Department of Mathematics
and Statistics

MTHE 472 Project

Asymptotically Efficient Adaptive Allocation Rules for the Multiarmed Bandit Problem with Switching cost

Charlie Horn
Mert Mumtaz
Liam Vovk

Submitted on: October 8, 2019

1 Introduction

The multiarmed bandit problem is one of the most investigated resource allocation problems in literature today. Such problem were introduced by Robbins (1952) and have since been used extensively to model the trade-offs faced by an automated agent which aims to gain new knowledge by exploring its environment and/or exploiting its current, reliable knowledge. These problems arise frequently in practical real world scenarios, for example in the context of online advertising, clinical trials, and even recommendation systems utilized by global giants such as Netflix [1]. The problem can be intuitively understood in the context of slot machines as follows. Let there be a certain number of slot machines present, and let each of them have associated expected values and variances. Initially these quantities are unknown. The player is viewed as a gambler whose goal is to collect as much reward as possible by pulling the arms of the slot machines over many turns. The player has two-fold goal: finding out which slot machines have the highest expected value, and maximizing rewards as much as possible while playing. With this, a more precise mathematical formulation for the problem can be assigned. In it's simplest form, there are $p \geq 2$ statistical populations (e.g., slot machines) whose distributions are parametrized by an unknown $\theta \in \Theta$. How should one sample $x_1, x_2, ...$ sequentially from the p populations in order to maximize the expected value of the sum $J_n = x_1 + x_2 + ... + x_n$ as n approaches infinity? It is important to note here that this corresponds to the setting of incomplete information.

Numerous varying versions of the above problem have been addressed in literature. At the time of this paper, numerous solutions had been provided by Lai and Robbins, and by Anantharam, Varaiya, and Walrand [2][3]. The latter developed theoretical lower bounds for regret associated with any uniformly good scheme (defined later), while the former developed the idea of optimism in the face of uncertainty in the context of upper confidence bounds (UCB) as well as coming up with different schemes for the concept of regret. Later on, Sutton and Barto developed reinforcement comparison methods which are built on top of the idea of pursuit algorithms [1]. Most recently, Fisher, Auer and Cesa-Bianchi (2002) proposed a simpler, more elegant implementation of UCB-style algorithms, which were extended on by means of tree-based planning algorithms by Kocsis Szepesvari (2006), which have proven very successful in the game of Go [1]. Regardless of the time setting, perhaps the most well known result from this problem is due to Gittins [4]. According to Gittins, the index of an arm (e.g., slot machine arm) can be defined as the maximal attainable payoff rate of the arm. He then showed that in such a sequential problem, it is optimal to choose the arm with highest index. However, the optimality of the index policy does not carry over to the problem where switching arms involves a cost.

Banks and Sundaram (1994) showed that no index policy is indeed optimal when switching arms incurs costs [1]. Moreover, none of the previous formulations of the multiarmed bandit problem, presented at the time of the paper under study, had addressed the issue of switching cost. Nonetheless, for the switching cost version, in 1996, Asawa and Teneketzis showed that if it is optimal to play an arm at a certain time, then it is optimal to continue with it until its stopping time when the appropriate index is achieved, where the index is a modified version of the Gittins index [1]. Van Oyen and Pichitlamken later drew upon the results from Asawa and Teneketsiz and incorporated setup times for switching arms, yet established a qualitatively identical result [1] Numerous other results were proposed in very restricted environments of the problem. Dusonchet and Hongler explicity derived optimal switching thresholds, but this was for an explicit two-armed bandit [1]. Benkherouf studied the special case where switching cost is paid only at the first the arm is played and showed that the Gittins index is actually still optimal [1]. Most recently, following

in theme of asymptotical limits developed by Robbins, Anathram, and the paper under study by Agrawal et al, Ofer proved that a player's N round minimax regret can be made to be $\Theta(N^{\frac{2}{3}})$ [5] In what follows, the problem definition, main results, and critique will be given for the paper under study *Asymptotically Efficient Adaptive Allocation Rules for the Multiarmed Bandit Problem with Switching cost* [6].

2 Problem Formulation

Let $\Pi_1, ..., \Pi_p$ be statistical populations specified respectively by uni-variate density functions $f(x;\theta_1), ..., f(x;\theta_p)$ with respect to some measure v, where $f(.;.))$ is known and the θ_j are unknown parameters in some set Θ. Assume that $\int_{-\infty}^{\infty} |x| f(x;\theta) dv(x) < \infty$ for all $\theta \in \Theta$. Let:

$$\mu(\theta) := \int_{-\infty}^{\infty} |x| f(x;\theta) dv(x) \tag{2.0.1}$$

and define:

$$\mu^* := max\{\mu(\theta_1), ..., \mu(\theta_p)\} = \mu(\theta^*) \tag{2.0.2}$$

An adaptive allocation rule ϕ is a sequence of random variables $\phi_1, \phi_2, ...$ taking values in the set $\{1, ..., p\}$ such that the event $\{\phi_n = j\}$ ("sample from Π_j at stage n") belongs to the σ-field F_{n-1} generated by $\phi_1, x_1, ..., \phi_{n-1}, x_{n-1}$. For $n \leq \infty$ let:

$$T_n(j) = \sum_{i=1}^{n} 1\{\phi_i = j\} \tag{2.0.3}$$

denote the number of times that the rule ϕ samples from $\Pi_j$through stage n, and let

$$J_n := \sum_{i=1}^{n} x_i \tag{2.0.4}$$

Then by Wald's lemma:

$$E * J_n := \sum_{j=1}^{p} \mu(\theta_j) E T_n(j) \tag{2.0.5}$$

Define the sampling regret

$$R'_n(\theta) := n\mu^* - EJ_n = \sum_{j:\mu(\theta_j)<\mu^*} (\mu^* - \mu(\theta_j))(ET_n(j)) \tag{2.0.6}$$

Also, let

$$S_n(j) := \sum_{i=2}^{n} 1\{\phi_i = j, \phi_{i-1} \neq j\} \tag{2.0.7}$$

and define the switching regret

$$SW_n(\theta) := C \sum_{j=1}^{p} ES_n(j) \tag{2.0.8}$$

where $C > 0$ is the fixed switching cost. And finally define the total regret as:

$$R_n(\theta) := R'_n(\theta) + SW_n(\theta) \tag{2.0.9}$$

Thus, the problem is in maximixing $(EJ_n - SW_n)$ which is equivalent to minimizing the total regret $R_n(\theta)$. Note that a rule is called uniformly good if for every θ, and $a > 0$

$$R_n(\theta) = o(n^a) \tag{2.0.10}$$

Such rules do not allow the total regret to increase very rapidly. The problem formulation is now complete. Before the main results of the paper, some set up is necessary. The idea is that since the introduction of a switching cost discourages frequent switching, one needs to sample in blocks. The block allocation scheme consists of two steps. First, determine *a priori* intervals of time, and over each interval sample from the same population. At the beginning of each interval, adaptively decide which population to sample from. The intervals are chosen so that if one ensure the expected number of samples from each inferior population is $O(log(n))$, the expected number of switches is automatically controlled to $o(log(n))$. First, time is divided into "frames" $0, 1, 2, ...$ Each time frame f is further divided into "blocks" numbered $1, 2, 3, ...$ All the blocks in a frame are equal length. Then if N_f denotes the time instant at the of frame f, b_f denotes block length of each block in frame f, and k_f denotes the number of blocks in frame f, the block lengths and frame lengths $(N_f - N_{f-1})$ as are chosen as:

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

Figure 1: Block Allocation Scheme [6]

Where the lower and upper bounds are computed in detail by Agrawal [6]. Second, in starting up the allocation scheme, in frame 0, the player samples once from each population. From then on, the beginning of each block is a comparison instant; at such a time, the player decided which population to sample from and samples from this population for the entire block, i.e., b_f times. To decide which population this corresponds to, each each comparison instant n, one employs upper confidence bounds $U_n(j)$ for $\mu(\theta_j)$, the mean of each population Π_j. The leader is defined as the population Π_{j_n} for which:

$$U_n(j_n) \geq U_n(j), j \in \{1, ..., p\} \tag{2.0.11}$$

At each comparison instant, the player samples from the leader. The actual construction of the upper confidence bounds is briefly discussed below, although the reader is referred to the works of Robbins for a more thorough discussion. Let $Y_1, Y_2, ...$ be i.i.d random variables with density function $f(y; \theta)$ with respect to some measure v, where θ is defined as before. Let g_{ni} map $\mathbb{R}^i$ to $\mathbb{R}$ with $n = 1, 2, ...; i = 1, 2, ..., n$ be Borel functions such that for all θ:

$$P_\theta\{g_{ni}(Y_1,...,Y_i) \geq \mu(\theta), i \leq n\} = 1 - o(n^{-1}) \tag{2.0.12}$$

and,

$$\lim_{n\to\infty}[E_\theta[sup\{1 \leq i \leq n | g_{ni}(Y_1,...,Y_n) \geq \mu(\lambda)\}]/\log(n)] \leq \frac{1}{I(\theta,\lambda)} \tag{2.0.13}$$

whenever $\mu(\lambda) > \mu(\theta)$,where $I(.;.)$ is the Kullback-Leibler number. Moreover, g_{ni} is non-decreasing in $n \geq i$. Then at each comparison instant, the upper confidence bound $U_n(j)$ for $\mu(\theta_j)$, the mean of population Π_j is given by:

$$U_n(j) = g_{nTn(j)}(Y_{j1},...,Y_{jTn(j)}) \tag{2.0.14}$$

3 Results

This paper shows that the achievable minimum regret under with the addition of a switching cost is the same as that when there is no switching cost. This is shown in Equation 3.0.1.

$$R_n(\theta) = \big[\sum_{j:\mu(\theta_j)\mu*} \frac{(\mu* - \mu(\theta_j))}{I(\theta_j,\theta*)}\big] log(n) \tag{3.0.1}$$

This result is displayed in Figure 2

Theorem 4.1: Under the block allocation rule ϕ^*, for all θ and every j such that $\mu(\theta_j) < \mu^*$

i) $E_\theta T_n(j) \leq \left(\frac{1}{I(\theta_j, \theta^*)} + o(1)\right) \log n,$ (4.8)

ii) $E_\theta S_n(j) \leq o(\log n),$ (4.9)

and consequently, for all θ satisfying Assumption A4

iii) $\limsup_{n\to\infty} R_n(\theta)/\log n \leq \sum_{j:\mu(\theta_j)<\mu^*} (\mu^* - \mu(\theta_j))/I(\theta_j, \theta^*).$ (4.10)

Figure 2: Theorem 4.1 [6]

The proof of this theorem relies on the setup described in 1.

3.1 Proof of i)

The intuition behind this portion of the theorem (Equation 3.1.1) is that the expected amount of times we choose from sub-optimal distribution grows much slower than $log(n)$.

$$E_\theta[T_n] \leq \left(\frac{1}{I(\theta_j,\theta*)} + o(1)\right) log(n) \tag{3.1.1}$$

The sketch of this proof is as follows:

1. First consider $n = N_f$ (time instances at the end of frame f).
   (a) For any j such that $j \neq j*$

$$T_{N_l(j)} = \sum_{f=0}^{l} b_f * |\{i \mid \phi \text{ samples from } \Pi_j\}| \tag{3.1.2}$$

   (b) Equation 3.1.2 is then upper bounded by splitting the sum into two scenarios:
      i. The upper confidence bound of the sub-optimal distribution j is greater than the true mean of the optimal distribution j^*, and the sub-optimal distribution is chosen.

$$Term\ 1 = \sum_{f=1}^{l} b_f * |\{i \mid U_n(j) \geq \mu(\theta*), \text{ and } \phi* \text{ samples from } \Pi_j \text{ at } n(f,i)\}| \tag{3.1.3}$$

      ii. The upper confidence bound of the optimal distribution is less than the true mean of the optimal distribution j^* (this scenario is not dependent on which distribution is chosen.

$$Term\ 2 = \sum_{f=1}^{l} b_f * |\{i \mid U_n(j*) \leq \mu(\theta*)\}| \tag{3.1.4}$$

   This upper bound is justified because there will be overlap between these two scenarios, therefore leading to doubly counting some particular scenarios.
   (c) The two scenarios are represented by their own summation over all frames. The proof then proposes two claims, which will provide the sufficient condition for this proof.
      i. Claim 1:

$$\limsup \left(\frac{E_\theta[Term\ 1]}{log(N_l)} \right) \leq \frac{1}{I(\theta_j, \theta*)} \tag{3.1.5}$$

      Proof:
      A. This term is upper bounded by 3.1.6.

$$Term\ 1 \leq \sup\{i | g_{N_l i}(Y_1, \ ..., \ Y_i) \geq \mu(\theta*)\} + b_f \tag{3.1.6}$$

         This upper bound relies on the non-decreasing properties of $g()$ (Section 2).
      B. Dividing the expectation of 3.1.3 by $log(N_l)$, and using the scheme in Figure 1 and the upper bounds on N_l, we can use 2.0.14 to achieve the desired result for Claim 1.
      ii. Claim 2:

$$E_\theta[Term\ 2] \leq o(1) * log(N_l) \tag{3.1.7}$$

      Proof:
      A. This claim uses 2.0.11 to upper bound the sum.
      B. the summations are then brought within the $o()$ function by applying a proof that is included in their appendix.

2. This notion is extended to any general n by considering $N_{l-1} \leq n \leq N_l$.
   Using the upper bounds on N_l and N_{l-1}, it is shown that the result also holds for all n.

3.2 Proof of ii)

1. Initially, the expected total cost of switching to a sub-optimal distribution j is upper bounded by assuming that, within each frame, we are never staying at that distribution for two blocks in a row, we are always switching to it from a different distribution.

2. This sum is then manipulated with arithmetic, and they use 3.1.1 to show 3.2.1.

$$E_\theta[(S_n(j))] \leq K(\epsilon) * 2 * l \ + \ M(\epsilon) \tag{3.2.1}$$

   Where :

$$M(\epsilon) \ = \ \sum_{f \ = \ 1}^{f_0 - 1} \frac{E_\theta T_{N_f}(j)}{f^2} \tag{3.2.2}$$

3. When the expected switching cost 3.2.1 is divided by $log(n)$, the desired result falls into place, proving ii)

3.3 Proof of iii)

The proof of iii) is achieved by directly using the definition of R_n as well as i) and ii). It is therefore omitted in the paper.

3.4 Discussion

The final results of this paper are novel and remarkable. By using an intelligent block allocation scheme, and action policy, the added cost function is rendered irrelevant. This is because the theoretical approachable limit under the imposed switching cost, is in fact the same as without that cost. The result given in 3.4.1 is entirely based on the idea of adaptive decision making, given new information of the previous actions.

$$R_n(\theta) = \left[\sum_{j:\mu(\theta_j)\mu*} \frac{(\mu * - \mu(\theta_j))}{I(\theta_j, \theta*)} \right] log(n) \tag{3.4.1}$$

4 Review

The problem of finding an asymptotically optimal solutions to the multi-armed bandit problem is quite an interesting one. This is because solutions of this kind offer a nearly optimal solution to a problem that has an exact optimal solution that is NP-hard to find [1]. Asymptotically optimal solutions are not the only possible ways to solve this problem, though. Jun offers 2 other potential methods for solving the problem [1]. First, by characterization of the optimal policy, whereby an index is defined for each bandit in a similar way to the bandit problem without switching cost. Second, by an exact solution in a restricted environment, whereby imposing certain constraints on the problem, an exact solution can be obtained. Each of these methods have various benefits and trade offs. The asymptotically optimal solution is appealing because taking such an approach does not require destroying the general structure of the problem, as is done to find an exact solution. However in doing so, the finite-sample performance is often not as good as its asymptotic performance [1]. This may well be the case in many applications, such as job search and labor

mobility.

In one example of this application, McCall and McCall explained migration behavior combined with job search among a set of cities. Workers can observe the existing wage and match them only by moving to the city [7]. Here the cost of switching is the cost of moving cities, and the unknown payoffs are the wages. In trying to implement this model for yourself, or for a population of people, it would likely be extremely difficult to collect enough data to sufficiently reach the asymptotically optimal limit. This makes sense because the average number of times a person moves in their lifetime is approximately 11 times [8], however the number of potential cities one could move to in Canada alone is 233 [9]. These applications limit the potential usefulness of this paper. There are, however, other applications where the asymptotically optimal model provides a good solution for optimality. In the classic bandit problem of a set of slot machines, the approach given in this paper would work well. This is because the slots can be tested an arbitrary amount of times, and knowledge of all of the previous payouts is retained. In this way, as time goes on, the adaptive allocation rule can be improved and will approach the asymptotically optimal solution.

A possible generalization of this paper is to make it robust to handling the case when multiple populations are sampled from at a time. This directly applies to the slot machine example again. A person playing multiple slot machines is likely to play more than one at a time, an issue that this paper does not address. The paper also does not directly address the case of variable switching costs, which is very important for application like job search. Different cities will naturally have different switching costs with respect to a number of factors. It is also noted that in defining a distance between distributions, the Kullback-Leibler Divergence is used. However, according to Cover and Thomas, this is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality [10]. Further research may improve the solution by picking a more generic or true distance metric between distributions.

References

[1] T. Jun, "A survey on the bandit problem with switching costs," 2004.

[2] H. R. T. Lai, "Asymptotically efficient adaptive allocation rules," 1985.

[3] P. J. Walrand, V. Ananthram, "Asymptotically efficient allocation rules for multi-armed bandit problem with multiple plays," 1987.

[4] J. Gittins, "Bandit processes and dynamic allocation indices," 1979.

[5] T. K. Y. P. O. Dekel, J. Ding, "Bandits with switching costs," 2013.

[6] D. T. R. Agrawak, M. Hedge, "Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost," 1988.

[7] B. McCall and J. McCall, "A sequential study of migration and job search," 1987.

[8] M. Chalabi, "How many times does the average person move?" 2015. [Online]. Available: https://fivethirtyeight.com/features/how-many-times-the-average-person-moves/

[9] W. P. Review, "Population of cities in canada (2018)," 2018. [Online]. Available: http://worldpopulationreview.com/countries/canada-population/cities/

[10] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley and Sons, Inc, 2006.