Proceedings of 25th Conference
on Decision and Control
Athens, Greece • December 1986

Asymptotically Efficient Rules in Multiarmed Bandit Problems

V. Anantharam and P. Varaiya

Department of Electrical Engineering and Computer Sciences
and Electronics Research Laboratory
University of California, Berkeley CA 94720

Setup

We are given N discrete-time real-valued stochastic processes

$$X^1: X^1(1), X^1(2), \cdots$$

$$\cdots$$

$$X^N: X^N(1), X^N(2), \cdots.$$

The essential assumption is that these processes are *independent*. For historical reasons these processes are also called *arms*.

A fixed number m, $1 \leqslant m \leqslant N$, is specified. At each time t we must select m different arms. Let $T^j(t)$ be the number of times that arm j was selected during the interval $1, \cdots, t$; and let $U(t) \subset \{1, \cdots, N\}$ be the m arms that are selected at time t. Then at time t we receive the reward

$$Y(t) = \sum_{j \in U(t)} X^j[T^j(t)],$$

and the selection of the arms at time $t+1$ can be based on the information

$$I(t) = \{X^j(s) \mid s = 1, \cdots, T^j(t);\ j = 1, \cdots, N\}.$$

Let Φ be a selection rule. Our aim is to find a rule Φ so as to maximize the expected cumulative reward

$$J(\Phi, t) = \sum_{s=1}^{t} EY(s). \tag{1}$$

Suppose each arm X^j is a sequence of iid variables $X^j(1), X^j(2), \cdots$ with probability density $f(x, \theta_j)\nu(dx)$ relative to some common measure ν on R. Let $\mu(\theta) = \int x f(x, \theta)\nu(dx)$ be the mean. The parameters θ_j that characterize the arms are *not* known. Let $C = (\theta_1, \cdots, \theta_N) \in R^N$ be the unknown parameter configuration. Then the reward (1) corresponding to C is

$$J(\Phi, C, t) = \sum_{j=1}^{N} \mu(\theta_j) ET^j(t). \tag{2}$$

Let σ be a permutation of $\{1, \cdots, N\}$ depending on C so that

$$\mu[\theta_{\sigma(1)}] \geqslant \cdots \geqslant \mu[\theta_{\sigma(N)}]. \tag{3}$$

If C were known from the beginning, the best choice would be always to select arms $\sigma(1), \cdots, \sigma(m)$, and the total expected reward up to t would then be

$$\sum_{j=1}^{m} \mu[\theta_{\sigma(j)}]t.$$

So we may define the *regret* of Φ at t and C as

$$R(t, C, \Phi) = \sum_{j=1}^{m} \mu[\theta_{\sigma(j)}]t - \sum_{j=1}^{N} \mu[\theta_j] ET^j(t), \tag{4}$$

and we want to find a rule to

$$\underset{\Phi}{\text{Minimize}}\ R(t, C, \Phi) \text{ for all } t \text{ and } C. \tag{5}$$

It is evident that there will not exist Φ that minimizes the regret "for all t and C". Indeed if such Φ exists it must achieve identically zero regret because the rule that always selects a fixed set of m arms gives zero regret for configurations C for which those m arms have the largest means, but for every other configuration this rule is quite bad since it will have regret *proportional* to t. This suggests that in order to exclude such non-learning or non-adaptive rules from consideration we should modify (5) keeping "for all C" while relaxing the condition "for all t". One way of doing this is to replace (5) with an expected average reward over time in a Bayesian setting.

In such a Bayesian setting we suppose given a prior distribution $P_j(d\theta_j)$ for θ_j and we try to minimize

$$J(\Phi) = \limsup \frac{1}{t} \int R(t, \theta_1, \cdots, \theta_N, \Phi)$$

$$P_1(d\theta_1) \cdots P_N(d\theta_N). \tag{6}$$

It is an easy matter to construct near-optimal rules for this problem. Note that (under some simple restrictions on the density $f(x, \theta)$) the mean $\mu(\theta_j)$ can be accurately estimated by the sample mean, i.e., for $\delta > 0$, there exists $T < \infty$ such that for every j

$$P_j\{|\mu(\theta_j) - \frac{1}{T}\sum_{t=1}^{T} X^j(t)| < \delta\} > 1 - \delta. \tag{7}$$

Now consider the following two-phased rule Φ_δ: In the first or estimation phase the rule selects each of the N arms at least T times, and in the second phase the rule selects the m arms with the largest sample means at the end of the estimation phase. From (7) it follows that this rule is near-optimal since

$$J(\Phi_\delta) \to 0 \text{ as } \delta \to 0.$$

Although Φ_δ is better than a fixed, non-adaptive rule, there are good reasons for not considering it close to optimal. First, observe that no matter how small $\delta > 0$ is, with positive probability the m arms selected will not have the largest mean values for a set of configurations C, and for each of these configurations the regret will grow proportional to t. Second, as will be seen below, there *do* exist rules Φ for which $\frac{1}{t}R(t, C, \Phi) \to 0$ for *every* C. Such a rule is qualitatively superior to Φ_δ, and leads us to conclude that the Bayesian loss (6) while it excludes non-adpative rules, it does not adequately discriminate among adaptive rules.

This discussion suggests that we should impose the *adaptation* requirement on admissible rules:

$$\limsup \frac{1}{t} R(t, C, \Phi) = 0, \text{ for all } C. \tag{8}$$

This is a significant restriction. For example, it excludes all rules that, like Φ_δ, stop learning after a predetermined finite time. Indeed, it is not at all clear whether there exist rules satisfying (8). Furthermore, if there is such a rule Φ, then any rule Φ' that selects the same arms as Φ except over $n(t)$ time instants during $1, \cdots, t$ with $\frac{n(t)}{t} \to 0$ will also satisfy (8). This brings us to finally to the prob-

CH2344-0/86/0000-0251 $1.00

lem of distinguishing among arms that satisy (8) and to the work of Lai and Robbins.

Asymptotically efficient adaptive rules

In a remarkable study [2-4] Lai and Robbins posed and answered the question of asymptotically efficient adaptive rules. Their work deals with the case $m = 1$. We summarize here its extension to $m > 1$ by Anantharam [1].

Recall that an arm is described by iid rewards with distribution $f(x,\theta)\nu(dx)$ and mean $\mu(\theta) = \int x f(x,\theta)\nu(dx)$. For a configuration $C = (\theta_1, \cdots, \theta_N)$ let σ be a permutation so that (3) holds. Let $0 \leq l < m \leq n \leq N$ be such that

$$\mu[\theta_{\sigma(1)}] \geq \cdots \geq \mu[\theta_{\sigma(l)}] > \mu[\theta_{\sigma(l+1)}] = \cdots = \mu[\theta_{\sigma(m)}] =$$
$$= \cdots = \mu[\theta_{\sigma(n)}] > \mu[\theta_{\sigma(n+1)}] \geq \cdots \geq \mu[\theta_{\sigma(N)}].$$

We call $\sigma(1), \cdots, \sigma(l)$ the *best* arms, $\sigma(l+1), \cdots, \sigma(n)$ the *border* arms, and $\sigma(n+1), \cdots, \sigma(N)$ the *worst* arms.
[Note: If $\mu[\theta_{\sigma(m)}] > \mu[\theta_{\sigma(m+1)}]$, then $\sigma(m)$ is simultaneously a best and border arm.]

A selection rule Φ is said to be *uniformly good* if $R(t,C,\Phi) = o(t^\alpha)$ for every $\alpha > 0$ and every C. From (4) it follows that Φ is uniformly good iff

$$E[t - T^j(t)] = o(t^\alpha) \text{ for every best arm } j,$$
$$E[T^j(t)] = o(t^\alpha) \text{ for every worst arm } j,$$

for every $\alpha > 0$ and every C.

The Kullback-Liebler number

$$I(\theta,\lambda) = \int \log\frac{f(x,\theta)}{f(x,\lambda)} f(x,\theta)\nu(dx)$$

is a well-known measure of dissimilarity between two distributions. In general $0 \leq I(\theta,\lambda) \leq \infty$ Define conditions A1-A4.

A1. $\mu(\theta)$ is strictly increasing in θ.
A2. $0 < I(\theta,\lambda) < \infty$ for $\lambda > \theta$.
A3. $I(\theta,\lambda)$ is continuous in $\lambda > \theta$ for fixed θ.
A4. For all λ, and all $\delta > 0$, there exists λ' with $\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta$.

Theorem 1. Suppose A1-A4 hold. Let Φ be any uniformly good rule and $C = (\theta_1, \cdots, \theta_N)$ be any configuration. Then

$$\liminf \frac{R(t,C,\Phi)}{\log t} \geq \sum_{j \text{ is worst}} \frac{[\mu(\theta_{\sigma(m)}) - \mu(\theta_j)]}{I(\theta_j, \theta_{\sigma(m)})}. \qquad (9)$$

Thus every uniformly good rule must select each worst arm j at least $[I(\theta_j, \theta_{\sigma(m)})]^{-1}\log t$ times during $1, \cdots, t$. This number decreases as the "information distance" $I(\theta_j, \theta_{\sigma(m)})$ between arm j and the arm $\sigma(m)$ with the mth largest mean increases.
[Remark: Unlike the mean $\mu(\lambda)$, the information distance $I(\theta,\lambda)$ need not increase with λ; however, that assumption is needed in Theorem 3.]

As example, in the Gaussian case, $f(x,\theta)\nu(dx) = N(\theta,\sigma^2)$ so $\mu(\theta) = \theta$. Then $I(\theta,\lambda) = (\theta-\lambda)^2/2\sigma^2$ and we get

$$\liminf \frac{R(t,C,\Phi)}{\log t} \geq \sum_{j \text{ is worst}} \frac{2\sigma^2}{\theta_{\sigma(m)} - \theta_j}.$$

Say that a rule Φ is **asymptotically efficent** if its regret achieves the lower bound (9) for every C.

The crucial feature in constructing an asymptotically efficient rule is this. At time t we have $T^j(t)$ observations of arm j from which we can estimate its mean. At $t+1$ we must decide whether to select the m arms whose estimated mean values are the largest -- "play the winners" rule -- or to select an apparently losing arm. The idea is to consider an apparently losing arm, say arm j, to estimate an *upper* bound for its mean value, and to compare that estimate with the estimate of the least best of the apparent winners.

We now describe a rule that is asymptotically efficient under the additional conditions A5, A6.

A5. $\log f(x,\theta)$ is concave in θ for each fixed x.
A6. $\int x^2 f(x,\theta)\nu(dx) < \infty$ for each θ.
Assumption A5 implies that $I(\theta,\lambda)$ is convex in λ, and since I is minimized at $\lambda = \theta$, it is increasing in λ for $\lambda > \theta$.

Let $X(1), X(2), \cdots$ be the sequence of rewards from an arm. Let $h:(0,\infty) \to (0,\infty)$ be a fixed continuous function with $\int h(s)ds = 1$, and let

$$W(a,\theta) = \int_0^\infty \prod_{b=1}^{a} \frac{f(X(b),\theta - s)}{f(X(b),\theta)} h(s)ds.$$

[A5 implies that $W(a,\theta)$ is increasing in θ.]

For $K > 0$, let

$$U(a, X(1), \cdots, X(a), K) = \inf\{\theta \mid W(a,\theta) > K\},$$

and, lastly, for a fixed $p > 1$, let

$$g(t, a, X(1), \cdots, X(a)) = \mu[U(a, X(1), \cdots, X(a), t(\log t)^p].$$

Now consider the following rule:

1. In the first N steps select each arm m times in order to establish an initial estimate.

2. Fix $0 < \delta < 1/N^2$. At any time t say that arm j is *well-sampled* if $T^j(t) > \delta t$. Then there are at least m well-sampled arms when $t > N$. At each t, from among the well-sampled arms choose the m *leaders* ranked by the sample mean $\mu^j(t)$ for arm j:

$$\mu^j(t) = \frac{X^j(1) + \cdots + X^j[T^j(t)]}{T^j(t)}.$$

Now consider the decision at $t+1$. Consider the arm j for which $t+1 = j \bmod N$, and estimate the upper bound for its mean:

$$\bar{\mu}^j(t) = g[t, T^j(t), X^j(1), \cdots, X^j(T^j(t))].$$

(a) If arm j is already one of the m leaders at time t, then select the m leaders at $t+1$.
(b) If arm j is not one of the leaders at t, and if its upper bound $\bar{\mu}^j(t) < \mu^k(t)$ for every m leader k, then again select the m leaders at $t+1$.
(c) If arm j is not one of the leaders at t, and if $\bar{\mu}^j \geq \mu^k(t)$ where k is a leader with the least mean estimate, then at $t+1$ select the $(m-1)$ leaders other than k and arm j.

Note that at each time $(m-1)$ well-sampled arms with the largest estimated means are always selected.

Theorem 2. Suppose A1-A6 hold. Then this rule is asymptotically efficient.

Final remarks

Theorems 1 and 2 also hold without the "denseness" condition A4. They have been extended to the important case where the arms are finite Markov chains with stationary transition probability matrix depending upon one unknown parameter, see [1]. For several families of distributions, including Bernoulli, Poisson, Gaussian and double exponential, the statistics $g(t,a)$ can be calculated recursively, see [4].

Condition A5 is essential in the proof of Theorem 2. It would seem, however, that asymptotically efficient rules should exist under the weaker condition that $I(\theta,\lambda)$ is increasing in λ for $\lambda > \theta$.

Acknowledgements

Research supported in part by the Joint Services Electronics Program Contract F49620-84-C-0057. This paper was written while Varaiya was visiting the International Institute of Applied Systems Analysis (IIASA). The kind hospitality of IIASA is acknowledged with pleasure.

References

[1] Anantharam, V., Ph.D Dissertation, Univ. of California, Berkeley, 1986.
[2] Lai, T.L., "Some thoughts on stochastic adaptive control," *Proc. 23rd IEEE Conf. on Decision and Control*, Las Vegas, Dec. 1984, 51-56.
[3] Lai, T.L. and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, 1985, 4-22.
[4] Lai, T.L. and H. Robbins, "Asymptotically efficient allocation of treatments in sequential experiments," in Santner, T.J. and A.C. Tamhane (eds) *Design of Experiments*, New York, Marcel Dekker, 1985, 127-142.