. Then $\mathbb{E}_\theta T_n(j) \sim \frac{\log n}{I(\theta_j,\theta^*)}$ for every j such that $\mu(\theta_j) < \mu(\theta^*)$, and the regret of ϕ^* satisfies (3.0).

Statement (i) shows that there is an upper bound on the rule constructed previously, and the only difference between the lower bound and the upper bound is the $o(1)\log n$ appearing in upper bound. (ii) states that as long as Θ satisfies A3, the rule is in the class of uniformly good rule and the small additional term in the upper bound does not affect the performance, so that rule is asymptotically efficient. One more important feature can be observed is that the more inferior the population is, the less frequently will it be sampled, as indicated by $\frac{1}{I(\theta_j,\theta^*)}$. It is mainly (3.1.9) that ascribes the rule this property. Moreover, even the worst population would be sampled infinite amount of times as n becomes infinite, and it is mainly (3.1.10) makes this property possible.

3.2 Multiple Plays with I.I.D. Rewards

In [Lai and Robbins, 1985], the problem of single play was studied, but now we are required to play a fixed number, m, of the arms, $1 \le m \le k$, at each time. The assumptions A1, A2 and A3 are followed by [Lai and Robbins, 1985]. However, the assumption A4 is modified in this setting. It is now assumed that there is a permutation σ of $\{1, 2, \ldots, k\}$ such that

$$\mu(\theta_{\sigma(1)}) \ge \mu(\theta_{\sigma(2)}) \ge \ldots \ge \mu(\theta_{\sigma(k)}) \tag{3.2.1}$$

Moreover, [Anantharam et al., 1987a] separated arms into three types.

- If $\mu(\theta_{\sigma(m)}) > \mu(\theta_{\sigma(m+1)})$, the arms $\theta_{\sigma(1)}, \ldots, \theta_{\sigma(m)}$ are called the distinctly $m-best$ arms and $\theta_{\sigma(m+1)}, \ldots, \theta_{\sigma(N)}$ the distinctly $m-worst$ arms.
- If $\mu(\theta_{\sigma(m)}) = \mu(\theta_{\sigma(m+1)})$, there exist l, n such that

  $$\mu(\theta_{\sigma(1)}) \ge \ldots \ge \mu(\theta_{\sigma(l)}) > \mu(\theta_{\sigma(l+1)}) = \ldots = \mu(\theta_{\sigma(m)}) = \mu(\theta_{\sigma(k)}) > \mu(\theta_{\sigma(n+1)}) \ge \ldots \ge \mu(\theta_{\sigma(k)})$$

  for $0 \le l < m$ and $m \le n \le k$. We call the arms $\theta_{\sigma(1)}, \ldots, \theta_{\sigma(l)}$ are called the distinctly $m-best$ arms and $\theta_{\sigma(n+1)}, \ldots, \theta_{\sigma(k)}$ the distinctly $m-worst$ arms.
- The arms with mean equal to $\mu(\theta_{\sigma(m)})$ are called the $m-border$ arms.

Essentially, we have relaxed the assumption A4 because one must play the arms with m highest means at one time instead of the best one. It is somewhat more general than the single play case discussed in [Lai and Robbins, 1985]. On the other hand, we can think of the separation of arms into three types as an extension of the uniqueness assumption A4, where the $m-best$ arms must be distinct from the $m-worst$ arms. When $m = 1$, it degenerates to the assumption A4. According to the current assumption, one can write down the new version of regret.

$$R_n(\theta_1, \ldots, \theta_k) = n \sum_{i=1}^{m} \mu(\theta_{\sigma(i)}) - \mathbb{E}S_n \tag{3.2.2}$$

Again, the goal of the problem is to minimize the regret by using some uniformly good rules, but the definition of uniformly good need to be extended in the setting of multiple plays.

We call an adaptive allocation rule Φ is uniformly good for multiple plays if and only if for every distinctly $m-best$ arm j

$$\mathbb{E}(n - T_n(j)) = o(n^a)$$

and for every distinctly $m-worst$ arm j

$$\mathbb{E}(T_n(j)) = o(n^a)$$

for every real $a > 0$.

For each distinctly $m - best$ arm, the expectation of the number of plays under which the arm j is excluded goes to infinity, but we enforce its increasing rate of that expectation to be controlled by the term $o(n^a)$, i.e. $\lim_{n\to\infty} \frac{\mathbb{E}(n-T_n(j))}{n^a} = 0$. On the other hand, the increasing rate of the expectation of the number of plays for each $m - worst$ arm should also be controlled by the term $o(n^a)$, i.e. $\lim_{n\to\infty} \frac{\mathbb{E}(T_n(j))}{n^a} = 0$ for every $m - worst$ arm. Furthermore, a lower bound of that expectation for uniformly good rules can be thus derived. In particular, the expression is very similar to the one in Theorem 1 except that θ^* is replaced by $\theta_{\sigma(m)}$ because we are measuring the dissimilarity between distributions with parameters θ_j and the least $m - best$ arm $\theta_{\sigma(m)}$ instead of the unique best arm θ^*. Consequently, a generalized Theorem 2 can also be stated as follows:

Theorem 4. *Assume that $I(\theta, \lambda)$ satisfies A1 and A2 and that Θ satisfies A3. Then for any uniformly good rule,*

$$\liminf_{n\to\infty} \frac{\mathbb{E}T_n(j)}{\log n} \geq \frac{1}{I(\theta_j, \theta_{\sigma(m)})} \Leftrightarrow \tag{3.2.3}$$

$$\liminf_{n\to\infty} \frac{R_n(\theta_1, ..., \theta_k)}{\log n} \geq \sum_{j \ is \ m-worst} \frac{[\mu(\theta_{\sigma(m)}) - \mu(\theta_j)]}{I(\theta_j, \theta_{\sigma(m)})} \tag{3.2.4}$$

for every parameter configuration θ.

The adaptive allocation rule is asymptotically efficient if the expectation of the number of plays from the distinctly $m - worst$ arms has the lower bound with (3.2.3). Following [Lai and Robbins, 1985], we need an upper bound for (3.2.4) to achieve asymptotic efficiency. [Anantharam et al., 1987a] proposed a scheme in constructing an upper confidence bound for exponential families of distributions. If $g_{nT_n(j)}$ is greater than any of the $m - best$ estimated parameter values, it is worthy to experiment with arm j. Two important assumptions are considered during the construction of g_{ni}

- **A5** $\log(f(x;\theta))$ is concave in θ for each fixed x;
- **A6** $\int x^2 f(x;\theta) d\nu(x) < \infty$ for each $\theta \in \mathbb{R}$.

Let $Y_1, Y_2, ...$ be the sequence of rewards from an arm. Let

$$W_i(\theta) = \int_{-\infty}^{0} \prod_{b=1}^{i} \frac{f(Y_b, \theta + t)}{f(Y_b, \theta)} h(t) dt \tag{3.2.5}$$

where $h(t)$ is a strictly positive continuous function with $\int_{-\infty}^{0} h(t)dt = 1$. For any $K > 0$, Let

$$U(a, Y_1, ..., Y_i, K) = \inf\{\theta | W_i(\theta) \geq K\} \tag{3.2.6}$$

The assumption A6 simply shows that the variance of reward is bounded. The log concavity assumption A5 implies that $W_i(\theta)$ is an increasing function in θ. Based on the assumption A5, we can have an equivalent relation

$$U(a, Y_1, ..., Y_i, K) \leq \theta \Leftrightarrow W_i(\theta) \geq K, \tag{3.2.7}$$

The function $W_i(\theta)$, originally motivated by Pollack and Siegmund, is a commonly used statistic in hypothesis testing. When we use $U(a, Y_1, ..., Y_i, K)$ to decide whether it is necessary to experiment, for a large value of K, we will be more sure that the samples have been generated by parameter values below θ before we reject

the possibility that they have been generated by θ. Based on the heuristics, [Anantharam et al., 1987a] defined

$$g_{ni} = \mu[U(i, Y_1, ..., Y_i, n(\log\, n)^p)] \tag{3.2.8}$$

for some $p > 1$.

If we construct the upper confidence bound through the function U, g_{ni} satisfies the properties (3.1.8)-(3.1.10). On the other hand, h_i is defined as an estimate for the mean reward of an arm satisfying the same condition (3.1.12).

Consider the following adaptive allocation rule.

Define:

$$\mu_n(j) = h_{T_n(j)}(Y_{j,1}, \ldots, Y_{j,T_n(j)})$$
$$U_n(j) = g_{n,T_n(j)}(Y_{j,1}, \ldots, Y_{j,T_n(j)})$$

R2 ϕ^*: Choose $0 < \delta < 1/k^2$.

1. In the first k steps sample m times from each of the arms to establish an initial sample.
2. Choose the $m-leaders$ with the arms having m best values of the statistic $\mu_n(j)$ for $j = 1, \ldots, k$. For $n > k$, calculate the statistic $U_n(j)$.
   - If arm j is in the one of the $m - leaders$, then at time $t + 1$, play the $m - leaders$.
   - If arm j is not in the one of the $m - leaders$, and $U_n(j) < \mu_n(j)$ for all the $m - leaders$, then again play the $m - leaders$.
   - If arm j is not in the one of the $m-leaders$, and $U_n(j) \geq \mu_n(j)$ of the least best of the $m-leaders$, play the $m - 1$ best $m - leaders$ and arm j.

When we remove the denseness condition A3, the situation becomes more subtle, in which with the "asymptotically efficient" rule R2, the least best μ_n among the $m - best$ leaders will decrease so rapidly often below $\mu(\theta_{\sigma(m)})$ that we have to play the $m - worst$ arms. Thus, R2 is no longer asymptotically efficient. To tackle this problem, [Anantharam et al., 1987a] required the statistic g_{ni} to exceed the least best μ_n among the $m - best$ leaders by a margin $\gamma(t)$, where $\gamma(t)$ decreases monotonously to zero suitably, in order to sample from the inferior arms. For example, $\gamma(n) = Kn^{-a}$ for $K > 0$ and $0 < a < \frac{1}{4}$. Furthermore, a small modification is made in step 2 to re-achieve asymptotic efficiency.

Let μ_t^* be the least best μ_t of the $m - leaders$ and calculate

$$\mu_n^+ = \inf_{\theta \in \Theta}\{\mu(\theta) | \mu(\theta) > \mu_n^* + \gamma(n)\} \tag{3.2.9}$$

Then replace $\mu_n(j)$ by μ_t^+, and other steps follow. It is proved that after modification the rule is asymptotically efficient.

3.3 Multiple Plays with Markovian Rewards

In [Lai and Robbins, 1985] and [Anantharam et al., 1987a], the reward statistics are specified by univariate density functions $f(x; \theta_j)$ with the unknown parameters $\theta_j \in \Theta$. However, we are now required to play the arms with Markovian rewards. In particular, each arm has an independent Markovian reward process. Before we reformulate the problem, it is necessary to consider the following lemma.

Lemma 1. *Let F_n denote the $\sigma-$algebra generated by $Y_1, ..., Y_n$ and G a $\sigma-$algebra independent of $F_\infty = \vee_n F_n$. Let τ be a stopping time of $\{F_n \vee G\}$. Let $N(x,\tau) = \sum_{i=1}^{\tau} \mathbf{1}(Y_i = x)$ and $N(x,y,\tau) = \sum_{i=1}^{\tau-1} \mathbf{1}(Y_i = x, Y_{i+1} = y)$. Then*

$$|\mathbb{E}N(x,\tau) - \pi(x)\mathbb{E}\tau| \leq K \tag{3.3.1}$$

$$|\mathbb{E}N(x,y,\tau) - \pi(x)P(x,y)\mathbb{E}\tau| \leq K \tag{3.3.2}$$

for all τ with $\mathbb{E}\tau < \infty$

Then the problem can be summarized as follows.

- Let $P(\theta)$ be the one-parameter family of probability transition matrices, defined by $[P(x,y,\theta)], \theta \in \mathbb{R}, x, y \in X$, where $X \subset \mathbb{R}$ is a finite set of rewards. The first play of an arm with parameter θ has reward distribution $p(\theta)$. We assume that for $x, y \in X$, $\theta, \theta' \in \mathbb{R}, P(x,y,\theta) > 0 \Rightarrow P(x,y,\theta') > 0$, $P(\theta)$ is irreducible and aperiodic for all $\theta \in \mathbb{R}$, and $p(x,\theta) > 0$, for all $x \in X$ and $\theta \in \mathbb{R}$.
- For $\theta \in \mathbb{R}$, $\pi(\theta) = [\pi(x,\theta)]$, $x \in X$, denotes the invariant probability distribution on X and the mean reward
$$\mu(\theta) = \sum_{x \in X} x\pi(x,\theta) \tag{3.3.3}$$
is assumed to be strictly monotone increasing in θ.
- The total reward is
$$S_n = \sum_{j=1}^{k} \sum_{i=1}^{T_n} Y_{ji} = \sum_{j=1}^{k} \sum_{x \in X} xN(x, T_n(j)) \tag{3.3.4}$$
- The regret is defined by (3.2.2) with θ satisfying (3.2.1).

In the case of i.i.d. rewards, [Lai and Robbins, 1985] and [Anantharam et al., 1987a] established the relationship between the regret R_n and $T_n(j)$ by Wald's lemma. Here, Lemma 1 plays the same role as Wald's lemma. In particular, when we apply Lemma 1 to the total reward S_n, we can obtain

$$|\mathbb{E}S_n - \sum_{j=1}^{k} \mu(\theta_j)\mathbb{E}T_n(j)| \leq K_1 \tag{3.3.5}$$

By (3.2.2), we can have

$$|R_n(\theta_1, \ldots, \theta_k) - [n\sum_{i}^{m} \mu(\theta_{\sigma(i)}) - \sum_{j=1}^{k} \mu(\theta_{\sigma(j)})\mathbb{E}T_n(j)]| \leq K_2 \tag{3.3.6}$$

where K_1, K_2 may depend on the parameter configuration θ. Having (3.3.5) and (3.3.6), one can easily define uniformly good in term of $T_n(j)$ for the $m - best$ and the $m - worst$ arms.

The Kullback-Liebler number is defined by

$$I(P,Q) = \sum_{x \in X} \pi(x) \sum_{y \in X} P(x,y)\log \frac{P(x,y)}{Q(x,y)} \tag{3.3.6}$$

where P and Q are irreducible and aperiodic probability transition matrices with P having invariant distribution π, which satisfies $P(x,y) > 0 \Leftrightarrow Q(x,y) > 0$. Moreover, the assumption A1, A2, A3 and modified A4 in Section 3.2 are followed. Surprisingly, R2 proposed in Section 3.2 is also asymptotically efficient for multiple plays with Markovian rewards.

Despite the introduction of Markov processes in [Anantharam et al., 1987b], under the formulation above, the definition of uniformly good follows and one can obtain the same lower bound (3.2.3) for a uniformly good rule. On the other hand, the construction of g_{ni} and h_i are similar with [Anantharam et al., 1987a]. Consequently, R2 is still asymptotically efficient under Markovian rewards. Furthermore, when the denseness condition A3 is removed, same modification discussed in Section 3.2 can be applied.

3.4 Switching Cost

The work presented in [Agrawal et al., 1988] adds on to the multi-armed bandit problem a switching cost, which is incurred when a switch from one bandit to another bandit takes place. Without any additional assumption, they proposed an allocation scheme that can also achieve asymptotic efficiency, despite the inclusion of switching cost.
In the paper, the switching cost is define as:

$$SW_n(\theta) := C \sum_{j=1}^{P} \mathbb{E} S_n(j) \tag{2.4.1}$$

Where $C > 0$ is the constant switching cost, and $S_n(j)$ is the number of times of switching to bandit j within n steps, which can be expressed as $S_n(j) := \sum_{i=2}^{n} \mathbf{1}\{\phi_i = j, \phi_{i-1} \neq j\}$.
Further, the total regret is defined as:

$$R_n(\theta) := R'_n(\theta) + SW_n(\theta) \tag{2.4.2}$$

where $R'_n(\theta)$ is the regret as defined previously.

We can actually associate each arm with a fixed switching cost, and the change thus made will not affect our result as long as none of the switching cost goes to infinity. Suppose $C_1, \ldots, C_k$ are the switching costs for arm $1, \ldots, k$, respectively. We can take $C = \max\{C_1, \ldots, C_k\}$. If the rule is asymptotically efficient with a switching cost C for all arms, then it is obvious that it is also asymptotically efficient with $C_1, \ldots, C_k$ assigned for each arm.

In view of the analysis presented in [Lai and Robbins, 1985], for an asymptotically efficient rule ϕ, the number of samples that takes from any inferior population $\prod_j$ up to stage n is about $(\log n)/I(\theta_j, \theta^*)$, where θ^* is the population with greatest mean. However, when the switching cost is introduced, the rule constructed in [Lai and Robbins, 1985] is no longer asymptotically efficient. To see how the switching cost sabotages [Lai and Robbins, 1985]'s rule, we can estimate the total switching cost incurred. Guided with [Lai and Robbins, 1985]'s rule, we won't keep sampling from an inferior population for many consecutive steps. When a switch to an inferior population happens, a cost C incurred. Then, with a probability close to one, it will switch to the dominant population with another cost C incurred. Therefore, the total switching cost incurred by an inferior population up to stage n under Lai and Robbin's rule is at most about $2 \log n / I(\theta_j, \theta^*)$, and the total regret is roughly $\left\{ \sum_{j:\mu(\theta_j)<\mu^*} (\mu^* - \mu(\theta_j) + 2)/I(\theta_j, \theta^*) \right\} \log n$ as n goes to infinity. Unfortunately, this result is far from desirable.

To solve the issue, [Agrawal et al., 1988] proposed a block allocation rule, under which, the expected numbers of samples from each inferior population is ensured to be $O(\log n)$, and the total switching cost is ensured to be $o(\log n)$ so that the contribution of switching cost to the total regret can be infinitesimal and meanwhile the total cost is still proportional to $\log n$. The allocation rule they presented is as follows:
Step 1:
Before starting sampling from any population, we first divide all the time-stamps into frames, numbered, $0, 1, 2, \ldots$ and each frame is further divided into blocks of equal size, numbered, $0, 1, 2, \ldots$. Denote (f, i) as the i'th block in f'th frame.
Then, we make several more notations:

N_f : the time instant at the end of frame f

b_f : the block length of each block in frame f

Frame (f)	b_f	$N_f - N_{f-1}$
0	1	p
1	1	$\lceil \frac{2^{1^2}-2^{0^2}}{1} \rceil p.1 = p$
2	2	$\lceil \frac{2^{2^2}-2^{1^2}}{2} \rceil p.2 = 14p$
f	f	$\lceil \frac{2^{f^2}-2^{(f-1)^2}}{f} \rceil p.f$

Step 2:
During frame 0, each arm is sampled. For frames $1, 2, \ldots$, at the beginning of each block, the leader is determined via the UCBs $U_n(j), j = 1, \ldots, p$, and the leader is sampled for the entire block.

[Agrawal et al., 1988] show that their algorithm achieves the same asymptotic regret bound as without switching cost by controlling the switching cost to a logarithmic rate. They also derive identical general form of the upper confidence bound as [Lai and Robbins, 1985].

4 Conclusion

Our term project aims to give a literature review of foundational work on the UCB family of algorithms. [Lai and Robbins, 1985], first proposes regret analysis and uniform goodness as performance measure for any allocation rule and then put forward an optimal algorithm. [Anantharam et al., 1987a] relax a restriction on the allocation rule and provides a modified policy that gives uniformly good performance. [Anantharam et al., 1987b] relaxes an assumption on the reward-generating mechanism. [Agrawal et al., 1988] introduces the concept of switching cost, redefines regret to account for switching cost, gives upper and lower regret bounds for the switching cost, and provides an asymptotically efficient allocation rule and upper confidence bound for the switching cost scenario.

References

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

[Anantharam et al., 1987a] Anantharam, V., Varaiya, P., and Walrand, J. (1987a). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976.

[Anantharam et al., 1987b] Anantharam, V., Varaiya, P., and Walrand, J. (1987b). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards. *IEEE Transactions on Automatic Control*, 32(11):977–982.

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