The Annals of Mathematical Statistics
1970, Vol. 41 No. 1, 78–86

DISCRETE-TIME MARKOVIAN DECISION PROCESSES WITH INCOMPLETE STATE OBSERVATION

BY YOSHIKAZU SAWARAGI AND TSUNEO YOSHIKAWA

Kyoto University

1. Introduction. Discrete-time Markovian decision processes (MDP's) with complete state observation and an infinite planning horizon, have been investigated by many authors (for example [4], [5], [9], [10]).

MDP's with incomplete state observation have also been investigated by several authors [1], [2], and [7]. Dynkin [7] has treated a very general discrete-time problem which includes MDP's with and without complete state observation as special cases. However, the relation between [7] and [5], [10] is not clear. Åström [2] and Aoki [1] have treated the case of a finite planning horizon (control interval).

In this paper it is shown that MDP's with incomplete state observation, countable possible states, uncountable available actions and an infinite planning horizon, can be transformed to MDP's with complete state observations and uncountable possible states. The latter MDP's are those which have been treated in [5] and [10]. The states of the latter MDP's are the probability distributions on the set of the states of the former. Similar transformations have been pointed out by several authors [1]–[3]. However, the above transformation should be formulated explicitly.

2. Probabilistic definitions and notation. In this section we develop the probabilistic notation to be used throughout the paper. We follow [5] and [10] as closely as possible.

A *Borel set* X is a Borel subset of a complete separable metric space. For any Borel set X, $\mathscr{B}(X)$ denotes the σ-field of Borel subsets of X. *Measurable* means measurable with respect to $\mathscr{B}(X)$. A *probability* on a non-empty Borel set X is a probability measure defined on $\mathscr{B}(X)$, and the set of all probabilities on X is denoted by $P(X)$. If X and Y are non-empty Borel sets, a *conditional probability* on Y given X is a function $q(\cdot \mid \cdot)$ such that for each $x \in X$, $q(\cdot \mid x)$ is a probability on Y and for each $E \in \mathscr{B}(Y)$, $q(E \mid \cdot)$ is a Baire function on X. The set of all conditional probabilities on Y given X is denoted by $Q(Y \mid X)$. $p \in P(X)$ and $q \in Q(Y \mid X)$ are also denoted by $p([x])$ and $q([y] \mid x)$ respectively, using the coordinate variables x, y of X, Y, in order to indicate explicitly the spaces where these probabilities are defined. We denote the Cartesian product of X and Y by XY. Every probability $p \in P(XY)$ has a factorization $p = p'q$, where $p' \in P(X)$ is the marginal distribution of the first coordinate variable under p, and $q \in Q(Y \mid X)$ is a version of the conditional distribution of the second coordinate variable given the first. $F(X)$ denotes either the set of all bounded Baire functions on X or the set of all non-positive, extended real-valued Baire functions on X. Unless otherwise noted, statements made about elements of $F(X)$ are valid for either definition.

Received December 2, 1968.

For any $u \in F(XY)$ and any $q \in Q(Y \mid X)$, qu denotes the element of $F(X)$ whose value at $x \in X$ is given by $qu(x) = \int_Y u(x,y)\, dq([y] \mid x)$. For any $p \in P(X)$, $q \in Q(Y \mid X)$, pq is the probability on XY such that for all $u \in F(XY)$, $pq(u) = p(qu)$. If $u, v \in F(X)$, $u \geqq v$ means $u(x) \geqq v(x)$ for all $x \in X$.

The above notation extends in an obvious way to a finite or infinite sequence of non-empty Borel sets $X_1, X_2, \cdots$. If $q_n \in Q(X_{n+1} \mid X_1 \cdots X_n)$ for $n \geqq 1$ and $p \in P(X_1)$, then $pq_1 \cdots q_n \in P(X_1 \cdots X_{n+1})$, $pq_1q_2 \cdots \in P(X_1X_2 \cdots)$, $q_2q_3 \in Q(X_3X_4 \mid X_1X_2)$ and for any $u \in F(X_1 \cdots X_{n+1})$, $l \leqq n$, $q_l \cdots q_n u \in F(X_1 \cdots X_l)$, etc. To avoid further complicating the notation we shall use the following convention: for any function u on Y, we shall use the same symbol u to denote the function v on XY such that $v(x,y) = u(y)$ for all y. Thus, for example, if $q \in Q(Y \mid X)$, $u \in F(Y)$, then $qu \in F(X)$ and $q \in Q(Y \mid X)$ will also denote the element $q' \in Q(Y \mid ZX)$ such that $q'(\cdot \mid z, \cdot) = q(\cdot \mid \cdot)$, etc.

A $p \in P(X)$ is *degenerate* if for some $x \in X$, $p(\{x\}) = 1$. A $q \in Q(Y \mid X)$ is *degenerate* if $q(\cdot \mid x)$ is degenerate for each x, and this happens if and only if there is a measurable function f from X to Y such that $q(\{f(x)\} \mid x) = 1$ for each x. We will also denote by f the associated degenerate q, so that for $u \in F(XY)$, $fu(x) = u(x, f(x))$ for all $x \in X$. Throughout the paper, we shall denote the completion of a proof by □.

3. Definitions and notation on MDP's with incomplete state observation. In this section we develop the definitions and notation on a class of MDP's with incomplete state observation in a similar way to [5] and [10]. This class of MDP's will be referred to as MDP–II, and MDP's with complete state observation treated in [5] and [10] will be referred to as MDP–I.

MDP–II is defined by $S, A, M, q^s q^m, \varphi_0, r^s$ and β. S is the *set of states*, M is the *set of observation signals*, and $S = M = \{1, 2, \cdots\} \in \mathscr{B}(R)$ where R is 1-dimensional Euclidean space. The *set of actions* A is a non-empty Borel set. The *law of motion* q^s is an element of $Q(S \mid SA)$, the *characteristic of the measuring system* $q^m \in Q(M \mid S)$, the *initial information* $\varphi_0 \in \Phi = P(S)$, the *return function* $r^s \in F(SA)$, and the *discount factor* β *is* $0 \leqq \beta \leqq 1$. The restriction of S and M to countable sets is for the sake of simplifying mathematical treatment. For probabilities on S and M, one point sets $\{s\}$ and $\{m\}$ will be denoted by s and m without parentheses { }; for example, $q^m(\{m\} \mid s)$ is denoted by $q^m(m \mid s)$ etc.

Assume that the state of the system at time $n, n = 0, 1, 2, \cdots$, is $s_n \in S$, and that we choose an action $a_n \in A$, then the system moves to a new state s_{n+1}, selected according to $q^s([s_{n+1}] \mid s_n, a_n)$ and we receive a reward $r^s(s_n, a_n)$. We cannot observe the state s_{n+1} directly. We can only obtain an observation signal $m_{n+1} \in M$ generated according to $q^m([m_{n+1}] \mid s_{n+1})$. This is the very point in which MDP–II is different from MDP–I. At time 0 we are given a probability φ_0 as the initial information on the initial state s_0. Since $S = \{1, 2, \cdots\}$, φ_0 is specified by $\{\varphi_0(1), \varphi_0(2), \cdots\}$.

The set Φ is metrizable by introducing the distance

$$d(\varphi', \varphi'') = \sum_{i=1}^{\infty} |\varphi'(i) - \varphi''(i)|, \qquad \varphi', \varphi'' \in \Phi. \tag{1}$$

Then the following lemma can easily be proved.

LEMMA 1. *The metric space Φ is complete and separable.*

Therefore, Φ itself is a Borel set.

A *policy* ω is a sequence $\{\omega_1, \omega_2, \cdots\}$, where $\omega_n \in Q(A \mid D_n)$ and $D_n = \Phi AMAM \cdots M(2n+1 \text{ factors})$ is the set of possible data concerning the history of the system up to the nth stage. Given that we have obtained data $d_n = (\varphi_0, a_0, m_1, a_1, m_2, \cdots, m_n) \in D_n$, we choose the nth action a_n according to $\omega_n([a_n] \mid d_n)$.

We define a conditional probability $q^P \in Q(S \mid \Phi)$ by

$$q^P(i \mid \varphi_0) = \varphi_0(i), \qquad i = 1, 2, \cdots. \tag{2}$$

Any policy ω, together with q^s and q^m defines for each initial information φ_0 a conditional distribution on the set $S\Omega = SASMASM \cdots$ of the future of the system, i.e. it defines

$$e_\omega = q^P \omega_0 \, q^s q^m \omega_1 \, q^s q^m \ldots \in Q(S\Omega \mid \Phi). \tag{3}$$

Any return function r^s defines a *total discounted return function* on $SASA \cdots$ given by

$$v(s_0, a_0, s_1, a_1, \cdots) = \sum_{n=0}^{\infty} (\beta)^n r^s(s_n, a_n) \tag{4}$$

and an *expected return function* on Φ given by

$$J(\omega) = e_\omega v = \sum_{n=0}^{\infty} (\beta)^n q^P \omega_0 \, q^s q^m \omega_1 \cdots \omega_n \, r^s(s_n, a_n). \tag{5}$$

The value of $J(\omega)$ at $\varphi_0 \in \Phi$ is denoted by $J(\omega)(\varphi_0)$.

The problem is to maximize $J(\omega)(\varphi_0)$ with respect to ω given the initial information φ_0. In order that the problem be well-defined, we assume that one of the following three conditions is satisfied.

(a) *The discounted case.* r^s is bounded and $0 \leqq \beta < 1$.

(b) *The positive bounded case.* r^s is non-negative and bounded, $\beta = 1$, and the structure of the problem is such that $J(\omega) \leqq R$ for any ω, where R is a positive number independent of ω.

(c) *The negative case.* $0 \geqq r^s > -\infty$, and $\beta = 1$.

From now on, in the discounted and positive bounded cases $F(X)$ is to be understood to denote the set of bounded Baire functions, and in the negative case $F(X)$ is the set of non-positive, extended real-valued Baire functions. We introduce the discount factor $\beta = 1$ in cases (b) and (c) only to allow a common notation throughout the paper.

For any $p \in P(\Phi)$ and $\varepsilon > 0$, we say that ω^* is (p, ε)-*optimal* if $p\{\varphi_0; J(\omega^*)(\varphi_0) \geqq \sup_\omega J(\omega)(\varphi_0) - \varepsilon\} = 1$. The set $\{\varphi_0; J(\omega^*)(\varphi_0) \geqq \sup_\omega J(\omega)(\varphi_0) - \varepsilon\}$ is in general not Borel; however it is shown in [10] that it is in the completion of the Borel sets with respect to p, hence the statement has meaning. We say that ω^* is ε-*optimal* if $J(\omega^*) \geqq \sup_\omega J(\omega) - \varepsilon$, and that ω^* is *optimal* if $J(\omega^*) \geqq \sup_\omega J(\omega)$. We say that ω^* dominates ω if $J(\omega^*) \geqq J(\omega)$. These definitions of optimality correspond to those of [10], but we can also define them following [5].

Notice that if the initial state s_0 is regarded as a "state of nature" (unknown

parameter) these definitions of optimality are similar to those of ε-Bayes and Bayes decision rules in the statistical decision theory [8]. Hence the optimal policy is one which, whatever the a priori distribution $\varphi_0 \in \Phi$ given, is a Bayes decision rule for this distribution in the above-cited meaning.

As far as we consider policies on the data space D_n, we cannot expect fruitful results. In the next section, the notion "I-policy" (information policy) will be introduced and it will be shown that we can restrict our attention to the set of all I-policies. Intuitively, an I-policy is one which is based upon the past actions and conditional probabilities of the states of the system given the past data.

4. I-policy. Let the conditional probability of s_n, given a datum d_n, be denoted by $q_n = q_n([s_n] \mid d_n) \in Q(S \mid D_n)$. According to Bayesian formula, q_n satisfies, for any d_n, a_n, m_{n+1},

$$
\begin{aligned}
q_{n+1}(i \mid d_{n+1}) &= q_{n+1}(i \mid d_n, a_n, m_{n+1}) \\
&= \textstyle\sum_{j=1}^{\infty} q^m(m_{n+1} \mid i) q^s(i \mid j, a_n) q_n(j \mid d_n) \\
&\quad \div \textstyle\sum_{i'=1}^{\infty} \sum_{j'=1}^{\infty} q^m(m_{n+1} \mid i') q^s(i' \mid j', a_n) q_n(j' \mid d_n), \qquad i = 1, 2, \cdots .
\end{aligned} \tag{6}
$$

In the case in which the denominator of (6) vanishes, q_{n+1} can be assigned arbitrarily. If d_n is regarded as a parameter, q_n corresponds to an element $\varphi_n \in \Phi$ such that

$$
\varphi_n(i) = q^P(i \mid \varphi_n) = q_n(i \mid d_n). \tag{7}
$$

From (6) and (7) φ_{n+1} can be determined by φ_n, a_n, and m_{n+1}; that is, there is a function g mapping from ΦAM to Φ:

$$
\varphi_{n+1} = g(\varphi_n, a_n, m_{n+1}). \tag{8}
$$

Then

$$
\begin{aligned}
\varphi_{n+1}(i) &= \textstyle\sum_{j=1}^{\infty} q^m(m_{n+1} \mid i) q^s(i \mid j, a_n) q^P(j \mid \varphi_n) \\
&\quad \div \textstyle\sum_{i'=1}^{\infty} \sum_{j'=1}^{\infty} q^m(m_{n+1} \mid i') q^s(i' \mid j', a_n) q^P(j' \mid \varphi_n) \\
&= g(\varphi_n, a_n, m_{n+1})(i), \qquad i = 1, 2, \cdots .
\end{aligned} \tag{9}
$$

THEOREM 1. *The function g is measurable.*

PROOF. Since, for each i, the real-valued function $g(\varphi_n, a_n, m_{n+1})(i)$ is a Baire function on ΦAM, it follows from (2.1) of [6] that $g(\varphi_n, a_n, m_{n+1})$ is a measurable function from ΦAM to Φ. □

By repeated use of g, corresponding to any $d_n \in D_n$ an element $b_n = (\varphi_0, a_0, \varphi_1, a_1, \cdots, \varphi_n)$ of $B_n = \Phi A \Phi A \cdots \Phi$ ($2n+1$ factors) is determined, where B_n is the set of possible informations concerning the history of the system up to the nth stage.

An *I-policy* π is a sequence $\{\pi_1, \pi_2, \cdots\}$ where π_n is a conditional probability on A given B_n;

$$
\pi_n = \pi_n([a_n] \mid b_n). \tag{10}
$$

An I-policy is called *stationary* if each π_n is a degenerate element of $Q(A \mid S)$ and independent of n, i.e., if there is a measurable function f from Φ to A such that

$$\pi_n(\{f(\varphi_n)\} \mid \varphi_n) = 1 \quad \text{for all} \quad \varphi_n \in \Phi. \tag{11}$$

For any π, define $\omega^\pi = \{\omega_0{}^\pi, \omega_1{}^\pi, \cdots\}$ by

$$\begin{aligned} \omega_0{}^\pi([a_0] \mid \varphi_0) &= \pi_0([a_0] \mid \varphi_0) \\ \omega_n{}^\pi([a_n] \mid d_n) &= \pi_n([a_n] \mid b_n{}^{d}) \qquad n = 1, 2, \cdots \end{aligned} \tag{12}$$

where $b_n{}^d$ is a point in B_n which corresponds to $d_n \in D_n$. Then the policy ω^π is equivalent to π, that is, ω^π assigns the same conditional probability on A as that assigned by π for any datum d_n. Hence an I-policy can be regarded as a policy, and the set Π of all I-policies as a subset of the set W of all policies.

Due to Theorem 1 we can define a conditional probability $e_{\pi\varphi}$ by

$$e_{\pi\varphi} = q^P \pi_0 \, q^s q^m g \pi_1 \, q^s q^m g \pi_2 \cdots \in Q(S\Omega_\varphi \mid \Phi) \tag{13}$$

where $\Omega_\varphi = ASM\Phi ASM\Phi \cdots$. Similarly to the case of policy ω, for any I-policy π, an expected return function on Φ, $J(\pi)$, is given by

$$J(\pi) = e_{\pi\varphi} v = \textstyle\sum_{n=0}^{\infty} (\beta)^n q^P \pi_0 \, q^s q^m g \pi_1 \cdots \pi_n r^s(s_n, a_n). \tag{14}$$

Notice that $D_n, B_n \subset \Phi S\Omega_\varphi, \Omega \subset \Omega_\varphi$ and that

$$J(\omega) = e_\omega v = e_{\omega\varphi} v \tag{15}$$

where

$$e_{\omega\varphi} = q^P \omega_0 \, q^s q^m g \omega_1 \, q^s q^m g \omega_2 \cdots \in Q(S\Omega_\varphi \mid \Phi). \tag{16}$$

5. I-policies are enough. A subset W_c of W will be called *complete* if for any $\omega \in W$ there exists a policy $\omega^* \in W_c$ which dominates ω; $J(\omega^*) \geqq J(\omega)$, (in statistical decision theory, W_c is called essentially complete [8]).

We are now going to show that Π is complete. For this purpose we will make some preparations.

The conditional probability on $AM\Phi AM\Phi \cdots \Phi$ ($3n$ factors), given Φ obtained from $e_{\omega\varphi}$ (a marginal distribution under $e_{\omega\varphi}$), will be denoted by $q_\omega([\bar{\varphi}_n, \bar{a}_{n-1}\bar{m}_n] \mid \varphi_0)$, and that obtained from $e_{\pi\varphi}$ by $q_\pi([\bar{\varphi}_n, \bar{a}_{n-1}, \bar{m}_n] \mid \varphi_0)$; where, for the sake of simplicity of notation, $(a_0, m_1, \varphi_1, \cdots, \varphi_n)$ is denoted by $(\bar{\varphi}_n, \bar{a}_{n-1}, \bar{m}_n)$, and $\bar{\varphi}_0$ and $\bar{m}_0$ are empty by convention. Other conditional probabilities $q_\omega(\cdot \mid \cdot)$ and $q_\pi(\cdot \mid \cdot)$ also have the same meaning.

LEMMA 2. *For any $r^s \in F(SA)$, $n \geqq 0$, ω and π*

(a) $e_{\omega\varphi} r^s(s_n, a_n) = e_{\omega\varphi} r^\varphi(\varphi_n, a_n)$
(b) $e_{\pi\varphi} r^s(s_n, a_n) = e_{\pi\varphi} r^\varphi(\varphi_n, a_n)$

where

$$r^\varphi(\varphi, a) = q^P r^s(s, a) \in F(\Phi A). \tag{17}$$

PROOF. We shall prove (a); the proof of (b) is similar.

$$\begin{aligned} e_{\omega\varphi} r^s(s_n, a_n) &= q_\omega([s_n, \varphi_n, a_n] \mid \varphi_0) r^s(s_n, a_n) \\ &= q_\omega([\varphi_n, a_n] \mid \varphi_0) q_\omega([s_n] \mid \varphi_0, \varphi_n, a_n) r^s(s_n, a_n) \\ &= q_\omega([\varphi_n, a_n] \mid \varphi_0) q^P([s_n] \mid \varphi_n) r^s(s_n, a_n) \\ &= q_\omega([\varphi_n, a_n] \mid \varphi_0) r^\varphi(\varphi_n, a_n) \\ &= e_{\omega\varphi} r^\varphi(\varphi_n, a_n). \quad \square \end{aligned}$$

THEOREM 2. *For any fixed sequence of actions* $\{a_0, a_1, \cdots\}$, *the sequence of conditional probabilities considered in the space* Φ, $\{\varphi_0, \varphi_1, \cdots\}$, *is a Markov process. Its transition probability at the* nth *stage depends only upon* a_n *among* $\{a_0, a_1, \cdots\}$, *and is given by* $q^\varphi([\varphi_{n+1}] \mid \varphi_n, a_n) \in Q(\Phi \mid \Phi A)$:

$$q^\varphi(\Gamma \mid \varphi_n, a_n) = \sum_{k \in \Gamma m} \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} q^m(k \mid j) q^s(j \mid i, a_n) q^P(i \mid \varphi_n) \tag{18}$$

for any $\Gamma \in \mathscr{B}(\Phi)$, *where*

$$\Gamma_m = \Gamma_m(\varphi_n, a_n, \Gamma) = \{m_{n+1}; g(\varphi_n, a_n, m_{n+1}) \in \Gamma\}. \tag{19}$$

PROOF.

$$\begin{aligned} q(\varphi_{n+1} \in \Gamma \mid \varphi_0, \bar{\varphi}_n, \bar{a}_n) &= q(m_{n+1} \in \Gamma_m \mid \varphi_0, \bar{\varphi}_n, \bar{a}_n) \\ &= \sum_{s_{n+1} \in S} \sum_{s_n \in S} q(m_{n+1} \in \Gamma_m \mid s_{n+1}, s_n, \varphi_0, \bar{\varphi}_n, \bar{a}_n) \\ &\quad \times q(s_{n+1} \mid s_n, \varphi_0, \bar{\varphi}_n, \bar{a}_n) q(s_n \mid \varphi_0, \bar{\varphi}_n, \bar{a}_n) \\ &= q^\varphi(\Gamma \mid \varphi_n, a_n). \quad \square \end{aligned}$$

REMARK. This theorem is essentially the same as Lemma 1 of [2].

LEMMA 3. *For any policy* ω *there exists an I-policy* π *which satisfies, for any* $n \geqq 0$ *and* $u \in F(\Phi A)$

$$e_{\pi\varphi} u(\varphi_n, a_n) = e_{\omega\varphi} u(\varphi_n, a_n). \tag{20}$$

PROOF. For any given ω, define $\pi^\omega = \{\pi_0{}^\omega, \pi_1{}^\omega \cdots,\}$ by

$$\begin{aligned} &\pi_0{}^\omega([a_0] \mid \varphi_0) = \omega_0([a_0] \mid \varphi_0), \\ &\pi_k{}^\omega([a_k] \mid b_k) = \pi_k{}^\omega([a_k] \mid \varphi_0, \bar{\varphi}_k, \bar{a}_{k-1}) \\ &\quad = \int_M \cdots \int_M \omega_k([a_k] \mid \varphi_0, \bar{a}_{k-1}, \bar{m}_k)\, dq_\omega([\bar{m}_k] \mid \varphi_0, \bar{\varphi}_k, \bar{a}_{k-1}), \qquad k = 1, 2, \cdots. \end{aligned} \tag{21}$$

We shall prove that this π^ω satisfies (20). From now on in this proof, π^ω is denoted by π for simplicity of notation. Since, by construction,

$$q_\pi([a_n] \mid \varphi_0, \bar{\varphi}_n, \bar{a}_{n-1}) = q_\omega([a_n] \mid \varphi_1, \bar{\varphi}_n, \bar{a}_{n-1})$$

and by Theorem 2,

$$\begin{aligned} q_\pi([\varphi_n] \mid \varphi_0, \bar{\varphi}_{n-1}, \bar{a}_{n-1}) &= q_\omega([\varphi_n] \mid \varphi_0, \bar{\varphi}_{n-1}, \bar{a}_{n-1}) \\ &= q^\varphi([\varphi_n] \mid \varphi_{n-1}, a_{n-1}) \end{aligned}$$

we obtain $q_\pi([\varphi_n, a_n] \mid \varphi_0) = q_\omega([\varphi_n, a_n] \mid \varphi_0)$.

Therefore

$$e_{\pi\varphi}u(\varphi_n, a_n) = q_\pi([\varphi_n, a_n]\,|\,\varphi_0)u(\varphi_n, a_n)$$
$$= q_\omega([\varphi_n, a_n]\,|\,\varphi_0)u(\varphi_n, a_n) = e_{\omega\varphi}u(\varphi_n, a_n). \quad \square$$

THEOREM 3. *The set of all I-policies,* Π, *is complete.*

PROOF. From Lemmas 2 and 3, for any policy $\omega \in W$ there exists an I-policy $\pi \in \Pi \subset W$ which satisfies

$$J(\pi) = \sum_{n=0}^{\infty}(\beta)^n e_{\pi\varphi}\, r^s(s_n, a_n) = \sum_{n=0}^{\infty}(\beta)^n e_{\pi\varphi}\, r^\varphi(\varphi_n, a_n) \tag{22}$$
$$= \sum_{n=0}^{\infty}(\beta)^n e_{\omega\varphi}\, r^\varphi(\varphi_n, a_n) = \sum_{n=0}^{\infty}(\beta)^n e_{\omega\varphi}\, r^s(s_n, a_n) = J(\omega). \quad \square$$

Hereafter we can restrict our attention to only I-policies.

6. Transformation of MDP-II to MDP-I. In this section we show that MDP-II can be transformed to MDP-I.

Noting Theorem 2, we consider the following Markovian decision process, MDP–I′, which is one of MDP–I.

MDP–I′ is defined by Φ, A, q^φ, r^φ and β. Φ is now the set of states of the system. q^φ is given by (18) and r^φ by (17). At the nth stage we observe the current state $\varphi_n \in \Phi$ completely, then choose an action $a_n \in A$. Then the system moves to a new state φ_{n+1}, selected according to $q^\varphi([\varphi_{n+1}]\,|\,\varphi_n, a_n)$, and we receive a reward $r^\varphi(\varphi_n, a_n)$. The process is repeated from the new state φ_{n+1}, and we wish to maximize the total discounted expected reward with the discount factor β.

As defined in [5] and [10], a policy for MDP–I′ is a sequence of conditional probabilities on A given B_n, $n = 1, 2, \cdots$, where B_n is the set of possible histories $(\varphi_0, a_0, \varphi_1, a_1, \cdots, \varphi_n)$ of the system at the nth stage; this policy is the same as an I-policy for MDP–II. Hence a policy for MDP–I′ will also be denoted by π. Then the total discounted expected return function on Φ for MDP–I′ is given by

$$I(\pi) = \sum_{n=0}^{\infty}(\beta)^n\, e_\pi\, r^\varphi(\varphi_n, a_n) \tag{23}$$

where

$$e_\pi = \pi_0\, q^\varphi \pi_1\, q^\varphi \cdots \in Q(A\Phi A\Phi \cdots\,|\,\Phi). \tag{24}$$

We have

THEOREM 4. *MDP–II and MDP–I′ are equivalent in the sense that, for any* $\pi \in \Pi$, $J(\pi) = I(\pi)$.

PROOF. For any $n \geqq 0$ and π, the conditional probability $q_\pi([\bar{\varphi}_n, \bar{a}_{n-1}]\,|\,\varphi_0)$ can be rewritten as

$$q_\pi([\bar{\varphi}_n, \bar{a}_n]\,|\,\varphi_0) = q_\pi([a_0]\,|\,\varphi_0)q_\pi([\varphi_1]\,|\,\varphi_0, a_0)\cdots q_\pi([a_n]\,|\,\varphi_0, \bar{\varphi}_n, \bar{a}_{n-1}) \tag{25}$$
$$= \pi_0\, q^\varphi \pi_1\, q^\varphi \cdots \pi_n.$$

Hence

$$e_{\pi\varphi}\, r^s(s_n, a_n) = q_\pi([\bar{\varphi}_n, \bar{a}_n] \mid \varphi_0) r^\varphi(\varphi_n, a_n)$$

$$= \pi_0\, q^\varphi \pi_1\, q^\varphi \cdots \pi_n\, r^\varphi(\varphi_n, a_n) \tag{26}$$

$$= e_\pi\, r^\varphi(\varphi_n, a_n).$$

Therefore

$$J(\pi) = \sum_{n=0}^{\infty} (\beta)^n\, e_{\pi\varphi}\, r^s(s_n, a_n) \tag{27}$$

$$= \sum_{n=0}^{\infty} (\beta)^n\, e_\pi\, r^\varphi(\varphi_n, a_n) = I(\pi). \quad \square$$

Theorem 4 shows that MDP–II can be transformed to MDP–I. Through this transformation, the discounted case, the positive bounded case and the negative case of MDP–II correspond to those of MDP–I; if r^s is bounded $r^\varphi = q^P r^s$ is also bounded. Therefore, the results obtained for MDP–I can readily be translated to the results for MDP–II by only replacing the word "policy" by "I-policy". For instance, we have

THEOREM 5. (*The discounted case of MDP–II*).

(a) *For any* $p \in P(\Phi)$, $\varepsilon > 0$, *there exists a* (p, ε)*-optimal stationary I-policy.*
(b) *If A is finite, then there exists an optimal stationary I-policy.*
(c) *If there exists an optimal policy, then there exists an optimal stationary I-policy.*

PROOF. Taking Theorem 4 into account, (a), (b) and (c) correspond to Theorems 8.1, 9.1 (b) and 8.3 of [10]. $\square$

With a little modification, all contents of this paper except for Theorem 5 are valid also for the case where q^m, q^s and/or r^s vary with the stage n, including the case of finite planning horizon ($r^s(s_n, a_n) = 0$ for $n > n_f$ where n_f is the number of the final stage).

For MDP's with incomplete state observation where the sets of states and observation signals are finite and where the planning horizon is finite, Åström [2] has obtained a theorem similar to Theorem 4. However, his method of proof does not work in the case of MDP–II.

When the total expected return $J(\pi)(\varphi_0)$ in MDP–II is infinite or undefined, the problem formulated in section 3 may not be well defined. Even in such cases, however, Theorem 4 holds in the sense that $J(\pi)(\varphi_0)$ is infinite or undefined if and only if $I(\pi)(\varphi_0)$ is infinite or undefined respectively. This is obvious from the proof of Theorem 4.

Acknowledgment. We are grateful to Professor H. J. Kushner of Brown University for valuable discussion concerning Theorem 2. We wish to thank Professor N. Furukawa of Kyusyu University, Professor H. Akashi and Professor M. Matsumura of Kyoto University for their helpful advice. We also thank the referee for his various suggestions for improving the original version of the paper.

REFERENCES

[1] AOKI, M. (1965). Optimal control of partially observable Markovian systems. *J. Franklin Inst.* **280** 367–386.
[2] ÅSTRÖM, K. J. (1965). Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.* **10** 174–205.
[3] BELLMAN, R. (1968). New classes of stochastic control processes. *J. Math. Anal. Appl.* **22** 602–617.
[4] BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.
[5] BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.
[6] DUBINS, L. and FREEDMAN, D. (1964). Measurable sets of measures. *Pacific J. Math.* **14** 1211–1222.
[7] DYNKIN, E. B. (1965). Controlled random sequences. *Theor. Probability Appl.* **10** 1–14.
[8] FERGUSON, T. S. (1967). *Mathematical Statistics, a Decision Theoretic Approach.* Academic Press, New York.
[9] HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes.* Wiley, New York.
[10] STRAUCH, R. E. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37** 871–890.