MATHEMATICS OF OPERATIONS RESEARCH
Vol. 11, No. 1, February 1986
Printed in U.S.A.

LINEAR PROGRAMMING FOR FINITE STATE MULTI-ARMED BANDIT PROBLEMS*

YIH REN CHEN AND MICHAEL N. KATEHAKIS

State University of New York at Stony Brook

We consider the multi-armed bandit problem. We show that when the state space is finite the computation of the dynamic allocation indices can be handled by linear programming methods.

1. Introduction. An important sequential control problem with a tractable solution is the multi-armed bandit problem. It can be stated as follows. There are N independent projects, e.g., statistical populations (see Robbins 1952), gambling machines (or bandits) etc.. The state of the νth of them at time t is denoted by $x_\nu(t)$ and it belongs to a set of possible states S_ν which in this paper is assumed to be finite. Let $S_\nu = \{1, \ldots, K_\nu\}$. At each point in time one can work on one project only and if the νth of them is selected, one receives a reward $r(t) = r_x^\nu(t)$ and its state changes according to a stationary transition rule: $p_{ij}^\nu = P(x_\nu(t+1) = j \mid x_\nu(t) = i)$ while the states of all other projects remain unchanged: $x_\kappa(t+1) = x_\kappa(t)$ if $\kappa \neq \nu$. Let $x(t) = (x_1(t), \ldots, x_N(t))$ and let $\pi(t)$ denote the project selected at time t. The states of all projects are observable and the problem is to choose $\pi(t)$ as a function of $x(t)$, so as to maximize the expected total discounted reward, given an initial state $x(0)$:

$$E_\pi\left[\sum_{t=0}^{\infty} \alpha^t r(t) \mid x(0)\right].$$

This problem can be handled, in principle, by the methods of Markovian Decision Theory, see Derman (1970). However, a major difficulty in computations is the high dimension of the state space: $K_1 x \cdots x K_N$. Gittins and Jones (1974) (c.f. Gittins 1979, Whittle 1980) have shown that an optimal policy has the following form. There exist numbers $M_\nu(i)$, $k \in S_\nu$, $1 \leqslant \nu \leqslant N$, the dynamic allocation (or Gittins) indices, such that they define an optimal policy π^0 as follows: $\pi^0(x(t)) = \nu$ if and only if $M_\nu(x_\nu(t)) = \max\{M_\kappa(x_\kappa(t)), 1 \leqslant \kappa \leqslant N\}$. Furthermore the following two characterizations for $M_\nu(i)$ were given.

$$M_\nu(i) = \min\left\{ m \mid \sup_{\tau > 0} E\left(\sum_{t=0}^{\tau-1} \alpha^t r_{x_\nu}^\nu(t) + \alpha^\tau m \mid x_\nu(0) = i \right) = m \right\}, \tag{1}$$

$$M_\nu(i) = \sup_{\tau > 0} \frac{E\left(\sum_{t=0}^{\tau-1} \alpha^t r_{x_\nu}^\nu(t) \mid x_\nu(0) = i\right)}{1 - E(\alpha^\tau \mid x_\nu(0) = i)}, \tag{2}$$

where in equations (1), (2) above τ denotes a stopping time for $\{x_\nu(t), t \geqslant 0\}$.

In this paper we use (1) to show that, for any fixed ν and k, $M_\nu(k)$ can be computed by solving a single linear programming problem. Computational procedures for the

*Received March 6, 1984; revised March 14, 1985.
AMS 1980 subject classification. Primary: 90C47.
IAOR 1973 subject classification. Main: Programming: Markov Decision.
OR/MS Index 1978 subject classification. Primary: 117 Dynamic programming/Markov.
Key words. Dynamic programming, Linear programming, Markov decision chains.

0364-765X/86/1101/0180$01.25

indices when the state spaces are finite have also been developed by Beale (1979) and Varaiya *et al.* (1984). More recently Katehakis and Veinott (1985) have discovered a different interpretation for the indices which shows that standard algorithms (including linear programming for finite state spaces) of Markov Decision Theory can be used to do the computations.

2. Linear programming formulation. In this section we construct a linear program for the solution of problem (1) for any fixed ν and k. Since we consider a fixed project for notational convenience we drop the ν from r^ν, p_{ij}^ν, $x_\nu(t)$, K_ν, S_ν, $M_\nu(i)$. For α in $(0,1)$ define:

$$\phi_i(m) = \sup_{\tau > 0} E\left(\sum_{t=0}^{\tau-1} \alpha^t r_{x(t)} + \alpha^\tau m \mid x(0) = i \right). \tag{3}$$

The next lemma summarizes properties of $\phi_i(m)$, $M(i)$; proofs are given in Whittle (1982, p. 210) and Ross (1983, p. 131).

LEMMA 1. a. $\phi_i(m) = \max\{m, r_i + \alpha \sum_j p_{ij}\phi_j(m)\}$.
b. $M(i) = \min\{m \mid \phi_i(m) = m\}$.
c. *For fixed* i, $\phi_i(m)$ *is nondecreasing, convex in* m.

For fixed m let $P(m)$ denote the following linear program.

$$\textit{minimize} \quad \sum_{j\in S} u_j$$

subject to

$$\sum_{j\in S} (\delta_{ij} - \alpha P_{ij}) u_j \geq r_i, \qquad i \in S, \tag{4}$$

$$u_i \leq m, \qquad i \in S. \tag{5}$$

Let $\{u_i^0(m), i \in S\}$ be an optimal solution of $P(m)$. Following Derman (1970, pp. 114), see also Kallenberg (1983) and Hordijk and Kallenberg (1984) one can prove the following.

LEMMA 2. a. $u_i^0(m) = \phi_i(m)$.
b. *If* $\{u_i, i \in S\}$ *is any other feasible solution then* $u_i \geq u_i^0(m)$ *for all* $i \in S$.

Consider now the next linear program which we denote (P_k).

$$\textit{minimize} \quad \sum_{j\in S} y_j + Kz$$

subject to

$$(1-\alpha)z + \sum_{j\in S} (\delta_{ij} - \alpha p_{ij}) y_j \geq r_i, \qquad i \in S - \{k\}, \tag{6}$$

$$(1-\alpha)z - \alpha \sum_{j\in S} p_{kj} y_j \geq r_k, \tag{7}$$

$$y_i \geq 0,$$

$$z \text{ unrestricted}.$$

Let $\{z^0; y_i^0, i \in S\}$ be an optimal solution of (P_k). Then we have

LEMMA 3. a. $y_k^0 = 0$.
b. $z^0 \geq M(k)$.
c. *If* $z \geq M(k)$, *then* $\{z; u_i^0(z) - z, i \in S\}$ *is feasible for* (P_k) *and* $\sum_i \phi_i(z) \geq \sum_i \phi_i(z^0)$.

PROOF. a. It suffices to notice that if $\{z; y_i, i \in S\}$ is a feasible solution for (P_k) then $\{z; \hat{y}_i, i \in S\}$ is also feasible, where $\hat{y}_i = y_i$ if $i \neq k$ and $\hat{y}_k = 0$.

b. Notice that $\{y_i^0 + z^0, i \in S\}$ is a feasible solution of $P(z^0)$. Hence, by part (a) above, Lemma 2 and Lemma 1(a): $z^0 = y_k^0 + z^0 \geq u^0(z^0) = \phi_k(z^0) \geq z^0$, i.e., $z^0 = \phi_k(z^0)$. Now, from the definition of $M(k)$ it follows that $z^0 \geq M(k)$.

c. For the feasibility of (P_k), only inequality (6) is not trivial. To show that it holds it suffices to prove that

$$z - \alpha \sum_{j \in S} P_{kj} u_j(z^0) \geq r_k \tag{8}$$

holds. Now since $z \geq M(k)$ it follows from Lemma 1 that $u_k^0(z) = \phi_k(z) = z$, thus (7) is identical to (4) and therefore it holds. Furthermore,

$$\sum_i \phi_i(z) = \sum_i \left(u_i^0(z) - z\right) + Kz \geq \sum_i \left(y_i^0 + z^0\right) \geq \sum_i \phi_i(z^0)$$

where the first inequality follows since $\{z; u_i^0(z) - z, i \in S\}$ is feasible for (P_k) and the second one holds since $\{y_i^0 + z^0\}$ is feasible for $P(z^0)$.

We are now in position to prove the following:

THEOREM. $z^0 = M(k)$.

PROOF. From Lemma 3(b) we have that it suffices to show that $z^0 \leq M(k)$. Assume that $z^0 > M(k)$. Then, using Lemma 3(c) we obtain:

$$\sum_{i \neq k} \phi_i(M(k)) \geq \sum_i \phi_i(z^0) - M(k) > \sum_i \phi_i(z^0) - z^0 = \sum_{i \neq k} \phi_i(z^0)$$

and we reach to a contradiction to Lemma 1(c).

REMARKS. When we have obtained the solution to (P_k) (and thus $M(k)$) in order to compute $M(l)$, we need to replace only two constraints of (P_k). Thus one can construct an efficient sequential procedure to obtain all the indices. Even if one groups all programs (P_k), $k \in S$ in an obvious way to form a single linear program this program will contain $\sum_{\nu=1}^{N} K_\nu^2$ constraints. The linear program for the multi-armed bandit problem that can be obtained using standard Markovian Decision Theory methods will contain $N \prod_{\nu=1}^{N} K_\nu$ constraints.

Acknowledgement. The authors would like to thank an anonymous referee for his conscientious review that has greatly improved this paper.

References

Beale, E. M. L. (1979). Discussant of J. C. Gittins. 171–172.

Derman, C. (1970). *Finite State Markovian Decision Processes*. Academic Press, New York.

Gittins, J. C. (1979). Bandit Processes and Dynamic Allocation Indices. *J. Roy. Statist. Soc. Ser. B* 41 148–164.

——— and Jones, D. M. (1974). A Dynamic Allocation Index for the Sequential Design of Experiments. In J. Gani, K. Sarkadi and I. Vince (Eds.), *Progress in Statistics*, North Holland, Amsterdam, 241–266.

Hordijk, A. and Kallenberg, L. C. M. (1984). Transient Policy in Discrete Dynamic Programming; Linear Programming Including Suboptimality Tests and Additional Constraints. *Math Programming* 30 46–70.

Kallenberg, L. C. M. (1983). Linear Programming and Finite Markovian Control Problems. Mathematical Centre Tract No. 148, Mathematical Centre, Amsterdam.

Katehakis, M. N. and A. F. Veinott, Jr. (1985). The Multi-Armed Bandit Problem: Decomposition and Computation. Department of Operations Research, Stanford Univeristy, Technical Report No. 41.

Robbins, H. (1952). Some Aspects of the Sequential Design of Experiments. *Bull. Amer. Math. Monthly* 58 527–586.

Ross, S. M. (1983). *Introduction to Stochastic Dynamic Programming*. Academic Press, New York.

Whittle, P. (1980). Multi-Armed Bandits and the Gittins Index. *J. Roy. Statist. Soc. Ser. B.* **42** 143–149.
———. (1982). *Optimization over Time. Vol. 1.* John Wiley, New York.
Varaiya, P., Walrand, J. and Buyukkoc, C. (1984). Extensions of the Multi-Armed Bandit Problem. Electronic Research Laboratory, University of California, Berkeley, Technical Report, 41 pp.

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS, STATE UNIVERSITY OF NEW YORK AT STONY BROOK, STONY BROOK, NEW YORK 11794