The Annals of Statistics
1973, Vol. 1, No. 2, 334-341

ADAPTIVE POLICIES FOR MARKOV RENEWAL PROGRAMS

BY BENNETT L. FOX AND JOHN E. ROLPH

Université de Montréal and The RAND Corporation

We recast a class of denumerable-state, infinite-action Markov renewal programs with unknown parameters as one-state programs with actions corresponding to stationary policies in the original program. Under suitable conditions we find an adaptive (nonstationary) optimal policy in the sense of maximizing long-run expected reward per unit time.

1. Introduction. Finite-state, finite-action Markov renewal programs with all parameters known were defined by Jewell in [10]. We study the denumerable-state, infinite-action analog with the parameters unknown. One-step reward distributions, transition-time distributions, and transition probabilities are unknown a priori. Beginning in state $i \in S$, the decisionmaker takes action $k \in A_i$, moves to a new state with probability distribution $q(\cdot \mid i, k)$ and given that he moves to j, receives reward R_{ij}^k during a transition lasting T_{ij}^k after which he takes another action, has a transition, etc. These one-step rewards and transitions are random variables and actions are taken only at the completion of transitions. His objective is to find a policy which maximizes his expected long-run average reward. A policy $(\delta_1, \delta_2, \cdots)$ is a collection of functions such that the nth action at state i is $\delta_n(i)$ where δ_n may depend on the history of the process prior to n. A stationary policy δ is of the form $(\delta, \delta, \cdots)$ where now δ depends only on the current state and thus cannot be history-remembering. Define Δ to be the set of all stationary policies. Making certain assumptions, we construct a nonstationary adaptive policy which does as well as any stationary policy in maximizing expected reward per unit time no matter what values the unknown parameters have.

Thus, using the average reward rate criterion, our policy is optimal whenever a stationary optimal or stationary ε-optimal policy exists. Lippman [11] shows the existence of a stationary ε-optimal policy under essentially our assumptions, although in general a stationary optimal policy need not exist. Similar results are unattainable in the general multichain case because there is no way to be sure of optimizing the action in a transient state when the action determines which absorbing chain will be entered. A stationary optimal policy exists in the finite-state, finite-action case (Fox [5]), and sufficient conditions for existence have been given (Fox [6]) in the finite-state, infinite-action case.

Mallows and Robbins [13] give results analogous to ours in the discrete-time, one-state case. Part of our argument is an adaptation of theirs. Baños [1] and Shubert [17] treat similar problems from game theoretic and statistical decision theoretic viewpoints, respectively.

Received October 22, 1971; revised June 15, 1972.

2. The maximizing policy. Suppose a stationary policy $\delta \in \Delta$ is always used. For any fixed path let

$$R^{\delta}(t) = \text{total reward received up to time } t \text{ and } \bar{R}^{\delta}(t) = \frac{1}{t} R^{\delta}(t) .$$

The strong law of large numbers together with a standard renewal theory argument implies that for any $\delta \in \Delta$, $\lim_{t\to\infty} \bar{R}^{\delta}(t)$ exists and is a constant with probability 1. Since we later show that $\bar{R}^{\delta}(t)$ is uniformly integrable, the expected gain rate associated with δ exists and is defined as $g^{\delta} = \lim_{t\to\infty} E[\bar{R}^{\delta}(t)]$. Let $g^* = \sup_{\delta\in\Delta} g^{\delta}$. Using the adaptive policy defined later for any fixed path, let $R(t) =$ total reward received up to time t, and $\bar{R}(t) = t^{-1}R(t)$. Under the assumptions given below we show that the rewards from the adaptive policy satisfy

$$P[\lim_{t\to\infty} \bar{R}(t) = g^*] = 1 . \tag{1}$$

It will then follow that

$$\lim_{t\to\infty} E[\bar{R}(t)] = g^* . \tag{2}$$

ASSUMPTIONS.

1. There is an a priori known countable set of stationary policies $\Lambda \subset \Delta$ such that

$$\sup\{g^{\lambda} : \lambda \in \Lambda\} = g^* .$$

2. For each state, the expectation and the variance of the time and reward until state 1 is next reached, is uniformly bounded over Λ. Call this bound H.

3. For each state, the expected time to return to state 1 is uniformly bounded away from 0 over Λ.

In the denumerable state-denumerable action case, Assumptions 2 and 3 on Δ imply Assumption 1. It is shown in [16] that we can take Λ to be the union of sets of policies of the form $(\delta \mid \delta(i) = a_i,\ 1 \leqq i \leqq n)$ for some n and some fixed sequence of actions $a_1, \cdots, a_n$ with $\delta(i)$ arbitrary for $i > n$. An analysis of the structure of a particular problem with a given parameter set may establish the form of an optimal policy and thus yield an appropriate Λ directly. In lieu of Assumptions 2 and 3, conditions on the transition matrices, one-step reward distributions and one-step time distributions can be given which may be easier to check for some applications. The age replacement application given later is an example with an uncountable action set where Assumption 1 is easily checked.

To ensure that our adaptive policy will concentrate with probability 1 on the high gain rate stationary policies, only policies in Λ are tried and each policy in Λ is tried infinitely often. The basic unit used in defining our policy is the state 1-to-state 1 cycle. Within each state 1-to-state 1 cycle the same stationary policy is used.

Beginning in the initial state some fixed stationary policy is applied until state 1 is reached. From this point forward we have a one-state problem since policy choices are only made at state 1. Following Mallows and Robbins [13] our

strategy specifies a sequence of positive integers which number the forced-choice cycles in which predetermined stationary policies are applied. Let $s_{11}, s_{12}, \cdots$ be any increasing sequence of positive integers with $s_{11} = 1$. Let $s_{21}, s_{22}, \cdots$ be a second sequence disjoint from the first with $s_{31}, s_{32}, \cdots$ being a third sequence disjoint from the first two, and so on. If for the nth cycle $n = s_{\delta l}$, for some δ and l we use stationary policy δ for the nth cycle; otherwise we choose the stationary policy with the leading observed reward rate. The observed reward rate for policy δ at time t is given by

$$\bar{R}^{\delta}(t) = \frac{R^{\delta}(t)}{B^{\delta}(t)},$$

where $R^{\delta}(t)$ is the total reward received and $B^{\delta}(t)$ is the total time spent prior to t while δ was applied. (In the case when a stationary policy δ is always used, $B^{\delta}(t) = t$ and $\bar{R}^{\delta}(t) = t^{-1}R^{\delta}(t)$ as in the beginning of the section.) The relationship is defined only for those δ which have been applied in a forced-choice cycle prior to t. Let $s(n)$ be the total number of forced choices up to the nth cycle, i.e., the number of integers $s_{\delta l}$ which are $\leqq n$. We choose the $s_{\delta l}$ so that $s(n) \doteq \log n$.

In sampling policies directly rather than actions, we do not fully utilize all information since each action is associated with many different policies. We do not know a general remedy, but in Section 4 we give a modified policy for the finite case that uses information more efficiently.

3. Proof of optimality. We first show that taking the limit as $t \to \infty$ is the same as taking the limit as the number of state 1-to-state 1 cycles become infinite. By Assumption 2, we can neglect the contributions of the time and reward before reaching state 1 for the first time to the overall time and overall reward. Indexing consecutive cycles by m and excluding the time and rewards before reaching state 1 for the first time we define

$$\begin{aligned} V_i &= \text{time to complete } i\text{th cycle}, \\ V_i^{\delta} &= \text{time to complete } i\text{th cycle if policy } \delta \text{ is used}, \\ &= 0 \qquad \text{otherwise}. \end{aligned}$$

$$B(m) = \sum_{i=1}^{m} V_i, \qquad B^{\delta}(m) = \sum_{i=1}^{m} V_i^{\delta}.$$

Similarly for the reward sequence define U_i, U_i^{δ} and abusing notation $R(m)$, $R^{\delta}(m)$. (Strictly speaking we should use $R(B(m))$ and $R^{\delta}(B(m))$ to be consistent with $R(t)$ and $R^{\delta}(t)$ notation.)

LEMMA.

(a) $$P\left(\limsup_{m\to\infty} \frac{B(m)}{m} < \infty\right) = 1$$

(b) $$P\left(\liminf_{m\to\infty} \frac{B(m)}{m} > 0\right) = 1.$$

Analogous relations hold with B replaced by R.

PROOF. We can take a space Ω on which the following σ-fields are well defined:

$$\mathscr{F}_j = \sigma\{D_i, V_i, U_i; i = 1, 2, \cdots, j\}$$

and

$$\mathscr{F} = \bigvee_{i=1}^{\infty} \mathscr{F}_i ,$$

the smallest σ-field generated by all the $\mathscr{F}_i$'s, where D_i is the stationary policy used in the ith cycle and, $\sigma\{\cdot\}$ means the smallest σ-field with respect to which $\{\cdot\}$ is measurable. The probability measure P is induced on $\mathscr{F}$ by the distributions of the one-step rewards and transition times and the one-step transition probabilities given by the adaptive policy specified in the preceding section. Thus $E(V_i \mid \mathscr{F}_{i-1})$ is the expected time to complete the ith cycle given the previous history of the process, while $E(V_i \mid D_i)$ is the expected time to complete the ith cycle given the value of D_i, the stationary policy used. Note that $\mathscr{F} \supset \mathscr{F}_j \supset \mathscr{F}_{j-1}$, D_j is measurable over $\mathscr{F}_{j-1}$, and

$$E(V_j \mid \mathscr{F}_{j-1}) = E(V_i \mid D_j) .$$

Let

$$X_i = \sum_{j=1}^{i} (V_j - E(V_j \mid D_j)) .$$

Then $\{X_i, \mathscr{F}_i, i = 1, 2, \cdots\}$ is a martingale relative to the $\mathscr{F}_i$ (see Doob [3]) since $E(X_i \mid \mathscr{F}_{i-1}) = X_{i-1}$ with probability 1. Now

$$\frac{B(m)}{m} = X_m/m + 1/m \sum_{i=1}^{m} E(V_i \mid D_i) .$$

By Assumptions 2 and 3, $E(V_i \mid D_i)$ is bounded away from 0 and ∞. Since

$$\sum_{i=1}^{\infty} (1/i^2) E(V_i - E(V_i \mid D_i))^2 \leqq H \sum_{i=1}^{\infty} 1/i^2 < \infty ,$$

a standard martingale convergence theorem (e.g., Feller [4], page 238, Theorem 2) implies that $X_m/m \to 0$ w.p. 1 as $m \to \infty$, completing the proof. □

We first show that we can neglect the contribution to the average reward at time t (for large t) of the forced choice cycles and of the last (prior to t) cycle in which a particular policy was freely chosen. To see that the contribution of the forced choice cycles to the overall reward and the overall time can be neglected, note that in view of Assumption 2, for any $\delta \in \Lambda$

$$\limsup_{n\to\infty} \left| \frac{1}{N^{\delta}(n)} \sum_{i=1}^{n} U_i^{\delta} \right| \leqq H \quad \text{with probability 1,} \tag{3}$$

$$\limsup_{n\to\infty} \left(\frac{1}{N^{\delta}(n)} \sum_{i=1}^{n} V_i^{\delta} \right) \leqq H \quad \text{with probability 1,}$$

where $N^{\delta}(n)$ = the number of the first n cycles which use policy δ. Now observe that $s(n)/n \to 0$ and apply the lemma.

Define the last free choice cycle prior to cycle n for a policy δ as the last cycle, if any, for which policy δ was chosen as the leader. At the nth cycle, define $p(n)$ to be the number of different policies used prior to the nth cycle. Let $r(n)$ be the number of different policies used on free choices prior to the nth cycle.

Since $r(n) \leqq p(n) \leqq s(n)$, the above argument implies that the contribution of the last free choice cycles to the overall reward and overall time can be neglected.

Fix a policy $\sigma \in \Lambda$, then

$$\bar{R}^{\sigma}(m) = \frac{R^{\sigma}(m)}{B^{\sigma}(m)} = \frac{R^{\sigma}(m)/N^{\sigma}(m)}{B^{\sigma}(m)/N^{\sigma}(m)} .$$

By (3), both numerator and denominator have expectations. An application of the SLLN to the numerator and denominator separately yields that both converge to a constant w.p. 1 as $N^{\sigma}(m) \to \infty$. Since the union of the two null sets where convergence does not occur is a null set and $N^{\sigma}(m) \to \infty$ as $m \to \infty$, application of (3) and the lemma yields the value of the limit as:

$$\lim_{m\to\infty} \bar{R}^{\sigma}(m) = g^{\sigma} \quad \text{w.p. 1} . \tag{4}$$

This would be a special case of a result of Pyke and Schaufele ([15] Theorem 5.1), except that in their paper there is only one decision available in each state. From (4) for, $\varepsilon_1, \varepsilon_2 > 0$, there exists an m_0 such that

$$P\{\bar{R}^{\sigma}(m) \geqq g^{\sigma} - \varepsilon_2 \text{ for all } m > m_0\} \geqq 1 - \varepsilon_1 .$$

Let Γ be the set of policies used in free choices for $m \geqq m_0$. For $\gamma \in \Gamma$ and $m \geqq m_0$, define $l^{\gamma}(m)$ as the largest cycle index $\leqq m$ where γ was freely chosen. From the definitions,

$$\bar{R}^{\delta}(l^{\delta}(m) - 1) \geqq \bar{R}^{\sigma}(l^{\delta}(m) - 1) ,$$

for all $\sigma \in \Lambda$.

Set $\bar{R}(m) = R(m)/B(m)$.

By earlier arguments and neglecting the contributions from policies not used after cycle m_0,

$$\bar{R}(m) = \textstyle\sum_{\gamma\in\Gamma} R^{\gamma}(l^{\gamma}(m) - 1)/B(m) + o_p(1) ,$$

where the last term goes to 0 w.p. 1 as $m \to \infty$. Thus,

$$\begin{aligned} \bar{R}(m) &\geqq \textstyle\sum_{\gamma\in\Gamma} \bar{R}^{\sigma}(l^{\gamma}(m) - 1)B^{\gamma}(l^{\gamma}(m) - 1)/B(m) + o_p(1) \\ &\geqq (g^{\sigma} - \varepsilon_2)[\textstyle\sum_{\gamma\in\Gamma} B^{\gamma}(l^{\gamma}(m) - 1)/B(m)] + o_p(1) \\ &\geqq g^{\sigma} - \varepsilon_2 + o_p(1) , \end{aligned}$$

since the term in square brackets goes to 1 w.p. 1 as $m \to \infty$, and as ε_2 was arbitrary,

$$\liminf_{m\to\infty} \bar{R}(m) \geqq g^{\sigma} \quad \text{w.p. 1} .$$

Since σ was arbitrary and a denumerable union of null sets is null,

$$P(\liminf_{m\to\infty} \bar{R}(m) \geqq g^{*}) = 1 . \tag{5}$$

Now $\bar{R}^{\delta}(m) \to g^{\delta}$ w.p. 1 as $m \to \infty$ so that

$$P(\limsup_{m\to\infty} \bar{R}(m) \leqq g^{*}) = 1 , \tag{6}$$

since $\bar{R}(m)$ is a weighted average of the $\bar{R}^{\delta}(m)$. Thus we have proved

$$P(\lim_{m\to\infty} \bar{R}(m) = g^{*}) = 1 . \tag{7}$$

From the lemma, taking limits as $t \to \infty$ is the same as taking limits as $m \to \infty$ so that (7) implies (1).

Turning to the proof of (2), we let P be a measure on the reward sequence corresponding to our policy. By the lemma (stated for R as well as B), there is an $M < \infty$ such that for any $a > 0$,

$$\int_{|\bar{R}(t)|>a} |\bar{R}(t)|\, dP \leqq a \int_{|\bar{R}(t)|>a} [\bar{R}(t)/a]^2\, dP \leqq M/a \to 0 \qquad \text{as} \quad a \to \infty .$$

Thus, the random variables $\{\bar{R}(t)\}$ are uniformly integrable so (1) implies (2) by a theorem in Loève [12], page 163. (Similar reasoning shows that $R^{\delta}(t)$ is uniformly integrable.) We have proved the following

THEOREM. *Under Assumptions* 1, 2, *and* 3, *the average rewards from the above strategy satisfy*

(i) $P\{\lim_{t\to\infty} t^{-1}R(t) = g^*\} = 1$
(ii) $\lim_{t\to\infty} t^{-1}E[R(t)] = g^*$.

4. Remarks. In the finite-state, finite-action case we could sample actions on each transition rather than policies on each cycle. By sampling actions on forced-choice transitions, we can obtain consistent estimates of the parameters. One choice is the natural empirical estimates, which are consistent (Moore and Pyke [14]). On the free-choice transitions, we follow the leader obtained by substituting these estimates for the unknown parameters in the gain rate formula for stationary policies. The estimated optimal policy can be calculated using a policy improvement routine or linear program (Fox [5], Denardo and Fox [2]). The proof that g^* is attained is different from the foregoing one since we do not reduce the problem to one state, but the number of policies being finite leads to some simplifications. Intuition indicates a faster convergence rate for sampling actions directly rather than just sampling policies.

We can constract an adaptive policy for infinite-state problems based on sampling actions rather than policies if (i) we are able to calculate an optimal policy knowing all parameters, (ii) we can obtain *uniformly* consistent estimates of these parameters by sampling actions in forced choices, (iii) the gain rates are continuous functions of these parameters, and (iv) the reward and time on the forced choices can be neglected. This adaptive policy would probably converge faster than the one described in Section 2. For example, the conditions may be satisfied in some problems where the mean one-step rewards, transition probabilities, and mean transition times depend on only a finite number of unknown parameters such as the arrival rate in queueing problems. This is an open problem.

5. Examples. Many problems studied in the literature satisfy our assumptions. Typical applications include:

(i) Replacement problems where we return to state 1 (replace the item) whenever the state (observable deterioration) exceeds a certain level, to be determined.

(ii) Queueing problems where we activate the server whenever the queue

length exceeds a certain level, to be determined. Heyman [8] and others give conditions under which a policy of this form is optimal for the M/G/1 queue. To satisfy our assumptions, we rule out policies that do not activate the server when the queue length exceeds a given (large) number.

(iii) Inventory problems where we determine a reorder point and a reorder level. See Iglehart [9] or for a general reference [7].

(iv) The "streetwalker's dilemma", where the server must decide whether to accept a given proposition or wait for a more desirable one. Lippman [11] gives simple conditions under which our assumptions hold and the optimal policy has the form: accept an offer if and only if the ratio of expected reward to expected service time exceeds a certain number.

(v) Age replacement. We replace an item at a scheduled time or at failure, whichever occurs first, with respective costs c_1 and c_2 $(0 < c_1 < c_2)$. This is a one-state problem with actions selected from $[b, \infty]$. Here b is a small positive number and ∞ corresponds to no planned replacement. Assuming that the failure distribution is continuous, we could take Λ equal to a set of rationals dense in $[b, \infty]$.

6. Acknowledgments. We are indebted to Colin Mallows and Ralph Strauch, whose careful reading of earlier versions of this paper led to substantial improvements. Comments by the referee and associate editor markedly improved the exposition.

REFERENCES

[1] BAÑOS, A. (1968). On pseudo-games. *Ann. Math. Statist.* **39** 1932–1945.

[2] DENARDO, E. V. and FOX, B. L. (1968). Multichain Markov renewal programs. *SIAM J. Appl. Math.* **16** 468–487.

[3] DOOB, J. L. (1952). *Stochastic Processes*. Wiley, New York.

[4] FELLER, W. (1966). *An Introduction to the Theory of Probability and Its Applications*, **2**. Wiley, New York.

[5] FOX, B. L. (1966). Markov renewal programming by linear fractional programming. *SIAM J. Appl. Math.* **14** 1418–1430.

[6] FOX, B. L. (1967). Existence of stationary optimal policies for some Markov renewal programs. *SIAM Rev.* **9** 573–576.

[7] HADLEY, G. and WHITIN, T. M. (1963). *Analysis of Inventory Systems*. Prentice Hall, Englewood Cliffs.

[8] HEYMAN, D. P. (1968). Optimal operating policies for M/G/1 queuing systems. *Operations Res.* **16** 362–382.

[9] IGLEHART, D. L. (1963). Optimality of (s, S) policies in the infinite horizon inventory problem. *Management Sci.* **9** 259–267.

[10] JEWELL, W. S. (1963). Markov renewal programs, I and II. *Operations Res.* **11** 938–971.

[11] LIPPMAN, S. A. (1970). Maximal average reward policies for a class of semi-Markov decision processes with arbitrary state and action space. Western Management Science Institute Paper 162, Univ. of California, Los Angeles. Also in *Ann. Math. Statist.* **42** 1717–1726.

[12] LOÈVE, M. (1963). *Probability Theory* (3rd ed.). Van Nostrand, Princeton.

[13] MALLOWS, C. L. and ROBBINS, H. (1964). Some problems of optimal sampling strategy. *J. Math. Anal. Appl.* **8** 90–103.

[14] MOORE, E. H. and PYKE, R. (1968). Estimation of the transition distributions of a Markov renewal process. *Ann. Inst. Statist. Math.* **20** 90–103.

[15] PYKE, R. and SCHAUFELE, R. (1964). Limit theorems for Markov renewal processes. *Ann. Math. Statist.* **37** 1746–1764.

[16] ROLPH, J. E. and STRAUCH, R. E. (1972). A countable policy set for sequential decision problems. *Ann. Math. Statist.* **43** 2079–2083.

[17] SHUBERT, B. (1969). Bayesian model of decision-making as a result of learning from experience. *Ann. Math. Statist.* **40** 2127–2142.

THE RAND CORPORATION
1700 MAIN STREET
SANTA MONICA, CALIFORNIA 90406