Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 8584–8585, September 1995
Statistics

Sequential choice from several populations

MICHAEL N. KATEHAKIS AND HERBERT ROBBINS

Rutgers University, New Brunswick, NJ 08903

Contributed by Herbert Robbins, May 4, 1995

ABSTRACT **We consider the problem of sampling sequentially from two or more populations in such a way as to maximize the expected sum of outcomes in the long run.**

Introduction. Consider $N \geq 2$ populations specified by random variables Y_{ij}, $i = 1, \ldots, N$, $j = 1, 2, \ldots$; Y_{ij} denotes the outcome from population i the jth time it is sampled. Our objective is to sample sequentially from the N populations in such a way as to maximize the expected sum of outcomes in the long run.

We first consider the case in which the Y_{ij} are independent normal random variables with unknown means μ_i and known variances σ_i^2. At the end we indicate how these assumptions can be relaxed.

For any policy π and any $t \geq 1$, let π_t, $n_i(t) = \Sigma_{j=1}^{t} \mathbf{1}\{\pi_j = i\}$ denote, respectively, the population sampled at time t and the total number of times population i has been sampled during times 1, 2, ..., t. The expected sum of the first t outcomes under the policy π is $V_{\underline{\mu}}^{\pi}(t) = \mathbf{E}_{\underline{\mu}}^{\pi} \Sigma_{i=1}^{N} \Sigma_{j=1}^{n_i(t)} Y_{ij}$, a function of the true values $\underline{\mu} = (\mu_1, \ldots, \mu_N)$. The regret due to ignorance of $\underline{\mu}$ when the policy π is employed is $R_{\underline{\mu}}^{\pi}(t) = t\mu^* - V_{\underline{\mu}}^{\pi}(t)$, where $\mu^* = \max\{\mu_i\}$. Maximization of $V_{\underline{\mu}}^{\pi}(t)$ with respect to π is equivalent to minimization of $R_{\underline{\mu}}^{\pi}(t)$.

Let $\mathcal{C}$ denote the class of policies π for which $R_{\underline{\mu}}^{\pi}(t) = o(t^{\alpha})$ as $t \to \infty$ for all $\alpha > 0$ and all $\underline{\mu}$, and let π^0 denote the policy specified as follows. At times $t = 1, 2, \ldots, N$, $\pi_t^0 = t$, and at any subsequent time $t > N$, given sample estimates $\bar{Y}_{in_i(t)} = \Sigma_{j=1}^{n_i(t)} Y_{ij}/n_i(t)$ of size $n_i(t)$ for μ_i, π_{t+1}^0 is the i for which the "index"

$$u_i(t, \bar{Y}_{in_i(t)}) = \bar{Y}_{in_i(t)} + \sigma_i(2 \log t/n_i(t))^{1/2} \qquad [1]$$

is largest; for notational simplicity we have suppressed the dependence of $u_i(t, \bar{Y}_{in_i(t)})$ on $n_i(t)$. We prove that $\pi^0 \in \mathcal{C}$ and is optimal in the sense of *Theorem 1* below.

The ideas involved in this paper represent a considerable simplification of the adaptive policy and the proofs employed in Lai and Robbins (1). They have been extended in Burnetas and Katehakis (2) to sequential allocation problems with populations specified by densities that depend on a vector of parameters and in Burnetas and Katehakis (3) to dynamic programming. For related work see also Agrawal, Teneketzis, and Anantharam (4), Yakowitz and Lowe (5), Li and Zhang (6), and Burnetas and Katehakis (7).

The Optimality Theorem. We first prove two lemmata. Fix $\varepsilon > 0$ and define the statistics

$$n_i^1(t, \varepsilon) = \sum_{j=N}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, u_i(j, \bar{Y}_{in_i(j)}) > \mu^* - \varepsilon\},$$

$$n_i^2(t, \varepsilon) = \sum_{j=N}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, u_i(j, \bar{Y}_{in_i(j)}) \leq \mu^* - \varepsilon\}.$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Let $\mathbf{I}(\mu_i, \mu_i') = (\mu_i - \mu_i')^2/2\sigma_i^2$ denote the Kullback–Leibler information number for the normal densities $N(\mu_i, \sigma_i^2)$ and $N(\mu_i', \sigma_i^2)$.

LEMMA 1. *If* $\mu_i < \mu^*$ *then* $\lim_{\varepsilon \to 0} \overline{\lim}_{t \to \infty} \mathbf{E}_{\underline{\mu}}^{\pi^0} n_i^1(t, \varepsilon)/\log t \leq 1/\mathbf{I}(\mu_i, \mu^*) < \infty$.

Proof: Choose $\delta > 0$ and write $\mathbf{E}_{\underline{\mu}}^{\pi^0} n_i^1(t, \varepsilon)$ as $\mathbf{E}_{\underline{\mu}}^{\pi^0} \Sigma_1 + \mathbf{E}_{\underline{\mu}}^{\pi^0} \Sigma_2$, where

$$\Sigma_1 = \sum_{j=N}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, u_i(j, \bar{Y}_{in_i(j)}) > \mu^* - \varepsilon, \text{ and } \bar{Y}_{in_i(j)} \leq \mu_i + \delta\}$$

and

$$\Sigma_2 = \sum_{j=N}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, u_i(j, \bar{Y}_{in_i(j)}) > \mu^* - \varepsilon, \text{ and } \bar{Y}_{in_i(j)} > \mu_i + \delta\}.$$

From the definition of $u_i(j, \bar{Y}_{in_i(j)})$ we get sample-pathwise

$$\begin{aligned}
\Sigma_1 &\leq \sum_{j=N}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, \log j/n_i(j) > (\mu^* - \varepsilon - \mu_i - \delta)^2/2\sigma_i^2\} \\
&\leq \sum_{j=N}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, n_i(j) < \log j/\mathbf{I}(\mu_i, \mu^* - \varepsilon - \delta)\} \\
&\leq \sum_{j=N}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, n_i(j) < \log t/\mathbf{I}(\mu_i, \mu^* - \varepsilon - \delta)\} \\
&\leq \log t/\mathbf{I}(\mu_i, \mu^* - \varepsilon - \delta),
\end{aligned}$$

where the last inequality requires the following counting argument.

Let X_t and $c_t \geq 0$ be two sequences of constants (or random variables), and for any fixed i let $n(t) = \Sigma_{j=1}^{t} \mathbf{1}\{X_j = i\}$. The definition of $n(t)$ implies that the following inequality holds, pointwise in the case of random variables:

$$\sum_{j=1}^{t} \mathbf{1}\{X_{j+1} = i, n(j) \leq c_t\} \leq c_t. \qquad [2]$$

It follows from the above that

$$\mathbf{E}_{\underline{\mu}}^{\pi^0} \Sigma_1 \leq \log t/\mathbf{I}(\mu_i, \mu^* - \varepsilon - \delta). \qquad [3]$$

For Σ_2 note that the following relations hold sample-pathwise:

$$\begin{aligned}
\Sigma_2 &\leq \sum_{j=N}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, \bar{Y}_{in_i(j)} > \mu_i + \delta\} \\
&= \sum_{j=1}^{t} \sum_{k=1}^{j} \mathbf{1}\{\pi_{j+1}^0 = i, \bar{Y}_{in_i(j)} > \mu_i + \delta, n_i(j) = k\} \\
&= \sum_{k=1}^{t} \sum_{j=k}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, \bar{Y}_{ik} > \mu_i + \delta, n_i(j) = k\} \\
&\leq \sum_{k=1}^{t} \mathbf{1}\{\bar{Y}_{ik} > \mu_i + \delta\} \sum_{j=k}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, n_i(j) = k\}
\end{aligned}$$

$$\leq \sum_{k=1}^{t} \mathbf{1}\{\bar{Y}_{ik} > \mu_i + \delta\},$$

where the last inequality is due to the fact that $\Sigma_{j=k}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, n_i(j) = k\} \leq 1$. It follows that

$$\mathbf{E}_{\underline{\mu}}^{\pi^0} \Sigma_2 \leq \mathbf{E}_{\mu_i} \sum_{k=1}^{t} \mathbf{1}\{\bar{Y}_{ik} > \mu_i + \delta\}$$

$$= \sum_{k=1}^{t} \mathbf{P}_{\mu_i}\{\bar{Y}_{ik} > \mu_i + \delta\} = o(\log t) \text{ as } t \to \infty, \qquad [4]$$

where the last equality is a consequence of the tail inequality $1 - \Phi(x) < \phi(x)/x$ for the standard normal distribution, which implies that $\mathbf{P}_{\mu_i}\{\bar{Y}_{ik} > \mu_i + \delta\} = 1 - \Phi(k^{1/2}\delta/\sigma_i) \leq \phi(k^{1/2}\delta/\sigma_i)/(k^{1/2}\delta/\sigma_i) = (2\pi)^{-1/2} e^{-k\delta^2/2\sigma_i^2}/(k^{1/2}\delta/\sigma_i) = 0(1/k)$ as $k \to \infty$.

The proof is now easy to complete, using inequality **3** and Eq. **4**. □

LEMMA 2. *For any* $\underline{\mu}$, $lim_{\varepsilon \to 0} \overline{lim}_{t \to \infty} \mathbf{E}_{\underline{\mu}}^{\pi^0} n_i^2(t, \varepsilon)/log\ t = 0$.

Proof: Recall that for $j > N$, policy π^0 always chooses the population with largest index; i.e., if $\pi_{j+1}^0 = i$, then $u_i(j, \bar{Y}_{in_i(j)}) = \max_a u_a(j, \bar{Y}_{an_a(j)})$. Thus for $j > N$ the event $\{\pi_{j+1}^0 = i, u_i(j, \bar{Y}_{in_i(j)}) \leq \mu^* - \varepsilon\}$ is a subset of $\{\pi_{j+1}^0 = i, u_a(j, \bar{Y}_{an_a(j)}) \leq \mu^* - \varepsilon\}$, where a is any population with mean $\mu_a = \mu^*$. The latter event is obviously a subset of $\{u_a(j, \bar{Y}_{an_a(j)}) \leq \mu^* - \varepsilon\}$. Using this observation and the definition of $u_a(j, \bar{Y}_{an_a(j)})$, we obtain the following inequalities that hold sample-pathwise:

$$n_i^2(t, \varepsilon) \leq \sum_{j=N}^{t} \mathbf{1}\{\pi_{j+1}^0 = i, u_a(j, \bar{Y}_{an_a(j)}) \leq \mu_a - \varepsilon\}$$

$$= \sum_{j=N}^{t} \sum_{k=1}^{j} \mathbf{1}\{u_a(j, \bar{Y}_{ak}) \leq \mu_a - \varepsilon, n_a(j) = k\}$$

$$\leq \sum_{j=N}^{t} \sum_{k=1}^{j} \mathbf{1}\{\bar{Y}_{ak} \leq \mu_a - \varepsilon - \sigma_a(2 \log j/k)^{1/2}\}.$$

Hence

$$\mathbf{E}_{\underline{\mu}}^{\pi^0} n_i^2(t, \varepsilon) \leq \sum_{j=N}^{t} \sum_{k=1}^{j} \mathbf{P}_{\mu_a}\{\bar{Y}_{ia} \leq \mu^* - \varepsilon - \sigma_a(2 \log j/k)^{1/2}\}$$

$$\leq \sum_{j=N}^{t} o(1/j) = o(\log t) \text{ as } t \to \infty; \qquad [5]$$

the last inequality follows from the tail inequality and the fact that $\mu_a = \mu^*$. Indeed, for any population a, not necessarily with $\mu_a = \mu^*$, we have

$$\sum_{k=1}^{j} \mathbf{P}_{\mu_a}\{\bar{Y}_{ak} \leq \mu_a - \varepsilon - \sigma_a(2 \log j/k)^{1/2}\}$$

$$= \sum_{k=1}^{j} \Phi(-\varepsilon_a k^{1/2} - (2 \log j)^{1/2})$$

$$\leq \sum_{k=1}^{j} \phi(\varepsilon_a k^{1/2} + (2 \log j)^{1/2})/(\varepsilon_a k^{1/2} + (2 \log j)^{1/2})$$

$$\leq c/(j(2 \log j)^{1/2}) = o(1/j) \text{ as } j \to \infty,$$

where $\varepsilon_a = \varepsilon/\sigma_a$ and $c = c(\varepsilon_a) > 0$. The lemma follows. □

We can now prove the optimality theorem.

THEOREM 1.

(i) $\pi^0 \in \mathcal{C}$.

(ii) *For all* $\underline{\mu}$, $\mathrm{R}_{\underline{\mu}}^{\pi_0}(\mathrm{t}) = \mathrm{M}(\underline{\mu}) log\ \mathrm{t} + \mathrm{o}(log\ \mathrm{t})$ *as* $\mathrm{t} \to \infty$, *where* $\mathrm{M}(\underline{\mu}) = \Sigma_{\mathrm{i}:\mu_i < \mu^*} (\mu^* - \mu_i)/\mathbf{I}(\mu_i, \mu^*)$.

(iii) *For any* $\pi \in \mathcal{C}$, $\overline{lim}_{t \to \infty} \mathrm{R}_{\underline{\mu}}^{\pi^0}(\mathrm{t})/\mathrm{R}_{\underline{\mu}}^{\pi}(\mathrm{t}) \leq 1$.

Proof: For i note that sample-pathwise, $1 \leq n_i(t) \leq 2 + n_i^1(t, \varepsilon) + n_i^2(t, \varepsilon)$. Thus, using *Lemmata 1* and *2*, we obtain that if $\mu_i < \mu^*$, then $\overline{lim}_{t \to \infty} \mathbf{E}_{\underline{\mu}}^{\pi^0} n_i(t)/\log t \leq 1/\mathbf{I}(\mu_i, \mu^*) < \infty$.

The proof of i is now easy to complete, since for any π

$$R_{\underline{\mu}}^{\pi}(t) = t\mu^* - \mathbf{E}_{\underline{\mu}}^{\pi} \sum_{i=1}^{N} \sum_{j=1}^{n_i(t)} Y_{ij} = \sum_{i=1}^{N} (\mu^* - \mu_i) \mathbf{E}_{\underline{\mu}}^{\pi} n_i(t).$$

ii follows from the proof of i and Theorem 1 in Lai and Robbins (1), which implies that for all π in $\mathcal{C}$ and all $\underline{\mu}$

$$\overline{\lim}_{t \to \infty} R_{\underline{\mu}}^{\pi}(t)/\log t \geq \sum_{i:\mu_i < \mu^*} (\mu^* - \mu_i)/\mathbf{I}(\mu_i, \mu^*) \text{ as } t \to \infty.$$

To prove *iii* we need only divide $R_{\underline{\mu}}^{\pi^0}(t)$ and $R_{\underline{\mu}}^{\pi}(t)$ by $M(\underline{\mu})\log t$ when $M(\underline{\mu}) > 0$. If $M(\underline{\mu}) = 0$ then $R_{\underline{\mu}}^{\pi}(t) = 0$ for all π and t, so we define $0/0 = 1$. □

Remark: Note that *Lemmata 1* and *2* hold for arbitrary sequences of random variables when there exist constants $\mu_i = \lim_{k \to \infty} \mathbf{E}_i \bar{Y}_{ik}$ and index sequences

$$u_i(t, \bar{Y}_{in_i(t)}) = \bar{Y}_{in_i(t)} + h_i(\log t/n_i(t)) \qquad [6]$$

that satisfy the following conditions:

(*A1*) for any $\varepsilon > 0$ there exist constants $C_i = C_i(\mu^*, \mu_i, \varepsilon)$ such that $h_i(\log t/n_i(t)) > \mu^* - \mu_i - \varepsilon$ if and only if $n_i(t) < \log t/C_i(\mu^*, \mu_i, \varepsilon)$,

(*A2*) $\mathbf{P}_i\{\bar{Y}_{ik} > \mu_i + \delta\} = o(1/k)$ as $k \to \infty$, $\forall \delta > 0$,

(*A3*) $\Sigma_{l=1}^{k} \mathbf{P}_i\{\bar{Y}_{ik} \leq \mu^* - \varepsilon - h_i(\log j/k)\} = o(1/j)$ as $j \to \infty$, $\forall \varepsilon > 0$.

In this case the arguments in the proof of part i of the optimality theorem hold if the regret is defined as $R_{\underline{\mu}}^{\pi}(t) = \Sigma_{i=1}^{N} (\mu^* - \mu_i) \mathbf{E}_i^{\pi} n_i(t)$.

Thus, $\overline{\lim}_{t \to \infty} R_{\underline{\mu}}^{\pi^0}(t)/\log t \leq \Sigma_{i:\mu < \mu^*} (\mu^* - \mu_i)/C_i(\mu_i, \mu^*, 0)$; i.e., $R_{\underline{\mu}}^{\pi^0}(t) = O(\log t)$, and policy π^0, defined analogously, is in $\mathcal{C}$. However, there is no analog of Theorem 1 in Lai and Robbins (1). Note also that with this definition the regret $R_{\underline{\mu}}^{\pi}(t)$ no longer represents the quantity $t\mu^* - \mathbf{E}_i^{\pi} \Sigma_{i=1}^{N} \Sigma_{j=1}^{n_i(t)} Y_{ij}$.

1. Lai, T. & Robbins, H. (1985) *Adv. Appl. Math.* **6,** 4–22.
2. Burnetas, A. N. & Katehakis, M. N. (1994) *Optimal Adaptive Policies for Dynamic Programming* (Rutgers University, New Brunswick, NJ), Tech. Rep. **34**.
3. Burnetas, A. N. & Katehakis, M. N. (1994) *Adv. Appl. Math,* in press.
4. Agrawal, R., Teneketzis, D. & Anantharam, V. (1989) *IEEE Trans. Autom. Control* **34,** 1249–1259.
5. Yakowitz, S. & Lowe, W. (1991) *Ann. Oper. Res.* **28,** 297–312.
6. Li, Z. & Zhang, C. (1992) *J. R. Statist. Soc. Ser. B* **54,** 609–616.
7. Burnetas, A. N. & Katehakis, M. N. (1993) *Probl. Eng. Info. Sci.* **7,** 85–119.