*The Annals of Mathematical Statistics*
1972, Vol. 43, No. 6, 2078-2082

# A COUNTABLE POLICY SET FOR SEQUENTIAL DECISION PROBLEMS

BY JOHN E. ROLPH AND RALPH E. STRAUCH

*The RAND Corporation*

In denumerable state, denumerable action sequential decision problems in which the reward function has uniformly bounded 2nd moment, the optimal reward for the decisionmaker who restricts himself to the countable set of stationary policies consisting of those which choose some arbitrary action at all but a finite number of states will be the same as the optimal reward for the decisionmaker who optimizes over all stationary policies. Under some further restriction, he can do almost as well simply by solving a large finite state truncation of the original problem.

**1. Introduction.** In denumerable state, denumerable action sequential decision problems, the number of possible stationary policies is uncountably infinite. We show (Theorem 2) that so long as the reward function being optimized has a uniformly bounded 2nd moment, the optimal reward for the decisionmaker who restricts himself to the countable set of stationary policies consisting of those which choose some action arbitrarily at all but a finite number of states will be the same as the optimal reward for the decisionmaker who considers all possible stationary policies. (The term "optimal," as used herein, refers to optimality among all *stationary* policies. This is not necessarily the same as optimality among all policies, including non-stationary policies. See, for instance, page 58 of [3], example 3.9.2.)

We also show (Theorem 3) that under some further restriction, the optimal reward will be the limit of the optimal rewards in the finite state truncations of the countable state problem. Thus, the decisionmaker can get close to the optimal reward by solving a large finite state truncation of the original problem. Since this procedure is frequently followed in practice, this result may give practitioners who worry about such things more peace of mind. Fox ([4] and [5]) gives results in the same vein for some dynamic programming problems.

**2. Results.** Let $S = (1, 2, \cdots)$ be the states, let $A = (1, 2, \cdots)$ be the actions, and let $q(\cdot \mid s, a)$ be the transition probabilities to the new states for a sequential decision problem. A stationary policy $\delta$ is a function from $S$ to $A$. Let $\Delta$ be the set of stationary policies. The cylinder topology on $\Delta$ is the topology generated by sets of the form $(\delta \mid \delta(i) = a_i, 1 \leqq i \leqq m)$ for some $m$ and some fixed sequence of actions $a_1, \cdots, a_m$.

In what follows, we assume that the initial state of the process, $s_1$, is fixed. Our results are thus valid for each initial state individually, rather than uniformly across states. Functional dependence on the initial state is suppressed.

Received February 22, 1972; revised March 8, 1972.

Suppose there exists some event $B$ which occurs at some finite stage of the process with probability one for all $\delta \in \Delta$. (Examples are: first return to some fixed state in a recurrent chain, occurrence of stage $N$ in the process, reaching a terminal state, etc.) Let $h$ denote the path of the process (i.e., the sequence of states visited and actions taken) beginning at state $s_1$ and continuing until event $B$ occurs. Let $H$ denote the set of possible paths $h$, and let $P_\delta$ denote the probability induced on $H$ by $\delta$. For any function $v$ defined on $H$, let $E_\delta(v)$ denote the expectation of $v$ under $P_\delta$. Think of $v(h)$ as the reward accumulated along the path $h$. Examples are given in Section 3.

THEOREM 1. *If $E_\delta(v^2)$ is uniformly bounded for all $\delta \in \Delta$, then $E_\delta(v)$ is a continuous function of $\delta$ in the cylinder topology on $\Delta$.*

PROOF. Assume $v \geqq 0$, the general case follows by considering the positive and negative parts of $v$ separately.

Let $||h||$ be the largest state reached by $h$. Let $p_1(\delta) = P_\delta(||h|| = i)$ and let $V_i(\delta) = E_\delta(v \mid ||h|| = i)$. We need the following lemma.

LEMMA 1. *If $\delta$ and $\sigma$ are stationary policies such that $\delta(i) = \sigma(i)$ for $1 \leqq i \leqq m$ for some $m \geqq s_1$, then $p_i(\delta) = p_i(\sigma)$ and $V_i(\delta) = V_i(\sigma)$ for $1 \leqq i \leqq m$.*

PROOF. So long as state $m$ is not exceeded, the probability laws governing the process under policies $\delta$ and $\sigma$ are identical. □

Now let $\delta_m \to \delta$ in the cylinder topology. Without loss of generality, assume that $\delta_m(i) = \delta(i)$ for $1 \leqq i \leqq m$, and that $\lim_{m\to\infty} E_{\delta_m}(v)$ exists. (Choose a subsequence to achieve this, if necessary. This convention avoids having to keep track of an additional index.)

By the lemma, for all $m > s_1$,

$$\sum_{i=1}^{m} p_i(\delta_m)V_i(\delta_m) = \sum_{i=1}^{m} p_i(\delta)V_i(\delta) \,. \tag{1}$$

Hence, for any $r > s_1$ and all $m \geqq r$

$$E_{\delta_m}(v) = \sum_{i=1}^{\infty} p_i(\delta_m)V_i(\delta_m) \geqq \sum_{i=1}^{r} p_i(\delta)V_i(\delta) \,.$$

Hence,

$$\lim_{m\to\infty} E_{\delta_m}(v) \geqq \sum_{i=1}^{\infty} p_i(\delta)V_i(\delta) = E_\delta(v) \,.$$

Now assume $\lim_{m\to\infty} E_{\delta_m}(v) > E_\delta(v)$. Then there exists $M > s_1$, and $\varepsilon > 0$ such that for all $m > M$, $E_{\delta_m}(v) > E_\delta(v) + \varepsilon$. Thus, by (1), for $m > M$

$$P_{\delta_m}(||h|| > m)E_{\delta_m}(v \mid ||h|| > m) = \sum_{i=m+1}^{\infty} p_i(\delta_m)V_i(\delta_m) > \varepsilon \,. \tag{2}$$

But by Lemma 1, $P_{\delta_m}(||h|| > m) = P_\delta(||h|| > m) \to 0$ as $m \to \infty$. Hence $E_{\delta_m}(v \mid ||h|| > m) \to \infty$. From (2), then

$$P_{\delta_m}(||h|| > m)E^2_{\delta_m}(v \mid ||h|| > m) \to \infty \,.$$

But

$$\begin{aligned} P_{\delta_m}(||h|| > m)E^2_{\delta_m}(v \mid ||h|| > m) &\leqq P_{\delta_m}(||h|| > m)E_{\delta_m}(v^2 \mid ||h|| > m) \\ &\geqq E_{\delta_m}(v^2) \,. \end{aligned}$$

But the $E_{\delta_m}(v^2)$ are uniformly bounded. This is a contradiction; hence $\lim_{m\to\infty} E_{\delta_m}(v) = E_\delta(v)$. □

COROLLARY 1. *If $v$ is bounded, then $E_\delta(v)$ is a continuous function of $\delta$ in the cylinder topology on $\Delta$ regardless of the transition probabilities $q$.*

PROOF. If $v$ is bounded, $E_\delta(v^2)$ is uniformly bounded regardless of $q$. □

COROLLARY 2. *If $A$ is finite and $E_\delta(v^2)$ uniformly bounded, then there exists a $\delta^*$ such that*

$$E_\delta{}^*(v) = \sup_{\delta\in\Delta} E_\delta(v) .$$

PROOF. The cylinder topology on $\Delta$ is the product topology corresponding to the discrete topology on $A$. By Tychonoff's theorem, if $A$ is finite, this topology is compact, hence $E_\delta(v)$ achieves its maximum. □

For $m \geqq 1$, let $\Delta_m = (\delta \mid \delta(i) = 1 \text{ for } i > m)$. Think of $\Delta_m$ as policies defined only on the first $m$ states—the choice of action 1 thereafter being an arbitrary one. Let $\Delta_0 = \bigcup_{m=1}^\infty \Delta_m$. $\Delta_0$ is a countable dense subset of $\Delta$ in the cylinder topology. Our main result is

THEOREM 2. *If $E_\delta(v^2)$ is uniformly bounded in $\delta$, then*

$$\sup_{\delta\in\Delta} E_\delta(v) = \sup_{\delta\in\Delta_0} E_\delta(v) .$$

The theorem is an immediate corollary of Theorem 1. Thus, if the criterion for optimization in the sequential decision problem may be placed in the form "maximize $E_\delta(v)$", with $E_\delta(v^2)$ uniformly bounded in $v$, then optimization may be restricted to $\Delta_0$. Examples are given in Section 3.

THEOREM 3. *If $E_\delta(v^2)$ is uniformly bounded and* $\lim_{m\to\infty} P_\delta(||h|| > m) = 0$ *uniformly in $\delta$, then*

$$\sup_{\delta\in\Delta} E_\delta v = \lim_{m\to\infty} \sup_{\delta\in\Delta_m} E_\delta(v \mid ||h|| \leqq m) . \tag{3}$$

Thus, if the condition of Theorem 3 holds, the optimal return in the original problem is the limit of the optimal returns in the finite state problems obtained by considering only the first $m$ states of $S$, with normalized transition probabilities of $q(t \mid s, a)/\sum_{i=1}^m q(i \mid s, a)$ for $s$, $t \leqq m$. Thus, for $m$ sufficiently large, $\varepsilon/2$-optimal stationary policies in the finite state problem will be $\varepsilon$-optimal in the original problem.

PROOF OF THEOREM 3. By Lemma 1, $E_\delta(v \mid ||h|| \leqq m)$ depends only on $\delta(1), \cdots, \delta(m)$, hence

$$\sup_{\delta\in\Delta_m} E_\delta(v \mid ||h|| \leqq m) = \sup_{\delta\in\Delta} E_\delta(v \mid ||h|| \leqq m) .$$

It is thus sufficient to prove that

$$\lim_{m\to\infty} \sup_{\delta\in\Delta} E_\delta(v \mid ||h|| \leqq m) = \sup_{\delta\in\Delta} E_\delta(v) . \tag{4}$$

For any $\delta \in \Delta$,

$$E_\delta(v) = P_\delta(||h|| \leqq m) E_\delta(v \mid ||h|| \leqq m) + P_\delta(||h|| > m) E_\delta(v \mid ||h|| > m) .$$

As $m \to \infty$, $P_\delta(||h|| \leqq m) \to 1$, and the argument used in the proof of Theorem 1 shows that $P_\delta(||h|| > m)E_\delta(v \mid ||h|| > m) \to 0$. Hence, $E_\delta(v \mid ||h|| \leqq m) \to E_\delta(v)$. If $P_\delta(||h|| > m) \to 0$ uniformly in $\delta$, this convergence will also be uniform, and (4) follows. □

The condition $P_\delta(||h|| > m) \to 0$ uniformly in $\delta$ is sufficient to yield (3) but not necessary. To see that some condition beyond $E_\delta(v^2)$ uniformly bounded is necessary, however, consider the following.

EXAMPLE 1. At state 1, pick any state $i > 1$ and move there with probability one. At state $i$, if $i$ is even move to state 1 or state $i + 1$ with probability $\frac{1}{2}$ each, pay nothing. If $i$ is odd, move to state 1 which probability one, pay \$ 1. Let $v$ be the total received in a state 1 to state 1 cycle. Thus, for $m$ even,

$$\sup\nolimits_{\delta \in \Delta_m'} E_\delta(v \mid ||h|| \leqq m) = 0 > \sup\nolimits_{\delta \in \Delta} E_\delta v = -\tfrac{1}{2}\,,$$

where $\Delta_m' = \{\delta \mid \delta \in \Delta_m,\ P_\delta(||h|| \leqq m) > 0\}$. Note that there exist $\delta \in \Delta_m$ for which $P_\delta(||h|| \leqq m) = 0$ so that $E_\delta(v \mid ||h|| \leqq m)$ is undefined.
Note that in this case $\lim_{m \to \infty} \sup_{\delta \in \Delta_m'} E_\delta(v \mid ||h|| \leqq m)$ does not exist. This example shows that Theorem 2 can fail in discounted reward, total negative reward, or average reward problems. (In the total negative reward case, move to a terminal state instead of back to 1.) It can also fail for problems with positive reward, as shown in Example 2.

EXAMPLE 2. Same transition structure as Example 1. Move from an even state 1 and get \$ 1. Move from an odd state and get nothing. Then, for $m$ even

$$\sup\nolimits_{\delta \in \Delta_m'} E_\delta(v \mid ||h|| \leqq m) = 1 > \sup\nolimits_{\delta \in \Delta} E_\delta v = \tfrac{1}{2}\,.$$

**3. Applications.** The results given above are applicable to the following types of sequential decision problems.

a. *Terminating problems*, in which the decision problem terminates after some finite (possibly stochastic) length of time. This includes search problems, many learning models, leavable gambling houses [3], and all finite stage problems. Let the event $B$ be the event of termination, and let $v$ be the total reward or cost, terminal utility, etc.

b. *Discounted and positive bounded infinite stage problems.* Let $v$ be the total discounted reward or total reward. Theorems 2 and 3 as stated above are not directly applicable, since there is no satisfactory event $B$. However, they can be easily extended to cover these cases using the fact that the optimal finite stage rewards converge (uniformly in the discounted case and monotonically in the positive case) to the optimal infinite stage reward ([1] and [2]). Theorems 2 and 3 hold for each finite stage, and remain valid under passage to the limit. The equivalent of Theorem 3 for the discounted case is given by Fox in [5]. Corollary 2 does not extend in the same way, and in fact, is false in the positive case. See [2] for a counterexample.

c. *Average reward recurrent chains* (including renewal problems). Assume some state, say state 1, is reached from every state with probability one. Start at state 1 and let the event $B$ be the return to state 1. Let $v$ be the (time) average reward on a state 1 to state 1 cycle. For continuous time problems one must assume that the means and variances of the times between transitions are bounded above and the means of these times are bounded away from zero so that the expected time average reward coincides with the overall time average reward. See [6] for details.

## REFERENCES

[1] Blackwell, David (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

[2] Blackwell, David (1967). Positive dynamic programming *Proc. Fifth Berkely Symp. Math. Statist. Prob.* **1** Univ. of California Press.

[3] Dubins, Lester E. and Savage, Leonard J. (1965). *How to Gamble if You Must.* McGraw-Hill, New York.

[4] Fox, B. L. (1972). Discretizing dynamic programs. To appear in *J. Optimization Theory Appl.*

[5] Fox, B. L. (1971). Finite-state approximations to denumerable-state dynamic programs. *J. Math. Anal. Appl.* **34** 665–670.

[6] Fox, Bennett L. and Rolph, John E. (1973). Adaptive policies for Markov renewal programs. To appear in *Ann. Statist.* **1**.

The Rand Corporation
1700 Main Street
Santa Monica, California 90406