ARBITRARY STATE MARKOVIAN DECISION PROCESSES

Sheldon M Ross
Stanford University
Stanford, California
January 1968

Processed for...
DEFENSE DOCUMENTATION CENTER
DEFENSE SUPPLY AGENCY

CLEARINGHOUSE
FOR FEDERAL SCIENTIFIC AND TECHNICAL INFORMATION

U. S. DEPARTMENT OF COMMERCE / NATIONAL BUREAU OF STANDARDS / INSTITUTE FOR APPLIED TECHNOLOGY

UNCLASSIFIED
ARBITRARY STATE MARKOVIAN DECISION PROCESSES

by

Sheldon M. Ross

TECHNICAL REPORT NO 105
January 8, 1968

Supported by the Army, Navy, Air Force, and NASA under
Contract Nonr-225(53)(NR-042-002)
with the Office of Naval Research

Gerald J. Lieberman, Project Director

Reproduction in Whole or in Part is Permitted for
any Purpose of the United States Government

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
A Markovian Decision Process is a process which is observed at distinct time points to be in some state. After observing the state of the system an action is chosen - corresponding to the action (and the present state) a cost is incurred and the transition probabilities for the next state are determined. A policy is any rule for choosing actions. Corresponding to each policy there is an expected long run average cost per unit time. This paper is concerned with finding an optimal policy - i.e. one whose associated average cost is minimal.

For example we might have a tool which wears out with time. The state of the system could be the length of the tool, and the possible actions could be either to replace the tool or not. Associated with each state there would be an operating cost. Thus a policy is a rule for determining when to replace the tool and an optimal one is one which minimizes the long run average cost.

In the past most of the work in this area has been done under the assumption that the state space is countable. In this paper we let the state space be arbitrary. For example, in the tool problem given above it is natural to let the state space be the continuum of possible values of the length of the tool.

This paper presents sufficient conditions for the existence of an optimal policy and for it to be of simple type. This type - called stationary deterministic - is of the form of a function mapping the state space into the action space. For example, in the tool problem...
a stationary deterministic policy would replace whenever the length of the tool is in some specified set of real numbers. The method employed is to treat the average cost problem as a limit of either the discounted cost problem or the nondiscounted n-stage problem. We also show how, in a special case, the average cost problem may be reduced to a discounted cost problem.
ARBITRARY STATE MARKOVIAN DECISION PROCESSES
Sheldon M. Ross

1. Introduction

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

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

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 \( \{D_k(S_{t-1}, x)\}_{k \in \mathcal{A}} \) satisfying
\[
D_k(S_{t-1}, x) \geq 0 \text{ for all } k \in \mathcal{A}, \quad \sum_{k \in \mathcal{A}} D_k(S_{t-1}, x) = 1 \text{ for every } (S_{t-1}, x).
\]
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 $\phi(x,R) = \limsup_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n} \mathbb{E} [C(X_t, \Delta_t) | X_0 = x]$. Thus $\phi(x,R)$ 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 $x$ 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. We also show how to treat the average cost problem as a limit of the $n$-stage problem. One of the advantages of this approach is that it enables us to determine, for denumerable $x$, a necessary and sufficient condition for the existence of a bounded solution to a functional equation which characterizes the optimal policy.

2. Stationary Deterministic Optimal Policies

The following theorem was originally proven by Derman [2] for the special case that $x$ 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 $f(x)$, $x \in X$ and a constant $g$, such that

$$
g + f(x) = \min_{k \in A} \left\{ C(x,k) + \int f(y) \, dP(y|x,k) \right\} \quad x \in X
$$

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

$$
g = \phi(x,R^*) = \min_{R} \phi(x,R) \quad \text{for all } x \in X
$$
and \( R^* \) is any policy which, for each \( x \), prescribes an action which minimizes the right side of (1).

**Proof:** For any policy \( R \)

\[
E_R \left( \sum_{t=1}^{n} [f(X_t) - E_R(f(X_t)|S_{t-1})] \right) = 0
\]

But

\[
E_R[f(X_t)|S_{t-1}] = \int_{y \in X} f(y) dP(y|X_{t-1}, A_{t-1})
\]

\[
= C(X_{t-1}, A_{t-1}) + \int_{y \in X} f(y) dP(y|X_{t-1}, A_{t-1}) - C(X_{t-1}, A_{t-1})
\]

\[
\geq \min_{k \in A} \{ C(X_{t-1}, k) + \int_{y \in X} f(y) dP(y|X_{t-1}, k) \} - C(X_{t-1}, A_{t-1})
\]

\[
\geq g + f(X_{t-1}) - C(X_{t-1}, A_{t-1})
\]

with equality for \( R^* \) since \( R^* \) is defined to take the minimizing action.

Hence

\[
0 \leq E_R \left( \sum_{t=1}^{n} [f(X_t) - g - f(X_{t-1}) + C(X_{t-1}, A_{t-1})] \right)
\]

or

\[
g \leq E_R \left( \frac{\sum_{t=1}^{n} f(X_t)}{n} - \frac{f(X_0)}{n} \right) + E_R \left( \frac{\sum_{t=1}^{n} C(X_{t-1}, A_{t-1})}{n} \right)
\]

with equality for \( R^* \). Letting \( n \to 0 \) and using the fact that \( f \) is bounded, we have that \( g \leq \phi(R, X_0) \) with equality for \( R^* \), and for all possible values of \( X_0 \). QED.

**Remark:** Note that the above proof doesn't make use of the fact that \( A \) is finite or that \( C(x, k) \) is bounded.
Let $g_n(x)$, $n = 1, 2, \ldots$ satisfy

$$g_1(x) = \min_k C(x,k)$$
$$g_{n+1}(x) = \min_k \left\{ C(x,k) + \int g_n(y)dP(y|x,k) \right\}$$

Note that $g_n(x) = \min_{x \in X} E_{\mathcal{F}}[C(X_t, A_t) | X_0 = x]$. 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

$$|g_n(x) - ng| < M$$

for all $n,x$

**Proof:** Let $M'$ be such that $|f(x)| < M'$. By (2) we have that $g \leq 2M' + g_n(x)$. Again from (2), by letting $R = R^*$ we have that $g \geq g_n(x) - 2M'$. QED.

Fix some state - call it $0$ - and let

$$f_n(x) = g_n(x) - g_n(0)$$

all $n,x$

One has from (3) that

$$g_{n+1}(0) - g_n(0) + f_n(x) = \min_k \left\{ C(x,k) + \int f_n(y)dP(y|x,k) \right\}$$

We shall now determine sufficient (and in the denumerable case necessary and sufficient) conditions for the existence of a bounded Baire function $f(x)$ and a constant $g$ satisfying (1).

**Theorem 2:** If $\{f_n\}$ is a uniformly bounded equicontinuous family of functions then

(1) there exists a bounded continuous function $f(x)$ and a constant $g$ satisfying (1).
(ii) \( \lim_{n \to \infty} (g_{n+1}(x) - g_n(x)) = g \) for all \( x \in X \).

Proof: By the Ascoli Theorem there exists a subsequence \( \{f_{n_k}\} \) and a continuous function \( f \) such that \( f_{n_k}(x) \to f(x) \). Now \( g_{n+1}(0) - g_n(0) \) is bounded (since costs are bounded) and so we can also require that

\[ g_{n+1}(0) - g_n(0) = g. \]

Hence by (5) and the bounded convergence theorem we have that \( g + f(x) = \min \{C(x,k) + \int f(y)dP(y|x,k)\} \).

For any subsequence \( \{n\} \) of \( \{n_i\} \) there is a sub-subsequence \( \{n''\} \) such that \( \lim (g_{n''+1}(0) - g_n(0)) \) exists. By the above this limit must be \( g \). Thus \( g = \lim g_{n+1}(0) - g_n(0) \). The result follows since 0 is any arbitrary state. QED.

If \( X \) is denumerable, then \( \{f_n\} \) can always be taken to be equicontinuous by considering the discrete topology. We thus have

Corollary 2: If \( X \) is denumerable, then a necessary and sufficient condition for the existence of a bounded function \( f(x) \) and constant \( g \) satisfying (1) is that there is a \( M < \infty \) such that \( |g_n(x) - g_n(0)| < M \) for all \( n,x \).

Proof: Sufficiency follows from the above theorem and necessity follows from Corollary 1. QED.

For any policy \( R, \beta \in (0,1) \), let \( \psi(x,\beta,R) = \sum_{t=0}^{\infty} \beta^t E_R [C(X_t,0_t)|X_0 = x] \).

A policy \( R_\beta \) such that \( \psi(x,\beta,R_\beta) = \min_R \psi(x,\beta,R) \) for all \( x \in X \) is said to be \( \beta \)-optimal.

We shall need the following result given by Blackwell [1]:

If \( A \) is finite, and \( C(\cdot,\cdot) \) is bounded then, for each \( \beta \in (0,1) \), there is a stationary deterministic policy \( R_\beta \) which is \( \beta \)-optimal. Furthermore \( \psi(x,\beta,R_\beta) \) is the unique solution to

\[ C(x,\beta,0) + \int \psi(y,\beta,R_\beta)dP(y|x) \]
and any policy which, when in state $x$, selects an action which minimizes the right side of (6) is $\beta$-optimal.

Fix some state - call it 0 - and let

$$f_0(x) = \psi(x,\beta,R_0) - \psi(0,\beta,R_0)$$

then

$$g + f_0(x) = \min_{k \in A} \{C(x,k) + \beta \int f_0(y) \, dp(y|x,k)\}$$

where

$$g = (1-\beta) \psi(0,\beta,R_0)$$

In analogous fashion to Theorem 2 we have

**Theorem 3:** If \( \{f_\beta\} \) is a uniformly bounded equicontinuous family of functions then

1. there exists a bounded continuous function $f(x)$ and a constant $g$ satisfying (1).
2. $(1-\beta)g(x) + g$ as $\beta \to 1^-$ for all $x \in \mathcal{X}$.

**Proof:** Same as proof of Theorem 2.

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

The following was proven in [4] for denumerable $\mathcal{X}$. The proof for arbitrary $\mathcal{X}$ is identical

\[ (6) \quad \psi(x,\beta,R_0) = \min_{k \in A} \{C(x,k) + \beta \int \psi(y,\beta,R_0) \, dp(y|x,k)\} \]
Theorem 4: Under the conditions of Theorem 3

(i) for some sequence \( \beta_r \rightarrow 1^- \), \( R^* = \lim_{r \to \infty} R_{\beta_r} \)

(ii) if \( R = \lim_{r \to \infty} R_{\beta_r} \), where \( \beta_r \rightarrow 1^- \) then \( R \) is optimal - i.e.

\[ \phi(x,R) = g \text{ for all } x \in X. \]

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 \( x \in X \), \( k \in A \), \( P(x,k) \) is absolutely continuous with respect to some \( \sigma \)-finite measure \( \nu \) on \( B \) and it possesses a density \( p(y|x,k) \) also assumed to be a Baire function in \( x \). Furthermore, for every \( x \in X \), \( k \in A \)

\[ \lim_{x' \to x} \int |P(y|x,k) - P(y|x',k)|d\nu(y) = 0 \]

Theorem 5: If conditions (a) and (b) are satisfied then

(i) \( |f_\beta(x)| < M \) for all \( x, \beta = \{f_\beta\} \) is equicontinuous

(ii) \( |f_n(x)| < M \) for all \( x, n = \{f_n\} \) is equicontinuous

Proof: Follows directly from (5) and (8) 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 (1): There is a state - call it 0 - and \( a = 0 \), such that

\[ P(X_{t+1} = 0|X_t = x, \Delta_t = k) = \alpha \text{ for all } x \in X, k \in A. \]
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 by

\[
P'(x,k) = \begin{cases} 
P(B|x,k) & \text{for } O \notin B \\ 1 - \alpha & \text{for } B \in \mathbb{R} \\ P(B|x,k) - \alpha & \text{for } O \in B \\ 1 - \alpha & \end{cases}
\]

Let \( \psi'(x,\beta,\mathcal{R}) \) be the total expected \( \beta \)-discounted cost, and let \( R'_\beta \) be the \( \beta \)-optimal policy, all with respect to the new process.

Letting \( f'(x) = \psi'(x,1-\alpha,R'_1) - \psi'(0,1-\alpha,R'_1) \) we have by (8) that

\[
(9) \quad \alpha \psi'(0,1-\alpha,\mathcal{R}) + f'(x) = \min_k \{ C(x,k) + (1-\alpha) \int f'(y) dP(y|x,k) \},
\]

And thus the conditions of Theorem 1 are satisfied. It follows that \( g = \alpha \psi'(0,1-\alpha,R'_1) \) and the optimal average-cost policy is the one which selects the actions which minimize the right side of (9). 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 \( \alpha \psi'(0,1-\alpha,R'_1) \).

The above result was proven in [4] for the denumerable case by showing that \( \phi(x,\mathcal{R}) = \alpha \psi'(0,1-\alpha,\mathcal{R}) \) for any stationary policy \( \mathcal{R} \).
This result also holds for arbitrary $x$. However this in itself does not show that $R_{1-n}$ is optimal. (It does in the denumerable case because it can be shown that Assumption (I) implies that $\{f_{g}\}$ 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 $\varepsilon$-optimal policies and replacement processes (Sections 3 and 4) carry over to the more general spaces $x$ considered here. The proofs are identical (with integrals replacing sums in the obvious places).
REFERENCES


Arbitrary state, finite action Markovian decision processes are studied with respect to the (long-run) average cost criterion. The problem is treated both as a limiting case of the discounted cost problem and also as a limit of the n-stage problem. Sufficient conditions are given for the existence of an optimal rule and for it to be of stationary deterministic type.
Markovian Decision Process

Arbitrary State Space

Stationary Deterministic Optimal Rule