Stochastics and Stochastic Reports, Vol. 29, pp. 437–459
Reprints available directly from the publisher
Photocopying permitted by license only

MULTI-ARMED BANDIT PROBLEMS WITH MULTIPLE PLAYS AND SWITCHING COST

R. AGRAWAL

Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53706-1691, USA

M. HEGDE

Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803, USA

D. TENEKETZIS

Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122, USA

(*Received 10 March 1989; in final form 19 October 1989*)

We consider multi-armed bandit problems with switching cost and multiple plays, define "uniformly good" allocation rules and restrict attention to such rules. We consider i.i.d. as well as Markovian rewards. We present a lower bound on the asymptotic performance of "uniformly good" allocation rules and construct an allocation scheme that achieves the bound. We discover that despite the inclusion of a switching cost the proposed allocation scheme achieves the same asymptotic performance as the optimal rules for the bandit problem without switching cost. This is made possible by grouping together the samples in a fashion that makes the rate of switching negligible compared to the rate of experimentation.

KEY WORDS: Multi-armed bandit, asymptotically efficient, switching cost, regret, block allocation.

1. INTRODUCTION

In this paper we consider a version of the multi-armed bandit problem with multiple plays and with switching cost. There are $p \geqq 2$ arms and at each stage we have to play a fixed number, m, of these arms, and collect rewards from each of them. Successive plays of an arm $j, j=1,\ldots,p$, yield rewards $X_{j1}, X_{j2},\ldots$ whose distribution (either i.i.d. or Markovian) is parametrized by an unknown parameter θ_j belonging to a known parameter space Θ. Moreover, each time we decide to play a different set of m arms we incur a switching cost proportional to the number of arms that are different from the previous play. The problem then is how to select these m arms at each stage so as to maximize, in some sense, the long run sum of rewards minus the switching cost.

The set up of the problem in this paper is similar to the one addressed by Lai and Robbins [1,2]. Anantharam *et al.* [3] have considered a version of this problem with multiple plays but without switching cost. Agrawal *et al.* [4] have provided a solution to the multi-armed bandit problem with switching cost but with single plays. In this paper we consider the more general set up where we have both multiple plays and a switching cost. Even though the adaptive allocation scheme we propose combines the essential ideas behind the strategies in [3,4], the calculation of the performance of the proposed adaptive scheme (Theorem 4.1) is highly non-trivial and does not follow easily from the technical results in [3,4].

The paper is organized as follows: In Section 2 we provide a precise problem formulation. We consider two cases for the distribution of rewards obtained from the successive plays of any arm: Case A–i.i.d., and Case B–Markovian. In Section 3 we present a lower bound on the total regret, and in Section 4 we construct allocation rules which achieve this lower bound.

2. PROBLEM FORMULATION

Let there be p arms. Successive plays of arm $j, j=1,2,\ldots,p$ yields rewards X_{j1}, $X_{j2},\ldots$ whose joint distribution is parametrized by an unknown parameter θ_j belonging to a known parameter space Θ. An adaptive allocation rule ϕ consists of a sequence of $\{0,1\}^p$ valued random vectors $\{\phi_n\}_{n=1}^{\infty}$, indicating which m, $(1 \leqq m < p)$, of p arms have been selected for play at stage n on the basis of all the past actions and past observations, i.e. ϕ_n is a function of only the past actions $\phi_1,\ldots,\phi_{n-1}$ and the past rewards $X_{j1},\ldots,X_{jT_{n-1}(j)}$, $j=1,\ldots,p$, where

$$T_n(j) := \sum_{i=1}^{n} \phi_i(j) \tag{2.1}$$

is the number of times arm j was used up to stage n. Let $\mathscr{F}_n(j)$ denote the σ-algebra generated by $X_{j1},\ldots,X_{jn}$; let $\mathscr{F}_\infty(j)=\bigvee_n \mathscr{F}_n(j)$ and $G(j)=\bigvee_{i\neq j}\mathscr{F}_\infty(i)$. For an adaptive allocation rule ϕ, the number of plays we have made of arm j by time n, $T_n(j)$, is a stopping time with respect to $\{\mathscr{F}_t(j) \vee G(j), t \geqq 1\}$. Let

$$J_n := \sum_{j=1}^{p} \sum_{i=1}^{T_n(j)} X_{ji} \tag{2.2}$$

be the sum of rewards collected up to stage n. Also let

$$S_n := \sum_{i=1}^{n-1} d(\phi_i, \phi_{i+1}) \tag{2.3}$$

be the total number of switches up to stage n where

$$d(\phi_i, \phi_{i+1}) := 1/2 \sum_{j=1}^{p} 1\{\phi_i(j) \neq \phi_{i+1}(j)\}$$

is the number of switches at stage i. Our objective (in choosing an adaptive allocation scheme ϕ) is to maximize in some sense $E_\theta(J_n - CS_n)$ where $C>0$ is a fixed switching cost, and $\theta := (\theta_1, \ldots, \theta_p)$ is the parameter configuration. We shall make this notion of optimality more precise shortly. Before doing so we would like to describe, in detail, the two cases for the distribution of the rewards obtained from each arm, that we shall address in this paper.

CASE A: (I.I.D. rewards)

The successive plays of each arm $j, j=1,2,\ldots,p$ yield i.i.d. rewards $X_{j1}, X_{j2}, \ldots$ with a common marginal density $f(x;\theta_j)$ with respect to some measure ν, where $f(\cdot;\cdot)$ is known and the θ_j's are unknown parameters belonging to some set Θ. Assume that

$$\int_{-\infty}^{\infty} |x| f(x;\theta)\, d\nu(x) < \infty \quad \text{for all} \quad \theta \in \Theta.$$

Define

$$\mu(\theta) := \int_{-\infty}^{\infty} x f(x;\theta)\, d\nu(x), \tag{2.4}$$

the mean reward under the parameter θ, and

$$I(\theta, \lambda) := \int_{-\infty}^{\infty} [\log(f(x;\theta)/f(x;\lambda))] f(x;\theta)\, d\nu(x), \tag{2.5}$$

the Kulback–Leibler number, a well known measure of distance between distributions.

CASE B: (Markovian rewards)

The successive plays of each arm $j, j=1,\ldots,p$ yield Markovian rewards $X_{j1}, X_{j2}, \ldots$ with a stationary transition probability

$$P(\theta_j) := \{P(x,y;\theta_j) \colon x, y \in \mathscr{X}\}, \tag{2.6}$$

and initial probability distribution

$$p(\theta_j) := \{p(x;\theta_j) \colon x \in \mathscr{X}\}, \tag{2.7}$$

where $\mathscr{X} \subset \mathbb{R}$ is a finite set of rewards and θ_j's are unknown parameters belonging to some set Θ. Assume that for

$$x, y \in \mathscr{X}, \theta, \theta' \in \Theta, P(x, y; \theta) > 0 \Rightarrow P(x, y; \theta') > 0, \tag{2.8}$$

$P(\theta)$ is irreducible and aperiodic for all $\theta \in \Theta$ and

$$p(x, \theta) > 0 \quad \text{for all} \quad x \in \mathscr{X} \text{ and } \theta \in \Theta. \tag{2.9}$$

Let $\pi(\theta) := \{\pi(x; \theta) : x \in \mathscr{X}\}$ be the invariant probability distribution under the parameter θ, and let

$$\mu(\theta) := \sum_{x \in \mathscr{X}} x\pi(x; \theta)$$

be the mean reward under this distribution $\pi(\theta)$. Define the Kulback–Leibler number

$$I(\theta, \lambda) = \sum_{x \in \mathscr{X}} \pi(x, \theta) \sum_{y \in \mathscr{X}} P(x, y; \theta) \log \frac{P(x, y; \theta)}{P(x, y; \lambda)}. \tag{2.10}$$

Note that, by (2.8)

$$0 < I(\theta, \lambda) < \infty \quad \text{for} \quad \theta \neq \lambda.$$

Note that we used the notation $\mu(\theta)$ and $I(\theta, \lambda)$ to stand for the mean reward and the Kulback–Leibler number respectively under both Case A and Case B. This is an abuse of notation as $\mu(\theta)$ and $I(\theta, \lambda)$ have different definitions in the two cases. For the sake of convenience any unqualified statements using the above terms will be taken to hold for both cases.

Under the above two cases we can make the following observations.

1a) For the i.i.d. case, by Wald's Lemma (cf. [6])

$$E_{\boldsymbol{\theta}} J_n = \sum_{j=1}^{p} \mu(\theta_j) E_{\boldsymbol{\theta}} T_n(j). \tag{2.11a}$$

1b) For the Markovian case, by (2.9) of [3]

$$\left| E_{\boldsymbol{\theta}} J_n - \sum_{j=1}^{p} \mu(\theta_j) E_{\boldsymbol{\theta}} T_n(j) \right| \leqq \text{const. (independent of } n). \tag{2.11b}$$

Let σ be a permutation of $\{1, \ldots, p\}$ such that

$$\mu(\theta_{\sigma(1)}) \geqq \cdots \geqq \mu(\theta_{\sigma(m)}) \geqq \mu(\theta_{\sigma(m+1)}) \geqq \cdots \geqq \mu(\theta_{\sigma(p)})$$

and call $\sigma(1), \ldots, \sigma(m)$ the m-best arms. Clearly, if the parameter configuration $\boldsymbol{\theta}$

were known, then the optimal strategy would be to always use the m-best arms. If this were true then

2a) $$E_\theta J_n = n \sum_{j=1}^{m} \mu(\theta_{\sigma(j)}) \quad \text{for the i.i.d. case.} \tag{2.12a}$$

2b) $$\left| E_\theta J_n - n \sum_{j=1}^{m} \mu(\theta_{\sigma(j)}) \right| \leqq \text{const. (indep. of } n\text{) for the Markovian case.} \tag{2.12b}$$

In the absence of the knowledge of θ it is desirable to approach this performance as closely as possible. For this purpose we define

a) the "sampling regret"

$$R'_n(\theta) := n \sum_{j=1}^{m} \mu(\theta_{\sigma(j)}) - E_\theta J_n, \tag{2.13}$$

b) the "switching regret"

$$SW_n(\theta) := CE_\theta S_n, \tag{2.14}$$

and

c) the "total regret"

$$R_n(\theta) := R'_n(\theta) + SW_n(\theta). \tag{2.15}$$

Maximizing $E_\theta(J_n - CS_n)$ is thus equivalent to minimizing the "total regret" $R_n(\theta)$. More precisely we want to minimize the rate at which $R_n(\theta)$ increases with n (linear, logarithmic, finite etc.). Note that it is impossible to do this uniformly over all parameter configurations θ. We call a rule "uniformly good" if for every parameter configuration θ

$$R_n(\theta) = o(n^\alpha) \text{ for every } \alpha > 0. \tag{2.16}$$

Such rules do not allow the total regret to increase very rapidly for any θ. We restrict our attention to the class of uniformly good schemes, and consider any others uninteresting.

The main results of the paper, appearing in Section 3 and 4, are derived under the following technical assumptions:

A1 $0 < I(\theta, \lambda) < \infty$ whenever $\mu(\lambda) > \mu(\theta)$.

A2 For every $\varepsilon > 0$ and $\theta, \lambda \in \Theta$ such that $\mu(\lambda) \geqq \mu(\theta)$, there exists $\delta > 0$, such that

$$|I(\theta, \lambda) - I(\theta, \lambda')| < \varepsilon \text{ if } |\mu(\lambda) - \mu(\lambda')| < \delta.$$

A3 $\forall\lambda\in\Theta$ and $\forall\delta>0, \exists\lambda'\in\Theta$ such that $\mu(\lambda)<\mu(\lambda')<\mu(\lambda)+\delta$.

A4 The parameter configuration θ is such that $\mu(\theta_{\sigma(m)})>\mu(\theta_{\sigma(m+1)})$.

Assumption A1 is trivially satisfied for Case B. For Case A the assumption $I(\theta,\lambda)>0$ is automatically satisfied whenever $\mu(\lambda)>\mu(\theta)$. The condition $I(\theta,\lambda)<\infty$ implies that the distribution of the samples under the parameter θ is absolutely continuous with respect to the distribution of the samples under any parameter λ such that $\mu(\lambda)>\mu(\theta)$. Such a condition can be expexted to be satisfied for most parametric families of distributions which are mutually absolutely continuous. Assumption A2 is a continuity condition on $I(\theta,\lambda)$ for fixed θ and $\mu(\lambda)\geqq\mu(\theta)$. Assumption A3 is a denseness condition on the space Θ. Assumption A2–A3 are needed to obtain the lower bound on the total regret. Assumption A4 implies that there is a unique set of m-best arms amongst all of the p arms. This assumption is essential in obtaining the upper bound on the total regret.

3. A LOWER BOUND FOR THE TOTAL REGRET

In this section we note the extension of the lower bound obtained by Anantharam *et al.* [3] to our problem. We state this in the form of Theorem 3.1.

THEOREM 3.1 *Assume that A1–A3 hold. Let ϕ be any uniformly good allocation rule, i.e. ϕ satisfies* (2.16). *Then for any inferior arm j, i.e., $\mu(\theta_j)<\mu(\theta_{\sigma(m)})$,*

$$\liminf_{n\to\infty} E_\theta T_n(j)/\log n \geqq 1/I(\theta_j,\theta_{\sigma(m)}), \tag{3.1}$$

and consequently

$$\liminf_{n\to\infty} R_n(\theta)/\log n \geqq \sum_{j\in\{\sigma(m+1),\ldots,\sigma(p)\}} \frac{(\mu(\theta_{\sigma(m)})-\mu(\theta_j))}{I(\theta_j,\theta_{\sigma(m)})}. \tag{3.2}$$

Proof Follows from Theorem 3.1 of Anantharam *et al.* [3]. □

We shall call rules that attain the above lower bound asymptotically efficient, i.e.,

$$R_n(\theta)\sim\left(\sum_{j\in\{\sigma(m+1),\ldots,\sigma(p)\}} \frac{(\mu(\theta_{\sigma(m)})-\mu(\theta_j))}{I(\theta_j,\theta_{\sigma(m)})}\right)\log n. \tag{3.3}$$

4. CONSTRUCTION OF ASYMPTOTICALLY EFFICIENT RULES

In the first part of Section 4.1 we motivate the idea of "block allocation" and then introduce a specific "block allocation scheme". In this scheme we employ upper confidence bounds and point estimates that are constructed in Section 4.2. Finally, in Section 4.3 we derive an upper bound on the total regret of our allocation scheme. This bound is asymptotically equal to the lower bound of Theorem 3.1. Consequently, the proposed allocation scheme is asymptotically efficient.

4.1 Block Allocation Scheme

In view of Theorem 3.1, if ϕ is an asymptotically efficient rule, then the number of times that ϕ plays any inferior arm j up to stage n is about $(\log n)/I(\theta_j, \theta_{\sigma(m)})$. With no knowledge of the time instants at which the plays are made from the inferior arms, all we can infer about the contribution from arm j to the switching regret up to stage n is that it is at most about $(2\log n)/I(\theta_j, \theta_{\sigma(m)})$. (The largest contribution to the switching cost occurs when every play of arm j involves switching to and from it.) Clearly any asymptotically efficient rule must ensure that the plays of any arm are grouped together in blocks in such a fashion that the contribution to the switching cost is much smaller than the above upper bound, in fact $o(\log n)$. Furthermore, the block lengths must increase with n.

With this idea in mind we construct a "block allocation scheme" in two steps: We first determine, *a priori*, intervals of time, and over each interval we use the same arms. Then, at the beginning of each interval we *adaptively* decide which arms to use. The intervals are chosen so that if we ensure the expected numbers of plays of each inferior arm is $O(\log n)$, the expected number of switches is automatically controlled to $o(\log n)$.

Step 1 To facilitate analysis time is first divided into "frames" numbered 0,1,2,.... Each frame f is further subdivided into "blocks" numbered 1,2,3,.... All the blocks in a frame are of equal length. Each such block can thus be uniquely identified by (f, i) where f is the frame number to which it belongs, and i is the block number.

Furthermore, let

N_f denote the time instant at the end of frame f,

b_f denote the block length of each block in frame f,

k_f denote the number of blocks in frame f.

We choose the block lengths and frame lengths $(N_f - N_{f-1})$ as follows:

$$\begin{array}{ccc} \textit{Frame}\ \#(f) & b_f & N_f - N_{f-1} \\ 0 & 1 & p \\ f \geqq 1 & f & \left[\dfrac{2^{f^2} - 2^{(f-1)^2}}{f}\right] p.f. \end{array} \tag{4.1}$$

We use $[x]$ to denote the smallest integer $\geqq x$, and $[x]$ to denote the greatest integer $\leqq x$. Such a block allocation scheme is illustrated in Figure 1 for $p=4$.

We now derive lower and upper bounds for N_f, the time instants at the end of frame f. These bounds will be used extensively in subsequent analysis.

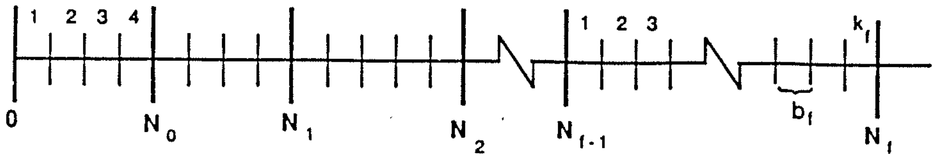

Figure 1 "Block allocation scheme" for $p=4$.

$$N_f=\sum_{i=1}^{f}\left[\frac{2^{i^2}-2^{(i-1)^2}}{i}\right]p.i.+p$$

$$\geqq\sum_{i=1}^{f}\left(\frac{2^{i^2}-2^{(i-1)^2}}{i}\right)p.i.+p$$

$$=p2^{f^2}. \tag{4.2}$$

$$N_f\leqq\sum_{i=1}^{f}\left(\frac{2^{i^2}-2^{(i-1)^2}}{i}+1\right)p.i.+p$$

$$=p2^{f^2}+p\sum_{i=1}^{f}i$$

$$\leqq p2^{f^2}+pf^2. \tag{4.3}$$

Therefore, $N_f\sim p2^{f^2}$, and $\log N_f\sim f^2$. On the other hand, the block lengths $b_f=f$.

Step 2 To start up the allocation scheme, in frame 0 use each arm m-times. From then on proceed as follows: The beginning of each block (f,i) is a "comparison instant", $n(f,i)(=N_{f-1}+(i-1)b_f)$; at that time decide when arms to play. Play the chosen arms for the entire block (f,i), i.e. b_f times. To decide which arms to use at each comparison instant n we employ point estimates $\hat{\mu}_n(j)$, and upper confidence bounds $U_n(j)$ for $\mu(\theta_j)$, the mean reward under each arm j. These are constructed in Section 4.2.

Choose $0<\delta<1/p$. Call the arms which have been used at least δn times (up to stage n) "well sampled". Note that at least m of the p arms have to be well sampled. Among these well sampled arms choose the "m-leaders" at comparison instant $n(f,i)$, namely the arms with m best values of the point estimate $\hat{\mu}_n(j)$, $j=1,\ldots,p$.

Consider for experimentation each of the p arms in a round-robin fashion. For this purpose associate each block i with arm j such that $j\equiv i \bmod p$, and compute the upper confidence bound $U_n(j)$ of this arm j at comparison instant $n(f,i)$. Now use the following rules to decide which arms to play at the comparison instant $n(f,i)$.

a) If arm j is already one of the m-leaders then at comparison instant $n(f,i)$ choose to play the m-leaders.

b) If arm j is not among the m-leaders and $U_n(j)$ is less than $\hat{\mu}_n(k)$ for every m-leader k, then again play the m-leaders.

c) If arm j is not among the m-leaders, and $U_n(j)$ exceeds or equals $\hat{\mu}_n(k)$ of the least best of the m-leaders, then play the $m-1$ best of m-leaders and arm j.

Note that in any case the $m-1$ best of the m-leaders always get played.

Remark The proposed allocation strategy is obtained by imposing the "block structure" (Step 1) of [4] on the adaptive allocation scheme of [3].

4.2 Construction of Upper Confidence Bounds and Point Estimates

To fix ideas, let $Y_1, Y_2, \ldots,$ be a sequence of random variables (under either Case A (i.i.d.) or Case B (Markovian)) whose distribution is parametrized by an unknown parameter θ belonging to a known parameter space Θ. Let

$$g_{ni}\colon \mathbb{R}^i \to \mathbb{R}(n=1,2,\ldots; i=1,2,\ldots,n)$$

be Borel functions such that for every $\theta \in \Theta$

$$P_\theta\{g_{ni}(Y_1,\ldots,Y_i) \geqq \mu(\theta) \quad \text{for all} \quad i \leqq n\} = 1 - o(n^{-1}), \tag{4.4}$$

$$\limsup_{n\to\infty}\,[E_\theta[\sup\{1 \leqq i \leqq n | g_{ni}(Y_1,\ldots,Y_i) \geqq \mu(\lambda)\}]/\log n] \leqq \frac{1}{I(\theta,\lambda)} \tag{4.5}$$

whenever $\mu(\lambda) > \mu(\theta)$, and

$$g_{ni} \text{ is nondecreasing in } n \geqq i \text{ for every fixed } i=1,2,\ldots. \tag{4.6}$$

Let $h_i\colon \mathbb{R}^i \to \mathbb{R}$ be Borel functions such that for every $\theta \in \Theta$

$$P_\theta\left\{\max_{\delta n \leqq i \leqq n} |h_i(Y_1,\ldots,Y_i) - \mu(\theta)| > \varepsilon\right\} = o(n^{-1}). \tag{4.7}$$

We now make use of the functions g_{ni} and h_i to define our upper confidence bounds and point estimates respectively. Let $Y_{j1},\ldots,Y_{jT_n(j)}$ be the successive rewards obtained from arm j up to stage n. Then at each comparison instant $n(f,i)$ the upper confidence bound $U_n(j)$ and the point estimate $\hat{\mu}_n(j)$ for $\mu(\theta_j)$, the mean reward under arm j, are given by

$$\begin{aligned} U_n(j) &= g_{nT_n(j)}(Y_{j1},\ldots,Y_{jT_n(j)}), \\ \hat{\mu}_n(j) &= h_{T_n(j)}(Y_{j1},\ldots,Y_{jT_n(j)}) \end{aligned} \tag{4.8}$$

for each $j \in \{1,\ldots,p\}$. Denote by ϕ^* the allocation rule constructed in Section 4.1 and 4.2.

For a heuristic explanation of the upper confidence bounds and point estimates constructed above, as well as for their explicit form under a special class of distributions, see [1,3,4,5,7].

4.3 Upper Bound on the Total Regret

THEOREM 4.1 *Assume that the arms have been reindexed and $l \geqq 0$ so that*

$$\mu(\theta_1) \geqq \cdots \geqq \mu(\theta_l) > \mu(\theta_{l+1}) = \cdots = \mu(\theta_m) > \cdots \geqq \mu(\theta_p).$$

Under the block allocation rule ϕ^*, for all *θ satisfying A4.*

$$E_\theta T_n(j) \leqq \left(\frac{1}{I(\theta_j,\theta_m)} + o(1)\right)\log n \quad \textit{for every } j > m, \tag{4.9}$$

$$E_\theta(n - T_n(j)) = o(\log n) \quad \textit{for every } j \leqq l, \tag{4.10}$$

$$SW_n(\theta) \leqq o(\log n), \tag{4.11}$$

and consequently,

$$\limsup_{n\to\infty} R_n(\theta)/\log n \leqq \sum_{j=m+1}^{p} (\mu(\theta_m) - \mu(\theta_j))/I(\theta_j,\theta_m). \tag{4.12}$$

Proof Throughout the proof let $\#\, A$ denote the number of elements of a set A. Also fix $\varepsilon > 0$ satisfying

$$\varepsilon < \frac{\mu(\theta_l) - \mu(\theta_m)}{2} \quad \text{if } l > 0$$

and

$$\varepsilon < \frac{\mu(\theta_m) - \mu(\theta_{m+1})}{2}.$$

We now proceed to prove each part of Theorem 4.1.

Proof of (4.9) We shall first prove (i) for $n = N_l$, the end of frame l, i.e. we shall show that

$$E_\theta T_{N_l}(j) \leqq \left(\frac{1}{I(\theta_j,\theta_m)} + o(1)\right)\log N_l.$$

For any fixed $j>m$, we have

$$T_{N_l(j)}=\sum_{f=0}^{l} b_f[\#\{1\leqq i\leqq k_f:\phi^* \text{ selects arm } j \text{ at comparison instant } n(f,i)\}]$$

$$=m+\sum_{f=1}^{l} b_f[\#\{1\leqq i\leqq k_f:\phi^* \text{ selects arm } j \text{ at comparison instant } n(f,i), \text{ all the } m\text{-leaders at comparison instant } n(f,i) \text{ are also the } m\text{-best arms, and for each arm } k \text{ which is an } m\text{-leader, } |\hat{\mu}_n(k)-\mu(\theta_k)|\leqq\varepsilon\}]$$

$$+\sum_{f=1}^{l} b_f[\#\{1\leqq i\leqq k_f:\phi^* \text{ selects arm } j \text{ at comparison instant } n(f,i), \text{ all the } m\text{-leaders at comparison instant } n(f,i) \text{ are also the } m\text{-best arms, and for at least one arm } k \text{ which is an } m\text{-leader } |\hat{\mu}_n(k)-\mu(\theta_k)|>\varepsilon\}]$$

$$+\sum_{f=1}^{l} b_f[\#\{1\leqq i\leqq k_f:\phi^* \text{ selects arm } j \text{ at comparison instant } n(f,i), \text{ and at least one of the } m\text{-leaders is not an } m\text{-best arm at comparison instant } n(f,i)\}]$$

$$\leqq m+\sum_{f=1}^{l} b_f[\#\{1\leqq i\leqq k_f:\phi^* \text{ selects arm } j \text{ at comparison instant } n(f,i), U_n(j)\geqq\mu(\theta_m)-\varepsilon\}]$$

$$+\sum_{k=1}^{m}\sum_{f=1}^{l} b_f[\#\{1\leqq i\leqq k_f:\text{arm } k \text{ is well sampled and } |\hat{\mu}_n(k)-\mu(\theta_k)|>\varepsilon\}]$$

$$+\sum_{f=1}^{l} b_f[\#\{1\leqq i\leqq k_f:\text{at least one of the } m\text{-leaders at comparison instant } n(f,i) \text{ is not an } m\text{-best arm}\}]$$

$$=m+\text{Term } 1(\varepsilon)+\text{Term } 2(\varepsilon)+\text{Term } 3 \text{ (say)}.$$

Claim 1 For any $\rho>0$ there exists $\varepsilon>0$ such that

$$E_\theta[\text{Term } 1(\varepsilon)]\leqq\left(\frac{1}{I(\theta_j,\theta_m)}+\rho+o(1)\right)\log N_l.$$

Claim 2 $E_\theta[\text{Term } 2(\varepsilon)]\leqq o(1)\log N_l$.

Claim 3 $E_\theta[\text{Term } 3]\leqq o(1)\log N_l$.

Proof of Claim 1

$$\begin{aligned}\text{Term } 1(\varepsilon) &= \sum_{f=1}^{l} b_f[\#\{1 \leqq i \leqq k_f : \phi^* \text{ selects arm } j \text{ at the comparison instant } n(f,i),\\ &\qquad U_n(j) \geqq \mu(\theta_m) - \varepsilon\}]\\ &= \sum_{f=1}^{l} b_f[\#\{1 \leqq i \leqq k_f : \phi^* \text{ selects arm } j \text{ at the comparison instant}\\ &\qquad n(f,i), g_{nT_n(j)}(Y_{j1}, \ldots, Y_{jT_n(j)}) \geqq \mu(\theta_m) - \varepsilon\}]\\ &\leqq \sum_{f=1}^{l} b_f[\#\{1 \leqq i \leqq k_f : \phi^* \text{ selects arm } j \text{ at the comparison instant}\\ &\qquad n(f,i), g_{N_l T_n(j)}(Y_{j1}, \ldots, Y_{jT_n(j)}) \geqq \mu(\theta_m) - \varepsilon\}]\\ &\leqq \sup\{1 \leqq i \leqq N_l | g_{N_l i}(Y_{j1}, \ldots, Y_{ji}) \geqq \mu(\theta_m) - \varepsilon\} + b_l.\end{aligned}$$

Then, by (4.1) and (4.2)

$$\frac{E_\theta[\text{Term } 1(\varepsilon)]}{\log N_l} \leqq E_{\theta_j}[\sup\{1 \leqq i \leqq N_l | g_{N_l i}(Y_{j1}, \ldots, Y_{ji}) \geqq \mu(\theta_m) - \varepsilon\}]/\log N_l + l/\log(p2^{l^2}).$$

Hence by (4.5) and A2

$$\limsup_{l \to \infty} \frac{E_\theta[\text{Term } 1(\varepsilon)]}{\log N_l} \leqq \frac{1}{I(\theta_j, \theta_m)} + \rho.$$

Proof of Claim 2

$$\begin{aligned}\text{Term } 2(\varepsilon) &= \sum_{k=1}^{m} \sum_{f=1}^{l} b_f[\#\{1 \leqq i \leqq k_f : \text{arm } k \text{ is well sampled and } |\hat{\mu}_n(k) - \mu(\theta_k)| > \varepsilon\}]\\ &\leqq \sum_{k=1}^{m} \sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} 1\{\max_{\delta n \leqq t \leqq n} |h_t(Y_{k1}, \ldots, Y_{kt}) - \mu(\theta_k)| > \varepsilon\}.\end{aligned}$$

$$\begin{aligned}\therefore E_\theta[\text{Term } 2(\varepsilon)] &\leqq \sum_{k=1}^{m} \sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} P_{\theta_k}\{\max_{\delta n \leqq t \leqq n} |h_t(Y_{k1}, \ldots, Y_{kt}) - \mu(\theta_k)| > \varepsilon\}\\ &= \sum_{i=1}^{m} \sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} o(n^{-1}) \qquad \text{by (4.7)}\\ &= o\left(\sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} \frac{1}{n(f,i)}\right)\\ &= o(\log N_l). \qquad \text{(by appendix } R_1).\end{aligned}$$

Proof of Claim 3

$$\text{Term } 3 = \sum_{f=1}^{t} b_f[\,\#\{1 \leqq i \leqq k_f \text{:at least one of the } m\text{-leaders is not an } m\text{-best arm at comparison instant } n(f,i)\}].$$

The steps of the proof of Claim 3 are summarized in the form of Lemma 4.1 below.

LEMMA 4.1 *Assume all the notation of Theorem 4.1. Let c be a positive integer such that* $c > (1-p\delta)^{-1}$.

For $r = 0, 1, \ldots$ *define*

$$A_r = \bigcap_{1 \leqq j \leqq p} \{ \max_{\delta c^{r-1} \leqq n \leqq c^{r+1}} |h_n(Y_{j1}, \ldots, Y_{jn}) - \mu(\theta_j)| \leqq \varepsilon \}$$

$$B_r = \bigcap_{j \leqq m} \{ g_{ni}(Y_{j1}, \ldots, Y_{ji}) \geqq \mu(\theta_j) - \varepsilon \quad \text{for} \quad 1 \leqq i \leqq \delta n \textit{ and } c^{r-1} \leqq n \leqq c^{r+1} \}$$

where $0 < \delta < 1/p$ *is the same as that used in* ϕ^*.

Then

i) $P_\theta(A_r^c) = o(c^{-r})$, $P_\theta(B_r^c) = o(c^{-r})$.

ii) On $A_r \cap B_r$, *at comparison instant* $n(f,i)$ *where* $k \equiv i \bmod p$ *for some* $k \leqq m$ *(k is an m-best arm) and* $c^{r-1} \leqq n(f,i) \leqq c^{r+1}$ *the rule* ϕ^* *plays k, and if* $r \geqq r_0$ *(sufficiently large) then all the m-best arms are well sampled at any comparison instant* $c^r \leqq n(f,i) \leqq c^{r+1}$.

iii) If $r \geqq r_0$ *(sufficiently large), then on the event* $A_r \cap B_r$, *for every* $c^r \leqq n \leqq c^{r+1}$, *the m-leaders are the m-best arms.*

iv) $E_\theta[\textit{Term 3}] = o(\log N_t)$.

Proof of Lemma 4.1

i) $A_r^c = \bigcup_{1 \leqq j \leqq p} \{ \max_{\delta c^{r-1} \leqq n \leqq c^{r+1}} |h_n(Y_{j1}, \ldots, Y_{jn}) - \mu(\theta_j)| > \varepsilon \}$

$$\therefore P_\theta(A_r^c) \leqq \sum_{j=1}^{p} P_\theta\{ \max_{\delta c^{r-1} \leqq n \leqq c^{r+1}} |h_n(Y_{j1}, \ldots, Y_{jn}) - \mu(\theta_j)| > \varepsilon \}$$

$= po(c^{-(r+1)})$ by (4.7)

$= o(c^{-r})$.

Let q be the smallest positive integer such that $[c^{r-1}/\delta^q] \geqq c^{r+1}$. For $t = 0, \ldots, q$ let $n_t = [c^{r-1}/\delta^t]$ and define

$$D_t = \bigcap_{j \leqq m} \{ g_{n_t i}(Y_{j1}, \ldots, Y_{ji}) \geqq \mu(\theta_j) - \varepsilon \,\forall i \leqq n_t \}.$$

Then by (4.4)

$$P_\theta(D_t^c) = o(n_t^{-1}) = o((c^{r-1})^{-1}) = o(c^{-r}) \quad \text{for} \quad t = 0,\dots,q. \tag{4.13}$$

Given $c^{r-1} \leqq n \leqq c^{r+1}$ and $1 \leqq i \leqq \delta n$, there exists $t \in \{0,\dots,q-1\}$ such that $n_{t+1} \geqq n \geqq n_t \geqq i$ (See appendix R2).

Therefore by (4.6) on the event $\bigcap_{0 \leqq t \leqq q} D_t$

$$g_{ni}(Y_{j1},\dots,Y_{ji}) \geqq g_{n_t i}(Y_{j1},\dots,Y_{ji}) \geqq \mu(\theta_j) - \varepsilon \ \forall j \leqq m.$$

It then follows that

$$B_r \supset \bigcap_{0 \leqq t \leqq q} D_t.$$

Therefore by (4.13) it follows that

$$P_\theta(B_r^c) \leqq \sum_{t=0}^{q} P_\theta(D_t^c)$$

$$= (q+1)o(c^{-r})$$

$$= o(c^{-r}) \qquad \text{by appendix R3.}$$

ii) Clearly if k is an m-leader then ϕ^* plays k. If k is not an m-leader then for the worst of the m-leaders, say j_n, since $T_n(j_n) \geqq \delta n$,

$$\hat{\mu}_n(j_n) \leqq \max_{j>m} \mu(\theta_j) + \varepsilon \qquad (\text{on } A_r)$$

$$< \mu(\theta_m) - \varepsilon \qquad (\text{by choice of } \varepsilon)$$

Now in case $T_n(k) \geqq \delta n$ we have on A_r

$$\mu(\theta_m) - \varepsilon \leqq \mu(\theta_k) - \varepsilon \leqq h_{T_n(k)}(Y_{k1},\dots,Y_{kT_n(k)}).$$

Hence k should be an m-leader which gives a contradiction.

In case $T_n(k) < \delta n$ we have on B_r

$$\mu(\theta_m) - \varepsilon \leqq \mu(\theta_k) - \varepsilon \leqq g_{nT_n(k)}(Y_{k1},\dots,Y_{kT_n(k)}).$$

Hence ϕ^* will play arm k.

Thus for all comparison instants $n(f,i)$ such that $c^{r-1} \leqq n(f,i) \leqq c^{r+1}$ and $k \equiv i \bmod p$ for some $k \leqq m$, ϕ^* chooses k on $A_r \cap B_r$. Then for all $c^r \leqq n \leqq c^{r+1}$, and $r \geqq r_0$ (sufficiently large) on the event $A_r \cap B_r$, for any $k \leqq m$

$$T_n(k) \geqq \frac{1}{p}(n - c^{r-1} - 2pb(c^{r+1}))$$

where $b(c^{r+1})$ is the block length of the frame $f(r)$ such that $N_{f(r)-1} \leqq c^{r+1} < N_{f(r)}$. By appendix R4 then,

$$T_n(k) \geqq \delta n$$

Thus k is well-sampled.

iii) Now for all $c^r \leqq n \leqq c^{r+1}$, and $r \geqq r_0$ (sufficiently large), on the event $A_r \cap B_r$

$$\begin{aligned}\max\{\hat{\mu}_n(j) : T_n(j) \geqq \delta n,\ j > m\} & \\ &\leqq \max_{j>m} \mu(\theta_j) + \varepsilon \qquad (\text{by } A_r) \\ &< \mu(\theta_m) - \varepsilon \qquad (\text{by choice of } \varepsilon) \\ &\leqq \hat{\mu}(\theta_k) \text{ for all } k \leqq m \qquad (\text{by } A_r \text{ and (ii)}).\end{aligned}$$

Thus the m-leaders are the m-best arms.

iv) Note that from (iii) it follows that, for $r \geqq r_0$ and $c^r \leqq n \leqq c^{r+1}$,

{at least one of the m-leaders is not an
m-best arm at comparison instant $n(f, i)\} \subseteq (A_r \cap B_r)^c = A_r^c \cup B_r^c$.

Thus, by (iii) it follows that

$$\begin{aligned}E_\theta[\text{Term } 3] = \sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} P_\theta\{&\text{at least one of the } m\text{-leaders is not an } m\text{-best arm at} \\ &\text{comparison instant } n(f, i)\} \\ &\leqq \sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} (P_\theta(A_r^c) + P_\theta(B_r^c)) + c^{r_0} + b(c^{r_0}) \\ &\leqq \sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} o(c^{-r}) + c^{r_0} + b(c^{r_0}) \\ &\leqq o(\log N_l) \qquad \text{by appendix R5.}\end{aligned}$$

This concludes the proof of Lemma 4.1 and Claim 3.

Thus, by Claims 1, 2 and 3,

$$E_\theta T_{N_l}(j) \leqq \left(\frac{1}{I(\theta_j, \theta_m)} + o(1)\right) \log N_l.$$

Now we extend this result for any arbitrary n. Let l be such that $N_{l-1} < n \leqq N_l$. Clearly,

$$\begin{aligned}\frac{E_\theta T_n(j)}{\log n} &\leqq \frac{E_\theta[T_{N_l}(j)]}{\log N_{l-1}} \\ &\leqq \left(\frac{1}{I(\theta_j, \theta_m)} + o(1)\right) \frac{\log(p2^{l^2} + pl^2)}{\log(p2^{(l-1)^2})} \\ &= \left(\frac{1}{I(\theta_j, \theta_m)} + o(1)\right).\end{aligned}$$

This completes the proof of (4.9).

Proof of (4.10) This proof is very similar to the proof of Claim 3 and is based on Lemma 4.1. From Lemma 4.1(iii) we know that all the m-leaders are the m-best arms at any comparison instant $n(f,i)$ such that $c^r \leqq n(f,i) \leqq c^{r+1}$ for some $r \geqq r_0$ (sufficiently large) on the event $A_r \cap B_r$.

Thus since all the m-leaders are well sampled, if follows, on A_r, that the point estimates of all the m-best arms lie within $\pm\varepsilon$ of their true means. So by the choice of ε we know that under these conditions all the arms $k \leqq l$ are among the m-leaders and none of them is least best of the m-leaders, hence they will be played. Consequently at the end of frame l.

$$\begin{aligned}E_\theta(N_l - T_{N_l}(j)) &= p - m + \sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} P_\theta\{\text{arm } j \text{ is not used at comparison instant } n(f,i)\} \\ &\leqq p - m + \sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} (P_\theta(A_r^c) + P_\theta(B_r^c) + c^{r_0} + b(c^{r_0})) \\ &\leqq p - m + \sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} o(c^{-r}) + c^{r_0} + b(c^{r_0}) \\ &\leqq o(\log N_l) \qquad \text{by appendix R5.}\end{aligned}$$

The result for general n follows just like as in the proof of (4.9).

Proof of (4.11) Let $\phi \in \{0,1\}^p$ with exactly m l's i.e. a vector representing a combination of m arms that have been selected for play. Let

$$S_n(\phi) := \sum_{i=1}^{n+1} 1(\phi_i = \phi, \phi_{i+1} \neq \phi)$$

and

$$T_n(\phi) := \sum_{i=1}^{n} 1(\phi_i = \phi).$$

Let l be such that $N_{l-1} < n \leqq N_l$. Then,

$$\begin{aligned} S_n(\phi) &\leqq T_{N_0}(\phi) + \sum_{f=1}^{l} \frac{T_{N_f}(\phi) - T_{N_{f-1}}(\phi)}{f} \\ &= T_{N_0}(\phi) + \sum_{f=1}^{l} \frac{T_{N_f}(\phi)}{f} - \sum_{f=0}^{l-1} \frac{T_{N_f}(\phi)}{f+1} \\ &= \frac{T_{N_l}(\phi)}{l} + \sum_{f=1}^{l-1} T_{N_f}(\phi)\left[\frac{1}{f} - \frac{1}{f+1}\right] \\ &\leqq \frac{T_{N_l}(\phi)}{l} + \sum_{f=1}^{l-1} T_{N_f}(\phi)\left[\frac{1}{f^2}\right]. \end{aligned}$$

Let ϕ^* be such that

$$\phi^*(i) = 1, i = 1, \ldots, m$$

$$\phi^*(j) = 0, j = m+1, \ldots, p.$$

Then, for $\phi \neq \phi^*, \phi(j) = 1$ for some $j = m+1, \ldots, p$. Thus,

$$T_n(\phi) \leqq T_n(j)$$

with the above found j.

By (4.9),

$$\limsup_{l \to \infty} \frac{E_\theta T_{N_l}(j)}{\log N_l} \leqq 1/I(\theta_j, \theta_m).$$

Therefore, for any given $\varepsilon > 0$ $\exists f_0$ such that $\forall f \geqq f_0$

$$\frac{E_\theta T_{N_f}(j)}{\log N_f} \leqq 1/I(\theta_j, \theta_m) + \varepsilon.$$

Hence, by (4.3)

$$E_\theta T_{N_f}(j) \leqq (1/I(\theta_j, \theta_m)+\varepsilon)\log(p2^{f^2}+pf^2)$$

$$\leqq K(\varepsilon) f^2 \qquad \text{for some } K(\varepsilon).$$

Thus, for $n > N_{f_0}$

$$E_\theta(S_n(\phi)) \leqq \frac{E_\theta T_{N_l}(j)}{l} + \sum_{f=1}^{l-1} \frac{E_\theta T_{N_f}(j)}{f^2}$$

$$\leqq \frac{K(\varepsilon) l^2}{l} + \sum_{f=f_0}^{l-1} K(\varepsilon) + \sum_{f=1}^{f_0-1} \frac{E_\theta T_{N_f}(j)}{f^2}$$

$$\leqq K(\varepsilon) 2l + M(\varepsilon)$$

$$\text{where } M(\varepsilon) = \sum_{f=1}^{f_0-1} \frac{E_\theta T_{N_f}(j)}{f^2}.$$

Consequently,

$$\frac{E_\theta S_n(\phi)}{\log n} \leqq \frac{E_\theta S_n(\phi)}{\log N_{l-1}} \leqq \frac{K(\varepsilon)2l + M(\varepsilon)}{(l-1)^2} = o(1).$$

Under A4 we have from (2.14) and (2.3)

$$SW_n(\theta) \leqq mC\left[\sum_{\phi \neq \phi^*} ES_n(\phi) + ES_n(\phi^*)\right]$$

$$\leqq mC\left[2 \sum_{\phi \neq \phi^*} ES_n(\phi) + 1\right]$$

$$= o(\log n).$$

This proves (4.11).

Proof of (4.12) By (2.13) we have

$$R'_n(\theta) = n \sum_{j=1}^{m} \mu(\theta_j) - \sum_{j=1}^{p} \mu(\theta_j) E_\theta T_n(j)$$

$$= \sum_{j=m+1}^{p} \mu(\theta_m) E_\theta T_n(j) - \sum_{j=m+1}^{p} \mu(\theta_j) E_\theta T_n(j)$$

$$+\sum_{j=1}^{m}\mu(\theta_j)E_\theta(n-T_n(j))-\sum_{j=m+1}^{p}\mu(\theta_m)E_\theta T_n(j)$$

$$=\sum_{j=m+1}^{p}(\mu(\theta_m)-\mu(\theta_j))E_\theta T_n(j)$$

$$+\sum_{j=1}^{m}(\mu(\theta_j)-\mu(\theta_m))E_\theta(n-T_n(j))$$

$$=\sum_{j=m+1}^{p}(\mu(\theta_m)-\mu(\theta_j))E_\theta T_n(j)$$

$$+\sum_{j=1}^{l}(\mu(\theta_j)-\mu(\theta_m))E_\theta(n-T_n(j)).$$

By (4.9) and (4.10) it then follows that,

$$R'_n(\theta)=\left[\sum_{j=m+1}^{p}(\mu(\theta_m)-\mu(\theta_j))/I(\theta_j,\theta_m)+o(1)\right]\log n. \tag{4.14}$$

Hence, (4.12) follows from (2.15), (4.14) and (4.11). □

In view of Theorems 3.1 and 4.1 the block allocation scheme ϕ^* that we propose in this section is asymptotically efficient, i.e.

$$R_n(\theta)\sim\left[\sum_{j\in\{\sigma(m+1),\ldots,\sigma(p)\}}(\mu(\theta_{\sigma(m)})-\mu(\theta_j))/I(\theta_j,\theta_{\sigma(m)})\right]\log n.$$

Thus, despite the imposition of a switching cost we are able to recapture the same asymptotically optimal performance as Anantharam *et al.* (cf. [3]) achieve in the non-switching cost case. The block allocation scheme proposed in this section is crucial in achieving this performance. By grouping together samples from each inferior population in blocks, we manage to maintain the number of samples from each inferior population at about $\log n/I(\theta_j,\theta_{\sigma(m)})$ and to limit the number of switches to $o(\log n)$.

5. CONCLUSIONS

Despite the inclusion of a switching cost, our allocation scheme achieves the same asymptotic performance as the optimal solutions for the case without switching cost. This is made possible by grouping together samples into blocks of increasing sizes, thereby reducing the number of switches to $o(\log n)$.

Notice that the block length and frame lengths are prescribed in advance and not generated adaptively from the data. With our block scheme if we can ensure that the number of samples from an inferior population is $O(\log n)$ then we automatically control the number of switches to $o(\log n)$. It is worthwhile to point out that the "block structure" we employ in our adaptive allocation scheme is the same as that of [4]. However, the adaptive scheme of [4], based only on upper confidence bounds, does not work in the case of multiple plays. More specifically, we need to prove part (ii) of Theorem 4.1, and for this purpose we need to employ point estimates.

Although in our problem formulation we consider a fixed switching cost, we can equally well handle switching costs which vary with time and with the pair of populations between which switching occurs, provided the switching cost is bounded.

Assumption A4 is essential to obtain asymptotic efficiency. If we do not have unique m-best populations, then the number of switches among the m-best populations can be arbitrarily large.

Acknowledgements

The research of Rajeev Agrawal and Demosthenis Teneketzis was supported in part by NSF Grant No. ECS-8517708 and ONR Grant No. N00014-87-K-0540.

References

[1] T. L. Lai and H. Robbins, Asymptotically efficient adaptive allocation rules, *Adv. Appl. Math.* **6** (1985), 4–42.

[2] T. L. Lai and H. Robbins, Asymptotically efficient allocation of treatments in sequential experiments, in: *Design of Experiments*, T. J. Santner and A. C. Tamhane (eds.), Marcel Dekker, New York, pp. 127–142.

[3] V. Anantharam, P. Varaiya and J. Walrand, Asymptotically efficient allocation rules for multi-armed bandit problem with multiple plays. Part I: I.I.D. Rewards, Part II: Markovian rewards, *IEEE Transactions on Automatic Control* AC-32 (11) (Nov. 1987), 968–982.

[4] R. Agrawal, M. Hegde and D. Teneketzis, Asymptotically efficient allocation rules for multi-armed bandit problem with switching cost, *IEEE Transactions on Automatic Control* AC-33 (10) (Oct. 1988), 899–906.

[5] T. L. Lai, Some thoughts on stochastic adaptive control, *Proc. 23rd IEEE Conf. on Decision and Control*, Las Vegas, Dec. 1984, pp. 51–56.

[6] S. M. Ross, *Stochastic Processes*, Wiley, 1983.

[7] R. Agrawal, Asymptotically efficient allocation schemes for stochastic adaptive optimization problems, Ph.D. Thesis, Dept. of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, August 1988.

APPENDIX

R1: $\sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} \frac{1}{n(f,i)}$

$$= \sum_{f=1}^{l} f \sum_{i=1}^{k_f} \frac{1}{n(f,i)}$$

$$= \frac{1}{N_o} + \frac{1}{N_o+1} + \cdots + \frac{1}{N_1-1} + \frac{1}{N_1} + \frac{1}{N_1} + \frac{1}{N_1+2} + \frac{1}{N_1+2} + \cdots$$

$$+ \frac{1}{N_2-2} + \frac{1}{N_2-2} + \frac{1}{N_2} + \frac{1}{N_2} + \frac{1}{N_2} + \cdots$$

$$\leqq \frac{1}{N_o} + \frac{1}{N_o+1} + \cdots + \frac{1}{N_1-1} + \frac{1}{N_1-1} + \frac{1}{N_1} + \frac{1}{N_1+1} + \frac{1}{N_1+2} + \cdots$$

$$+ \frac{1}{N_2-3} + \frac{1}{N_2-2} + \frac{1}{N_2-2} + \frac{1}{N_2-1} + \frac{1}{N_2} + \cdots$$

$$= \sum_{t=N_o}^{N_l-l-1} \frac{1}{t} + \sum_{j=1}^{l} \frac{1}{N_f - b_f}$$

$$\leqq \sum_{t=1}^{N_l} \frac{1}{t} + \sum_{f=1}^{l} \frac{1}{N_f - b_f}$$

$$\leqq \sum_{t=1}^{N_l} \frac{1}{t} + \sum_{f=2}^{l} \frac{1}{p2^{f^2} - f} + \frac{1}{N_1 - 1}$$

$$\leqq \sum_{t=1}^{N_l} \frac{1}{t} + \frac{1}{p} \sum_{f=2}^{l} \left(\frac{1}{2}\right)^f + \frac{1}{N_1-1}$$

$$\leqq \log N_l + 1 + \frac{1}{p} + \frac{1}{N_1-1}.$$

R2: By the definition of q and $n_t, t=0,\ldots,q$ we have

$$c^{r-1} = n_0 \leqq n_1 \leqq \cdots \leqq n_{q-1} < n_q = \left[\frac{c^{r-1}}{\delta^q}\right] \geqq c^{r+1}.$$

Since

$$c^{r-1} \leqq n \leqq c^{r+1}, \exists t \in \{0,1,\ldots,q-1\} \ni n_t \leqq n \leqq n_{t+1}.$$

Now assume $n_t < i$. Then,

$$\frac{c^{r-1}}{\delta^t} < i \Rightarrow \frac{c^{r-1}}{\delta^{t+1}} < \frac{i}{\delta}$$

$$\Rightarrow \left[\frac{c^{r-1}}{\delta^{t+1}}\right] < \frac{i}{\delta} \leqq n$$

$$\Rightarrow n_{t+1} < n, \text{ which is a contradiction.}$$

So $n_{t+1} > n \geqq n_t \geqq i$ for some $t \in \{0, 1, \dots, q-1\}$.

R3:

$$\left[\frac{c^{r-1}}{\delta^{q-1}}\right] < c^{r+1} \Rightarrow \frac{c^{r-1}}{\delta^{q-1}} < c^{r+1}$$

$$\Rightarrow \delta^{q-1} > c^{-2}$$

$$\Rightarrow (q-1)\log\delta > \log c^{-2}$$

$$\Rightarrow q+1 < \frac{\log c^{-2}}{\log\delta} + 2 = \text{const.} > 0.$$

Thus $(q+1)o(c^{-r}) = o(c^{-r})$.

R4: By the choice of c, $(1-c^{-1})/p > \delta$. Therefore,

$$\frac{1}{p}(n - c^{r-1} - 2pb(c^{r+1})) = \frac{n}{p}\left(1 - \frac{c^{r-1}}{n} - \frac{2pb(c^{r+1})}{n}\right)$$

$$\geqq \frac{n}{p}\left(1 - \frac{c^{r-1}}{c^r} - \frac{2pb(c^{r+1})}{c^r}\right) \quad (\text{as } n \geqq c^r)$$

$$= \frac{n}{p}\left(1 - c^{-1} - \frac{2pb(c^{r+1})}{c^r}\right).$$

Also

$$2^{(f(r)-1)^2} \leqq N_{f(r)-1} \leqq c^{r+1} < N_{f(r)},$$

and

$$b(c^{r+1}) = b_{f(r)} = f(r) \leqq \sqrt{\frac{\log c^{r+1}}{\log 2}} + 1.$$

Hence,

$$\frac{2pb(c^{r+1})}{c^r} \to 0 \quad \text{as} \quad r \to \infty.$$

Therefore for sufficiently large $r(\geqq r_0)$

$$\frac{n}{p}(1 - c^{r-1} - 2pb(c^{r+1})) > n\left(\frac{1}{p}(1 - c^{-1}) - \frac{\rho}{p}\right) > n\delta$$

where

$$\rho < \frac{1}{p}(1 - c^{-1}) - \delta.$$

R5: In what follows r is a function of $n(f, i)$ satisfying $c^r \leqq n(f, i) \leqq c^{r+1}$. However, we shall suppress this dependence for the sake of notational convenience.

$$\sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} o(c^{-r}) \leqq \sum_{r=0}^{r(l)} \left[\sum_{n=c^r}^{c^{r+1}} o(c^{-r}) + b(c^r)[o(c^{-(r-1)})]\right]$$

where $r(l)$ is such that

$$c^{r(l)} < N_l \leqq c^{r(l)+1},$$

and $b(c^r)$ is the block length of the frame $f(c^r)$ such that

$$N_{f(c^r)-1} \leqq c^r < N_{f(c^r)}.$$

Now, N_i increases as $p2^{i^2}$ where as c^i increases as c^i. Clearly there exists an $r' \ni \forall\, r \geqq r'$, $c^r < p2^{r^2}$, so that c^r lies in a frame before frame r. Thus $b(c^r) < b_r = r$. Thus,

$$\begin{aligned}\sum_{f=1}^{l} b_f \sum_{i=1}^{k_f} o(c^{-r}) &\leqq K(r') + \sum_{r=r'}^{r(l)} [o(1) + ro(c^{-r})] \\ &\leqq K(r') + r(l)o(1) + o\left(\sum_{r=r'}^{r(l)} rc^{-r}\right) \\ &\leqq K(r') + o\left(\frac{\log N_l}{\log c}\right) + o\left(\sum_{r=r'}^{r(l)} 1\right) \\ &= o(\log N_l).\end{aligned}$$