The Annals of Mathematical Statistics
1968, Vol. 39, No. 6, 2118–2122

ARBITRARY STATE MARKOVIAN DECISION PROCESSES[1]

BY SHELDON M. ROSS[2]

Stanford University

1. Introduction. We are concerned with a process which is observed at times $t = 0, 1, 2, \cdots$ and classified into one of a possible number of states. We let $\mathfrak{X}$ denote the state space of the process. $\mathfrak{X}$ is assumed to be a Borel subset of a complete separable metric space, and we let $\mathfrak{B}$ be the σ-algebra of Borel subsets of $\mathfrak{X}$. After each classification an action must be chosen and we let A, assumed finite, denote the set of all possible actions.

Let $\{X_t ; t = 0, 1, 2, \cdots\}$ and $\{\Delta_t ; t = 0, 1, 2, \cdots\}$ denote the sequence of states and actions; and let $S_{t-1} = (X_0 , \Delta_0 , \cdots , X_{t-1} , \Delta_{t-1})$. It is assumed that for every $x \in \mathfrak{X}$, $k \in A$ there is a known probability measure $P(\cdot \mid x, k)$ on $\mathfrak{B}$ such that, for some version, $P\{X_{t+1} \in B \mid X_t = x, \Delta_t = k, S_{t-1}\} = P(B \mid x, k)$ for every $B \in \mathfrak{B}$, and all histories S_{t-1}. It is also assumed that for every $k \in A$, $B \in \mathfrak{B}$, $P(B \mid \cdot, k)$ is a Baire function on $\mathfrak{X}$.

Whenever the process is in state x and action k is chosen then a bounded (expected) cost $C(x, k)$—assumed, for fixed k, to be a Baire function in x—is incurred.

A policy R is a set of Baire functions $\{D_k(S_{t-1} , x)\}_{k \in A}$ satisfying $D_k(S_{t-1} , x) \geqq 0$ for all $k \in A$, and $\sum_{k \in A} D_k(S_{t-1} , x) = 1$ for every (S_{t-1} , x). The interpretation being: if at time t the history S_{t-1} has been observed and $X_t = x$ then action k is chosen with probability $D_k(S_{t-1} , x)$. R is said to be stationary if $D_k(S_{t-1} , x) = D_k(x)$ for every S_{t-1} ; R is said to be stationary deterministic if $D_k(x)$ equals 0 or 1 for all x, k.

For any policy R, let

$$\varphi(x, R) = \limsup_{n \to \infty} (n + 1)^{-1} \sum_{t=0}^{n} E_R[C(X_t , \Delta_t) \mid X_0 = x].$$

Thus $\varphi(x, R)$ is the expected average cost per unit time when the process starts in state x and policy R is used.

In [4], under the assumption that $\mathfrak{X}$ is denumerable, a number of results dealing with the average cost criterion were proven. The method employed was to treat the average cost problem as a limit (as the discount factor approaches unity) of the discounted cost problem. In this paper we generalize some of these results to arbitrary state spaces.

2. Stationary deterministic optimal policies. The following theorem was originally proven by Derman [2] for the special case that $\mathfrak{X}$ is denumerable. The following proof is new; it makes use of a technique used by Taylor [5].

THEOREM 1. *If there exists a bounded Baire function $f(x)$, $x \in \mathfrak{X}$ and a constant g,*

Received 27 November 1968.

[1] This work was supported in part by the Army, Navy, Air Force and NASA, under contract Nonr 225(53)(NR-042-002) with the Office of Naval Research.

[2] Now at the University of California, Berkeley.

such that

$$g + f(x) = \min_{k \in A} \{C(x, k) + \int_{y \in \mathfrak{X}} f(y)\, dP(y \mid x, k)\}, \qquad x \in \mathfrak{X}, \tag{1}$$

then there exists a stationary deterministic policy R^* *such that*

$$g = \varphi(x, R^*) = \min_R \varphi(x, R) \qquad \text{for all} \quad x \in \mathfrak{X}$$

and R^* *is any policy which, for each* x, *prescribes an action which minimizes the right side of* (1).

PROOF. For any policy R.

$$E_R\{\textstyle\sum_{t=1}^{n} [f(X_t) - E_R(f(X_t) \mid S_{t-1}]\} = 0.$$

But

$$\begin{aligned} &E_R[f(X_t) \mid S_{t-1}] \\ &\quad = \int_{y \in \mathfrak{X}} f(y)\, dP(y \mid X_{t-1}, \Delta_{t-1}) \\ &\quad = C(X_{t-1}, \Delta_{t-1}) + \int_{y \in \mathfrak{X}} f(y)\, dP(y \mid X_{t-1}, \Delta_{t-1}) - C(X_{t-1}, \Delta_{t-1}) \\ &\quad \geqq \min_{k \in A} \{C(X_{t-1}, k) + \int_{y \in \mathfrak{X}} f(y)\, dP(y \mid X_{t-1}, k)\} - C(X_{t-1}, \Delta_{t-1}) \\ &\quad = g + f(X_{t-1}) - C(X_{t-1}, \Delta_{t-1}) \end{aligned}$$

with equality for R^* since R^* is defined to take the minimizing action. Hence

$$0 \leqq E_R\{\textstyle\sum_{t=1}^{n} [f(X_t) - g - f(X_{t-1}) + C(X_{t-1}, \Delta_{t-1})]\}$$

or

$$g \leqq E_R f(X_n) n^{-1} - E_R f(X_0) n^{-1} + E_R \textstyle\sum_{t=1}^{n} C(X_{t-1}, \Delta_{t-1}) n^{-1} \tag{2}$$

with equality for R^*. Letting $n \to \infty$ and using the fact that f is bounded, we have that $g \leqq \varphi(R, X_0)$ with equality for R^*, and for all possible values of X_0. QED.

REMARKS. The above proof doesn't make use of the fact that x is a complete separable metric space or that A is finite or even that $C(x, k)$ is bounded. $\mathfrak{X}$ may be any arbitrary probability space and A may be countably infinite. Also the boundedness of $f(\cdot)$ may be replaced by the condition that

$$n^{-1} E_R[f(X_n) \mid X_0 = x] \to 0$$

for all rules R and all x.

Let $g_n(x)$, $n = 1, 2, \cdots$, satisfy

$$\begin{aligned} g_1(x) &= \min_k C(x, k), \\ g_{n+1}(x) &= \min_k \{C(x, k) + \int_{y \in \mathfrak{X}} g_n(y)\, dP(y \mid x, k)\}. \end{aligned} \tag{3}$$

Note that $g_n(x) = \min_R \sum_{t=0}^{n-1} E_R[C(X_t, \Delta_t) \mid X_0 = x]$. The following corollary was proven by Derman [2] for the denumerable case.

COROLLARY 1. *Under the conditions of Theorem* 1, *there is a* M *such that*

$$|g_n(x) - ng| < M \qquad \text{for all} \quad n, x.$$

PROOF. Let M' be such that $|f(x)| < M'$. By (2) we have that $ng \leqq 2M' + g_n(x)$. Again from (2), by letting $R = R^*$ we have that $ng \geqq g_n(x) - 2M'$. QED.

For any policy R, $\beta \varepsilon (0, 1)$, let $\psi(x, \beta, R) = \sum_{t=0}^{\infty} \beta^t E_R[C(X_t, \Delta_t) \mid X_0 = x]$. A policy R_β such that $\psi(x, \beta, R_\beta) = \min_R \psi(x, \beta, R)$ for all $x \varepsilon \mathfrak{X}$ is said to be β-optimal.

We shall need the following result given by Blackwell [1]: If A is finite, and $C(\cdot, \cdot)$ is bounded then, for each $\beta \varepsilon (0, 1)$, there is a stationary deterministic policy R_β which is β-optimal. Furthermore $\psi(x, \beta, R_\beta)$ is the unique solution to

$$(4) \quad \psi(x, \beta, R_\beta) = \min_{k \varepsilon A} \{C(x, k) + \beta \int_{y \varepsilon \mathfrak{X}} \psi(y, \beta, R_\beta)\, dP(y \mid x, k)\}$$

and any policy which, when in state x, selects an action which minimizes the right side of (4) is β-optimal.

Fix some state—call it 0—and let

$$(5) \quad f_\beta(x) = \psi(x, \beta, R_\beta) - \psi(0, \beta, R_\beta)$$

then

$$(6) \quad g_\beta + f_\beta(x) = \min_k \{C(x, k) + \beta \int_{y \varepsilon \mathfrak{X}} f_\beta(y)\, dP(y \mid x, k)\}$$

where $$g_\beta = (1 - \beta)\psi(0, \beta, R_\beta).$$

The following theorem gives sufficient conditions for the existence of a bounded Baire function $f(x)$ and a constant g satisfying (1).

THEOREM 2. *If $\{f_\beta\}$ is a uniformly bounded equicontinuous family of functions then*

(i) *there exists a bounded continuous function $f(x)$ and a constant g satisfying* (1);

(ii) $(1 - \beta)\psi(x, \beta, R_\beta) \to g$ *as* $\beta \to 1^-$ *for all* $x \varepsilon \mathfrak{X}$

PROOF. By the Ascoli theorem there exists a sequence $\beta_\nu \to 1$ and a continuous function f such that $f_{\beta_\nu}(x) \to f(x)$. Now g_β is bounded (since costs are bounded) and so we can also require that $g_{\beta_\nu} \to g$. Hence by (6) and the bounded convergence theorem we have that

$$g + f(x) = \min \{C(x, k) + \int_{y \varepsilon \mathfrak{X}} f(y)\, dP(y \mid x, k)\}.$$

For any sequence $\beta_\nu \to 1$ there is a subsequence β_ν' such that $\lim g_{\beta_\nu'}$ exists. By the above this limit must be g. Thus $g = \lim_{\beta \to 1} g_\beta$. The result follows since 0 is any arbitrary state. QED.

For any stationary deterministic policy R let $x(R)$ be the action R chooses when in state x. We say that $\lim_n R_n = R$ if, for each x, there exists $N_x < \infty$ such that $x(R_n) = x(R)$ for all $n \geqq N_x$.

The following was proven in [4] for denumerable $\mathfrak{X}$. The proof for arbitrary $\mathfrak{X}$ is identical (with the Ascoli theorem replacing the diagonal argument in showing that the sequence $f_{\beta_r}(\cdot)$ has a convergent subsequence).

THEOREM 3. *Under the conditions of Theorem* 2

(i) *if for all but at most a countable number of x's there is a unique action minimiz-*

ing the right side of (1) *then for some sequence* $\beta_r \to 1^-$, *and some* R^*, $R^* = \lim_r R_{\beta_r}$,

(ii) *if* $R = \lim_r R_{\beta_r}$, *where* $\beta_r \to 1^-$ *then* R *is optimal* i.e.

$$\varphi(x, R) = g \quad \text{for all} \quad x \in \mathfrak{X}.$$

The following two conditions were given by Taylor [5] to prove equicontinuity of $\{f_\beta\}$ in the special case of a replacement process:

(a) For every $k \in A$, $C(\cdot, k)$ is continuous.

(b) For every $x \in \mathfrak{X}$, $k \in A$, $P(\cdot \mid x, k)$ is absolutely continuous with respect to some σ-finite measure μ on B and it possesses a density $p(y \mid x, k)$ also assumed to be a Baire function in x. Furthermore, for every $x \in \mathfrak{X}$, $k \in A$,

$$\lim_{x' \to x} \int |p(y \mid x, k) - p(y \mid x', k)| \, d\mu(y) = 0.$$

THEOREM 4. *If conditions* (a) *and* (b) *are satisfied then*

$$|f_\beta(x)| < M \quad \text{for all} \quad x, \beta \Rightarrow \{f_\beta\} \quad \text{is equicontinuous.}$$

PROOF. Follows directly from (6) and conditions (a), (b).

A sufficient condition for the uniform boundedness of $\{f_\beta\}$ is given in [4].

3. Reduction of average cost case to discounted cost case. We shall need the following assumption

ASSUMPTION (I). There is a state—call it 0—and $\alpha > 0$, such that $P\{X_{t+1} = 0 \mid X_t = x, \Delta_t = k\} \geqq \alpha$ for all $x \in \mathfrak{X}$, $k \in A$.

For any process satisfying the above Assumption consider a new process with identical state and action spaces, with identical costs, but with transition probabilities now given for $B \in \mathfrak{B}$ by

$$\begin{aligned} P'(B \mid x, k) &= P(B \mid x, k)/(1 - \alpha) && \text{for} \quad 0 \notin B \\ &= [P(B \mid x, k) - \alpha]/(1 - \alpha) && \text{for} \quad 0 \in B. \end{aligned}$$

Let $\psi'(x, \beta, R)$ be the total expected β-discounted cost, and let R_β' be the β-optimal policy, all with respect to the new process. Letting $f'(x) = \psi'(x, 1 - \alpha, R'_{1-\alpha}) - \psi'(0, 1 - \alpha, R'_{1-\alpha})$ we have by (6) that

$$\begin{aligned} &\alpha\psi'(0, 1 - \alpha, R) + f'(x) \\ &\qquad = \min_k \{C(x, k) + (1 - \alpha) \textstyle\int_{y \in \mathfrak{X}} f'(y) \, dP'(y \mid x, k)\} \\ &\qquad = \min_k \{C(x, k) + \textstyle\int_{y \in \mathfrak{X}} f'(y) \, dP(y \mid x, k)\}. \end{aligned} \tag{7}$$

And thus the conditions of Theorem 1 are satisfied. It follows that $g = \alpha\psi'(0, 1 - \alpha, R'_{1-\alpha})$ and the optimal average-cost policy is the one which selects the actions which minimize the right side of (7). But it is easily seen that $R'_{1-\alpha}$ does exactly this. Hence the optimal average cost policy is precisely the $1 - \alpha$-optimal policy with respect to the new process; and the optimal expected average cost per unit time is $\alpha\psi'(0, 1 - \alpha, R'_{1-\alpha})$.

The above result was proven in [4] for the denumerable case by showing that $\varphi(x, R) = \alpha\psi'(0, 1 - \alpha, R)$ for any stationary policy R. This result also holds for arbitrary $\mathfrak{X}$. However this in itself does not show that $R'_{1-\alpha}$ is optimal. (It

does in the denumerable case because it can be shown that Assumption (I) implies that $\{f_\beta\}$ is uniformly bounded and thus by Theorem 3 there exists a stationary deterministic policy which is optimal.)

4. Concluding remarks. Results given in [4] which dealt with ϵ-optimal policies and replacement processes (Sections 3 and 4) carry over to the more general spaces $\mathfrak{X}$ considered here. The proofs are identical (with integrals replacing sums in the obvious places).

REFERENCES

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

[2] DERMAN, CYRUS (1966). Denumerable state Markovian decision processes—average cost criterion. *Ann. Math. Statist.* **37** 1545–1554.

[3] DERMAN, CYRUS and LIEBERMAN, GERALD J. (1966). A Markovian decision model for a joint replacement and stocking problem. *Management Sci.* **13** 609–617.

[4] ROSS, SHELDON M. (1968). Non-discounted denumerable Markovian decision models. *Ann. Math. Statist.* **39** 412–423.

[5] TAYLOR, HOWARD (1965). Markovian sequential replacement processes. *Ann. Math. Statist.* **36** 1677–1694.