The Annals of Mathematical Statistics
1972, Vol. 43, No. 5, 1612-1622

MARKOVIAN DECISION PROCESSES WITH COMPACT ACTION SPACES

By Nagata Furukawa

Kyushu University

We consider the problem of maximizing the expectation of the discounted total reward in Markovian decision processes with arbitrary state space and compact action space varying with the state. We get the existence theorem for a (p, ε)-optimal stationary policy, and the relation between the optimality of a policy and the optimality equation. Assuming the action space is a compact subset of n-dimensional Euclidean space, the existence of an optimal stationary policy is established, and an algorithm is obtained for finding the optimal policy. The last two facts are based on the Borel implicit function lemma given in this paper.

1. Introduction. We shall be concerned with an optimization problem in a Markovian decision process specified by S, $\{A(s), s \in S\}$, q, r, β, where S is a nonempty Borel subset of a Polish space, the set of states of some system, for each s, $A(s)$ is a nonempty subset of a compact metric space A, the set of actions feasible at state s, q is a conditional probability measure on S given $S \times A$, the law of motion of the system, r is a bounded Baire function on $S \times A$, the immediate reward function, and $0 \leqq \beta < 1$, the discount factor.

A policy π is a sequence $\{\pi_1, \pi_2, \cdots\}$, where π_n is a conditional probability measure on A given the previous history $(s_1, a_1, \cdots, a_{n-1}, s_n)$, such that $\pi_n(A(s_n) \mid s_1, a_1, \cdots, s_n) = 1$ for all $s_1, a_1, \cdots, s_n$. A Markov policy is a sequence $\{f_1, f_2, \cdots\}$, where each f_n is Borel measurable function from S to A such that $f_n(s) \in A(s)$ for all $s \in S$. A stationary policy is a Markov policy in which $f_n = f$ for some Borel measurable function f for all n. A policy π associates each initial s with the average of the total discounted reward over the infinite future, $I(\pi)(s)$. A policy π^* is called optimal if $I(\pi^*)(s) \geqq I(\pi)(s)$ for all policies π and all $s \in S$. Our problem, then, is to find an optimal policy.

In general there may not always exist an optimal policy. Blackwell [1], Strauch [10], and Maitra [8] have made studies of this problem in the case when $A(s)$ is assumed to be independent of s. Blackwell [1] has shown that in the case of a finite action space there always exists an optimal stationary policy. Maitra [8] has given a sufficient condition for the existence of an optimal stationary policy in the case of a compact action space.

Special references have to be made to the work of Dubins and Savage [3], Strauch [11] and Sudderth [12]. In our formulation, being given a policy we can compute the transition probabilities of the states at each stage from the sequence of previous states. The previous papers are based on the transition probabilities of the states, but the actions between the adjacent states are deleted

Received December 22, 1970; revised March 7, 1972.

from their formulation. Hence the existence of a good strategy in their sense cannot easily be translated in terms of our formulation, where the actions are taken into account explicitly.

In this paper we treat the case of a compact action space varying with the state. In Section 3 we give the relation between the optimality of a policy and the optimality equation. Assuming the action space is a compact subset of n-dimensional Euclidean space, we give, in Section 4, a Borel implicit function lemma and the existence theorem for an optimal stationary policy, and, in Section 5, we establish an algorithm for finding an optimal policy, a generalization of the policy improvement routine by Howard [4].

2. Preliminaries. In this section we develop the basic notation and definitions to be used throughout the paper.

First we give general probabilistic notation and definitions following closely those of [1]. By a Borel set we mean a Borel subset of some complete separable metric space. By a probability measure on a nonempty Borel set X we mean a probability measure defined over the Borel field of X, and the set of all probability measures on X is denoted by $P(X)$. For any nonempty Borel sets X, Y, $Q(Y \mid X)$ is the set of all conditional probability measures $q(\cdot \mid \cdot)$ such that for each $x \in X$, $q(\cdot \mid x)$ is a probability measure on Y, and for each Borel subset $B \subset Y$, $q(B \mid \cdot)$ is a real-valued Borel function on X. For any nonempty Borel set X, $M(X)$ denotes the set of all bounded Baire functions on X. If $u, v \in M(X)$, $u \geqq v$ means $u(x) \geqq v(x)$ for all $x \in X$. For any $u \in M(XY)$, where XY is the product space of X and Y, and any $q \in Q(Y \mid X)$, qu denotes the element of $M(X)$ whose value at $x_0 \in X$ is given by

$$qu(x_0) = \int_Y u(x_0, y)\, dq(y \mid x_0) .$$

We extend the above notation in an obvious way to a finite or countable sequence of nonempty Borel sets. The details are omitted.

When we speak of the carrier of a probability measure, we use the following notation. For any nonempty Borel sets X, Y, and for any set-valued function $F(\cdot)$ defined on X, whose value at each $x \in X$ is a Borel subset of Y, $Q(\{F(x)\} \mid X)$ consists of all elements of $Q(Y \mid X)$ satisfying that $q(F(x) \mid x) = 1$ for each $x \in X$. In general, for any nonempty Borel sets $X_1, X_2 \cdots, X_{n+1}$, and for any function $F(\cdot)$ defined on X_n, whose value at each $x_n \in X_n$ is a Borel subset of X_{n+1}, $q \in Q(\{F(x_n)\} \mid X_1, X_2 \cdots X_n)$ iff $q \in Q(X_{n+1} \mid X_1, X_2 \cdots X_n)$ and for each $x_i \in X_i$ $(i = 1, 2, \cdots, n)$, $q(F(x_n) \mid x_1, x_2, \cdots, x_n) = 1$.

We now define an optimization problem for Markov decision processes. Our *optimization problem* is specified by S, $\{A(s); s \in S\}$, q, r, β, where S is a nonempty Borel set, the *set of states* of a system, for each $s \in S$, $A(s)$ is a nonempty Borel subset of some compact metric space A, the *set of actions feasible* to us at state s, q is an element of $Q(S \mid SA)$, the *law of motion* of the system, r is an element of $M(SA)$, the *immediate reward function*, and $0 \leqq \beta < 1$, the *discount factor*. A

policy π is a sequence $\{\pi_1, \pi_2, \cdots\}$, where each $\pi_n \in Q(\{A(s)\} \mid H_n)$ and $H_n = SASA \cdots S(2n-1$ factors). Here, $Q(\{A(s)\} \mid H_n)$ is an ambiguous notation, and precisely speaking $Q(\{A(s)\} \mid H_n)$ means $Q(\{A(x_{2n-1})\} \mid X_1 X_2 \cdots X_{2n-1})$ with letting $X_{2i-1} = S$ and $X_{2i} = A$ for $i = 1, 2, \cdots, n$. A policy π is *Markov*, if for each n, there is a Borel measurable function f_n from S into A such that $f_n(s) \in A(s)$ for all s and $\pi_n(\cdot \mid s_1, a_1, \cdots, s_n) = \delta(f_n(s_n))$, and then a Markov policy is denoted by $\{f_1, f_2, \cdots\}$. A Borel measurable function f from S into A such that $f(s) \in A(s)$ for all $s \in S$ define a policy: when in state s, take action $f(s)$ independently both of the time and the previous history. Such policies will be called *stationary*, and will be denoted by f^∞.

An expected discounted total reward from a policy $\pi = \{\pi_1, \pi_2, \cdots\}$ is given by

$$I(\pi) = e_\pi[\textstyle\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n)] ,$$

where $e_\pi = \pi_1 q \pi_2 q \cdots$. For any $p \in P(S)$, and $\varepsilon > 0$, a policy π^* will be called (p, ε)-*optimal* if $p\{I(\pi^*) \geqq I(\pi) - \varepsilon\} = 1$ for all policies π. A policy π^* will be called ε-*optimal* if $I(\pi^*) \geqq I(\pi) - \varepsilon$ for all policies π. A policy π^* will be called *optimal* if $I(\pi^*) \geqq I(\pi)$ for all policies π.

3. (p, ε)-optimal stationary policy and optimality equation. We shall prepare the lemmas concerning the selector in a Polish space, which are fundamental to the existence of a (p, ε)-optimal stationary policy.

Let X be a nonempty Borel set, $\mathscr{B}(X)$ the Borel field on X, Y a compact metric space, and let 2^Y denote the family of all nonempty closed subsets of Y following the notation in [5]. We denote the Hausdorff metric in 2^Y by h. It is, then, well known that the metric space $(2^Y, h)$ is compact. Let $\mathscr{B}(2^Y)$ denote the Borel field on 2^Y generated by the Hausdorff metric h. A function $F(\cdot): X \to 2^Y$ will be called measurable relative to $\mathscr{B}(X)$ and $\mathscr{B}(2^Y)$, if for every $B \in \mathscr{B}(2^Y)$, $F^{-1}(B) \in \mathscr{B}(X)$, and then we will write simply $F(\cdot) \in \mathscr{B}(X)/\mathscr{B}(2^Y)$.

LEMMA 3.1. *Suppose* $F(\cdot) \in \mathscr{B}(X)/\mathscr{B}(2^Y)$, *then*

$$\{(x, y);\ x \in X,\ y \in F(x)\} \in \mathscr{B}(X) \times \mathscr{B}(Y) .$$

PROOF. It is apparent that

$$\{(x, y);\ x \in X,\ y \in F(x)\} = \{(x, y);\ \rho(y, F(x)) = 0,\ x \in X\}$$

where ρ denotes the distance between a point and a set in Y.

Let us define $g(x, y) = \rho(y, F(x))$. Then, since $\rho(p, A)$ is continuous in $(p, A) \in Y \times 2^Y$ by Theorem 2 in Section 42—V of [6], it can be readily seen from Theorem 2 in Section 31—VI of [5] that g is $\mathscr{B}(X) \times \mathscr{B}(Y)$-measurable. Thus $\{(x, y);\ x \in X, y \in F(x)\} = g^{-1}(0) \cap XY$ which belongs to $\mathscr{B}(X) \times \mathscr{B}(Y)$. This completes the proof.

The following proposition is direct from Theorem 2 of [2].

PROPOSITION 3.1. *For any* $q \in Q(\{F(x)\} \mid X)$ *and any set* $\Gamma \in \mathscr{B}(X) \times \mathscr{B}(Y)$ *such that*

$$q(\Gamma_x \mid x) > 0 \qquad \textit{for all } x \in X$$

and

$$\Gamma_x \subset F(x) \qquad \text{for all } x \in X,$$

where Γ_x *denotes the x-section of* Γ, *there is a* $\mathscr{B}(X)$*-measurable function* f *whose graph is a subset of* Γ, *i.e.*, $(x, f(x)) \in \Gamma$ *for all* $x \in X$.

LEMMA 3.2. *Assume* $F(\cdot) \in \mathscr{B}(X)/\mathscr{B}(2^Y)$. *Then for any* $q \in Q(\{F(x)\} \mid X)$, *any* $u \in M(Z)$, *where* Z *denotes the set* $\{(x, y);\ x \in X,\ y \in F(x)\}$, *and any* $\varepsilon > 0$, (i) *there is a* $\mathscr{B}(X)$*-measurable* f_1, *whose graph is a subset of* Z, *satisfying*

$$f_1 u \geqq qu, \qquad \text{and}$$

(ii) *there is a* $\mathscr{B}(X)$*-measurable* f_2, *whose graph is a subset of* Z, *satisfying*

$$q(\{y \in F(x_0);\ u(x_0, y) \leqq u(x_0, f_2(x_0)) + \varepsilon\} \mid x_0) = 1 \qquad \text{for all } x_0 \in X.$$

PROOF.

(i) Let D be the set $\{(x, y) \in Z;\ u(x, y) \geqq qu(x)\}$. Since $Z \in \mathscr{B}(X) \times \mathscr{B}(Y)$ by Lemma 3.1, then it follows that $D \in \mathscr{B}(X) \times \mathscr{B}(Y)$ for any $u \in M(Z)$. It is easily verified that $q(D_x \mid x) > 0$ for all $x \in X$, where D_x denotes the x-section of D. Thus (i) follows directly from Proposition 3.1.

(ii) Let $\mathbf{u}(y) = \sup_{x \in Z_y} u(x, y)$, where Z_y stand for the y-section of Z.

It can be easily shown that $\mathbf{u}(y)$ is universally measurable by using Kuratowski's theorem that analytic sets are universally measurable (cf. [5]). Hence, for each $x \in X$, there is a measurable function $u_0(\cdot)_{(x)}$ such that $u_0(y)_{(x)} = \mathbf{u}(y)$ a.e. $(q(\cdot \mid x))$. Let

$$\nu(x) = \sup\{\text{rational } r;\ q(\{y;\ u_0(y)(x) > r\} \mid x) > 0\},$$

then

$$q(\{y;\ \mathbf{u}(y) > \nu(x)\} \mid x) = 0 \qquad \text{for all } x \in X. \tag{3.1}$$

It is apparent that for each real λ,

$$\{x \mid \nu(x) > \lambda\} = \bigcup_{r > \lambda} \{x \mid q(\{y;\ \mathbf{u}(y) > r\} \mid x) > 0\}, \tag{3.2}$$

where the union is taken over rational number r's. Since the right side of (3.2) is a measurable set, $\nu(x)$ is measurable.

Let for $\varepsilon > 0$,

$$D(\varepsilon) = \{(x, y) \in Z;\ u(x, y) > \nu(x) - \varepsilon\}$$

and

$$\Gamma(\varepsilon) = \{(x, y) \in Z;\ q(\{y^*;\ u(x, y^*) \leqq u(x, y) + \varepsilon\} \mid x) = 1\}.$$

If $(x_0, y_0) \in D(\varepsilon)$, by virtue of the definition of $D(\varepsilon)$ and (3.1) it follows that

$$q(\{y;\ u(x_0, y) \leqq u(x_0, y_0) + \varepsilon\} \mid x_0) + 1,$$

which implies $(x_0, y_0) \in \Gamma(\varepsilon)$, i.e., $D(\varepsilon) \subset \Gamma(\varepsilon)$. Thus, in order to prove (ii) of this lemma, it is sufficient to show the existence of a measurable function f whose graph is a subset of $D(\varepsilon)$. This, however, can be established by appealing to Proposition 3.1.

ASSUMPTION (I). For each $s \in S$, $A(s) \in 2^A$, where 2^A denotes the family of all nonempty closed subsets of a compact metric space A, and $A(\cdot) \in \mathscr{B}(S)/\mathscr{B}(2^A)$.

THEOREM 3.1. *Let Assumption* (I) *be satisfied. Then for any* $p \in P(S)$ *and any* $\varepsilon > 0$, *there is a* (p, ε)*-optimal Markov policy.*

PROOF. By virtue of Lemma 3.2 (ii), the proof follows in the same way as in the Corollary to Theorem 2 of [1], and the details are omitted.

DEFINITION 3.1. With any Borel function f from S into A such that $f(s) \in A(s)$ for all $s \in S$, we associate the operator L_f, mapping $M(S)$ into $M(S)$, defined by

$$L_f u = fq(r + \beta u) .$$

In particular, with any $a \in A(s)$ associate the operator L_a defined by

$$L_a u(s) = \int_S [r(s, a) + \beta u(t)]\, dq(t \mid s, a) .$$

It is apparent that L_f is monotone, and is a contraction operator with contraction coefficient β.

DEFINITION 3.2. We say that a $u \in M(S)$ satisfies the *optimality equation*, if for each $s \in S$,

$$u(s) = \sup_{a \in A(s)} L_a u(s) .$$

We can prove the following theorem in the same way as in Theorem 6 of [1].

THEOREM 3.2. *Let Assumption* (I) *be satisfied.*

(a) *For any* $p \in P(S)$ *and any* $\varepsilon > 0$, *there is a* (p, ε)*-optimal stationary policy.*

(b) *For any* $\varepsilon \geqq 0$, *if there is an* ε*-optimal policy, there is an* $\varepsilon/(1 - \beta)$*-optimal stationary policy.*

(c) *A policy* π *is optimal if and only if its return* $I(\pi)$ *satisfies the optimality equation.*

4. Existence of an optimal stationary policy. In this section we shall give a sufficient condition for the existence of an optimal stationary policy. In the preceding sections we assumed the action space to be an arbitrary compact metric space, but henceforth we will restrict it to be a compact subset of R^n (Assumption (I*)), while the state space S remains to be an arbitrary Borel set. The proofs of Lemmas 4.1 and 4.3 follow the line similar to those of Theorems 1 and 2 in [9].

Let X^m denote the family of all compact subsets of R^m, and $\mathscr{B}(X^m)$ the Borel field on X^m generated by the Hausdorff metric h.

LEMMA 4.1. *Let* A *be a compact subset of* R^n. *Suppose* $u(s, a)$ *satisfies that*

(i) $u(s, a)$ *is a bounded function from* SA *to* R^m,

(ii) $u(s, a)$ *is Borel measurable in* s *for each fixed* $a \in A$,

(iii) $u(s, a)$ *is continuous in* a *for each fixed* $s \in S$.

Then $U(s) = \{x;\ x = u(s, a),\ a \in A\}$ *is a function from* S *to* X^m *such that* $U(\cdot) \in \mathscr{B}(S)/\mathscr{B}(X^m)$.

ASSUMPTION (I*). For each $s \in S$, $A(s) \in 2^A$, and $A(\cdot) \in \mathscr{B}(S)/\mathscr{B}(2^A)$, where A is a compact subset of R^n.

LEMMA 4.2. *Let Assumption* (I*) *be satisfied. Let* $u(s, a)$ *be as in Lemma* 4.1. *Then* $\tilde{U}(s) = \{x;\ x = u(s, a),\ a \in A(s)\}$ *is a function from* S *to* X^m *such that* $\tilde{U}(\cdot) \in \mathscr{B}(S)/\mathscr{B}(X^m)$.

PROOF. By virtue of Assumption (I*) there is a sequence $\{A^n(s)\}$ of 2^A-valued $\mathscr{B}(S)$-simple functions such that $h(A(s), A^n(s)) \to 0$ as $n \to \infty$ for each $s \in S$, which implies that

$$h(\tilde{U}(s), \tilde{U}^n(s)) \to 0 \qquad \text{as } n \to \infty \text{ for each } s \in S, \tag{4.1}$$

where $\tilde{U}^n(s) = \{x;\ x = u(s, a),\ a \in A^n(s)\}$. Let I_E denote the characteristic function of a set E, and let us define a set K multiplied by 0 or 1 as follows: $K \times 1 = K$ and $K \times 0 =$ empty set. Then $A^n(s)$ is written as

$$A^n(s) = \sum_{i=1}^{k_n} K_{ni} I_{S_{ni}}(s),$$

where $K_{ni} \in 2^A$, $S_{ni} \in \mathscr{B}(S)$. By putting S_{ni}, K_{ni} into S, A of Lemma 4.1, respectively, we have $\tilde{U}^n(\cdot) \in \mathscr{B}(S)/\mathscr{B}(X^m)$. Hence it is clear from (4.1) that $\tilde{U}(\cdot) \in \mathscr{B}(S)/\mathscr{B}(X^m)$.

Now we shall give the definition of the lexicographic maximum of a set, following [9].

DEFINITION 4.1. Let G be any compact subset of n-dimensional real Euclidean space R^n, and $(x_1, x_2, \cdots, x_n)$ an orthonormal basis of R^n. We define a sequence $\{G_{x_1, x_2, \cdots, x_i}\}$ of subsets of G inductively as follows:

$$G_{x_1} = \{x;\ x \in G,\ (x_1 \mid x) = \max_{y \in G}(x_1 \mid y)\},$$

$$G_{x_1, x_2, \cdots, x_i} = \{x;\ x \in G_{x_1, x_2, \cdots, x_{i-1}},\ (x_i \mid x) = \max_{y \in G_{x_1, x_2, \cdots, x_{i-1}}}(x_i \mid y)\} \quad \text{for } i = 2, \cdots, n,$$

where $(\cdot \mid \cdot)$ means the inner product in R^n. Then $e(G) = G_{x_1, x_2, \cdots, x_n}$ reduces to a single point in R^n which will be called the *lexicographic maximum* of G.

LEMMA 4.3. *Let* A *be a compact subset of* R^n. *Suppose* $G(s)$ *is a function from* S *to* X^n *such that* $G(s) \subset A$ *for all* $s \in S$ *and* $G(\cdot) \in \mathscr{B}(S)/\mathscr{B}(X^n)$. *Then the function* $e(G(s)) = G_{x_1, x_2, \cdots, x_n}(s) : S \to R^n$ *is Borel measurable in* s.

PROOF. By induction, in order to prove the lemma it is enough to show that $G_{x_1}(s)$ is Borel measurable in s. For the sake of simplicity we put $x = x_1$. From the assumption there is a sequence $\{G^m(s)\}$ of X^n-valued $\mathscr{B}(S)$-simple functions such that

$$\{G^m(s)\} \to G(s) \qquad \text{for all } s \in S. \tag{4.2}$$

We let $G^m(s) = \sum_{i=1}^{k(m)} K_i I S_{mi}(s)$, where $K_i \in X^n$, $S_{mi} \in \mathscr{B}(S)$, $S_{mi} \cap S_{mj} = \phi$, $(i \neq j)$, $\sum_{i=1}^{k(m)} S_{mi} = S$. For each m and any $s \in S$, letting $i_m(s)$ be the number

such that $1 \leqq i_m(s) \leqq k(m)$ and $s \in S_{mi_m(s)}$, without loss of generality we may assume that $\{S_{mi_m(s)}\}$ is decreasing in m for each s. Let

$$F^m(s) = \overline{\bigcup_{t \in S_{mi_m(s)}} G_x(t)}$$

for $s \in S$ and $m = 1, 2, \cdots$, then it is clear that $F^m(s)$, $m = 1, 2, \cdots$, is an X^n-valued $\mathscr{B}(S)$-simple function, and that $\{F^m(s)\}$ is decreasing in m for each s. Putting $F(s) = \lim_{m\to\infty} F^m(s)$, we have

$$G_x(s) \subset F(s) \qquad \text{for } s \in S. \tag{4.3}$$

Since we may assume that the convergence in (4.2) is uniform from the hypotheses stated in the lemma, for any $\varepsilon > 0$ there is a number M such that for $t \in S_{mi_m(s)}$ and for $m \geqq M$,

$$\begin{aligned} h(G(s), G(t)) &\leqq h(G(s), G^m(t)) + h(G^m(t), G(t)) \\ &= h(G(s), G^m(s)) + h(G^m(t), G(t)) \\ &< \varepsilon . \end{aligned} \tag{4.4}$$

Hence it follows that $r(p, G(s)) < \varepsilon$ for $p \in G_x(t)$, $t \in S_{mi_m(s)}$, $m \geqq M$, therefore $r(p, G(s)) \leqq \varepsilon$ for $p \in F(s)$.
Since ε is arbitrary, it holds that

$$F(s) \subset G(s) \qquad \text{for all } s \in S. \tag{4.5}$$

We get now from (4.4) that for $t \in S_{mi_m(s)}$ and $m \geqq M$,

$$|\max_{z \in G(t)}(x \mid z) - \max_{z \in G(s)}(x \mid z)| \leqq h(G(s), G(t)) < \varepsilon ,$$

which implies that for $y \in F^m(s)$ and $m \geqq M$,

$$|(x \mid y) - \max_{z \in G(s)}(x \mid z)| < \varepsilon .$$

Since ε is arbitrary, the latter implies that

$$|(x \mid y) - \max_{z \in G(s)}(x \mid z)| = 0 \qquad \text{for } y \in F(s) \text{ and } s \in S. \tag{4.6}$$

We have from (4.5), (4.6) that $F(s) \subset G_x(s)$ for $s \in S$, which together with (4.3) implies $F(s) = G_x(s)$ for $s \in S$. But $F(s)$ is Borel measurable, therefore $G_x(s)$ is also Borel measurable. This completes the proof.

LEMMA 4.4. *Let A be a compact subset of R^n. Let $u(s, a)$ be as in Lemma* 4.1. *And let $U(s) = \{x;\ x = u(s, a),\ a \in A\}$. Then for any Borel measurable $\phi\colon S \to R^m$ such that $\phi(s) \in U(s)$ for all $s \in S$, it holds that $G(s) = \{a;\ a \in A,\ u(s, a) = \phi(s)\}$ is a Borel measurable map from S to X^n, i.e. $G(\cdot) \in \mathscr{B}(S)/\mathscr{B}(X^n)$.*

PROOF. This lemma can be easily verified by appealing to the proof of Theorem 2—(II) of [9].

LEMMA 4.5. *Let Assumption* (I*) *be satisfied. Let $u(s, a)$ be as in Lemma* 4.1. *And let $\tilde{U}(s) = \{x;\ x = u(s, a),\ a \in A(s)\}$. Then for any Borel measurable $\phi\colon S \to R^m$ such that $\phi(s) \in \tilde{U}(s)$ for all $s \in S$, it holds that $\tilde{G}(s) = \{a;\ a \in A(s),\ u(s, a) = \phi(s)\}$ is a Borel measurable map from S to X^n, i.e. $\tilde{G}(\cdot) \in \mathscr{B}(S)/\mathscr{B}(X^n)$.*

PROOF. This lemma can be proved in the same way as in Lemma 4.2, by taking a sequence of 2^A-valued $\mathscr{B}(S)$-simple functions which converges to $A(s)$.

The following lemma falls under the category of selection theorems (cf. [7]), and may be anticipated on the basis of other general theorems. The present author, however, can locate neither such an explicit statement like this one nor the proof in the literatures.

LEMMA 4.6. (*Borel implicit function lemma*). *Let Assumption* (I*) *be satisfied. Let* $u(s, a)$ *be as in Lemma* 4.1. *And let* $\tilde{U}(s) = \{x; x = u(s, a), a \in A(s)\}$. *Then for any Borel measurable* $\phi: S \to R^m$ *such that* $\phi(s) \in \tilde{U}(s)$ *for all* $s \in S$, *there is a Borel measurable* $f: S \to A$ *satisfying*

$$f(s) \in A(s) \qquad \text{for all } s \in S \tag{4.7}$$

and

$$\phi(s) = u(s, f(s)) \qquad \text{for all } s \in S. \tag{4.8}$$

PROOF. By virtue of Lemma 4.5, $\tilde{G}(s) = \{a; a \in A(s), u(s, a) = \phi(s)\}$ is a Borel measurable map from S to X^n.

Letting $f(s) = e(\tilde{G}(s))$, it is clear by Lemma 4.3 that $f(s)$ is a Borel measurable map satisfying (4.7) and (4.8).

THEOREM 4.1. *Let Assumption* (I*) *be satisfied. Let* $u(s, a)$ *be a bounded function from* SA *to* R^1 *satisfying* (ii), (iii) *in Lemma* 4.1. *Then there is a Borel measurable function* f *from* S *to* A *such that*

$$f(s) \in A(s) \qquad \text{for all } s \in S \tag{4.9}$$

and

$$u(s, f(s)) = \max_{a \in A(s)} u(s, a) \qquad \text{for all } s \in S. \tag{4.10}$$

PROOF. Because $u(s, a) \in R^1$, it holds that $\max_{a \in A(s)} u(s, a) = e(\tilde{U}(s))$, where $\tilde{U}(s) = \{x; x = u(s, a), a \in A(s)\}$.

Since $\tilde{U}(\cdot) \in \mathscr{B}(S)/\mathscr{B}(X^m)$ by Lemma 4.2, it follows from Lemma 4.3 that $e(\tilde{U}(s))$ is Borel measurable, which implies $\phi(s) = \max_{a \in A(s)} u(s, a)$ is Borel measurable. It is apparent that $\phi(s) \in \tilde{U}(s)$ for all $s \in S$. Thus by virtue of Lemma 4.6 there is a Borel measurable function f satisfying (4.9) and (4.10), which completes the proof.

Now we put the following assumptions.

ASSUMPTION (II). $r(s, a)$ is continuous in $a \in A$ for each fixed $s \in S$.

ASSUMPTION (III). For each fixed $s \in S$ and for any $w \in M(S)$, $\int_S w(\cdot)\, dq(\cdot \mid s, a)$ is continuous in $a \in A$.

DEFINITION 4.2. The contraction operator T; $M(S) \to M(S)$ is given by

$$Tw(s) = \max_{a \in A(s)} [\int_S \{r(s, a) + \beta w(t)\}\, dq(t \mid s, a)].$$

We close this section with the following theorem concerning the existence of an optimal stationary policy.

THEOREM 4.2. *Let Assumptions* (I*), (II), (III) *be satisfied. Then there exists an optimal stationary policy.*

PROOF. Let w^* denote the fixed point of the operator T. Because of Theorem 4.1 then there exists a Borel measurable map f from S to A satisfying (4.11) and the equality

$$L_f w^* = T w^* , \tag{4.11}$$

consequently $L_f w^* = w^*$, i.e. w^* is also the fixed point of L_f. On the other hand, by virtue of uniqueness of the fixed point, it follows that $w^* = I(f^\infty)$.

$I(f^\infty)$ satisfies the optimality equation, for from (4.11) we have

$$L_f I(f^\infty) = I(f^\infty) = T I(f^\infty) .$$

Thus it follows from Theorem 3.2 (c) that f^∞ is optimal. This completes the proof.

5. Algorithm for finding an optimal policy. Concerning the method of construction of an optimal policy in a Markov decision process, Howard's policy improvement routine [4] in the case of finite state and finite action is well known.

Our object in this section is to give an algorithm for finding an optimal policy in the decision process of our concern, a generalization of Howard's routine.

Let N denote a Markov kernel on the measurable space $(S, \mathscr{B}(S))$, and E a Markov identity kernel on $(S, \mathscr{B}(S))$.

DEFINITION 5.1. If f is a Baire function on S, and if

$$h = \lim_{n\to\infty} (E + \beta N + \cdots + \beta^n N^n) f$$

is well defined and finite, then f is called a *β-charge* and h is called a *β-Potential* of f, where $0 \leqq \beta < 1$. Consequently it follows that any $f \in M(S)$ is a *β-charge*. We shall call $G = \sum_{n=0}^{\infty} (\beta N)^n$ a *β-Potential kernel.*

LEMMA 5.1. *Let h be any element of $M(S)$, and let*

$$f = (E - \beta N) h ,$$

then h is a β-Potential of f, and $h = Gf$.

PROOF. Since $h = f + \beta N h$, inductively we have

$$h = (E + \beta N + \cdots + \beta^{n-1} N^{n-1}) f + \beta^n N^n h , \qquad n \geqq 1 .$$

Then by virtue of the boundedness of h we have

$$\lim_{n\to\infty} (E + \beta N + \cdots + \beta^{n-1} N^{n-1}) f = h . \tag{5.1}$$

Since the left side of (5.1) is well defined, by monotone convergence we have $Gf = h$.

LEMMA 5.2. *Let π be an arbitrary policy. Suppose f is a Borel measurable function from S into A such that $f(s) \in A(s)$ for all $s \in S$.*

(a) *If* $L_f I(\pi) \geqq I(\pi)$, *then* $I(f^\infty) \geqq I(\pi)$.
(b) *If* $L_f I(\pi) \leqq I(\pi)$, *then* $I(f^\infty) \leqq I(\pi)$.

PROOF. Let

$$J = L_f I(\pi) - I(\pi) ,$$
$$\Delta I = I(f^\infty) - I(\pi) ,$$
$$r_f(s) = r(s, f(s)) ,$$
$$N_f(s, \cdot) = q(\cdot \mid s, f(s)) .$$

Then it follows that

$$J = r_f + \beta N_f I(\pi) - I(\pi)$$

and

$$\begin{aligned} \Delta I &= r_f + \beta N_f I(f^\infty) - I(\pi) \\ &= J + \beta N_f \cdot \Delta I . \end{aligned}$$

Hence

$$J = (E - \beta N_f)\Delta I .$$

Because $\Delta I \in M(S)$, from Lemma 5.1 it follows that ΔI is a β-Potential of J and $\Delta I = G_f J$, where $G_f = \sum_{n=0}^{\infty} (\beta N_f)^n$. By the last statement we have $J \geqq 0$ only if $I(f^\infty) \geqq I(\pi)$, and $J \leqq 0$ only if $I(f^\infty) \leqq I(\pi)$, which completes the proof.

THEOREM 5.1. *Let Assumptions* (I*), (II), (III) *be satisfied. Let* $\{f_n\}_{n=0,1,2\cdots}$ *be defined in the iterative manner*:

(i) *Take an arbitrary Borel measurable map* $f_0 : S \to A$ *such that* $f_0(s) \in A(s)$ *for all* $s \in S$.

(ii) *For each* $n \geqq 0$ *select a Borel measurable map* $f_{n+1} : S \to A$ *such that* $f_{n+1}(s) \in A(s)$ *for all* $s \in S$ *and* $L_{f_{n+1}} I(f_n^\infty) = TI(f_n^\infty)$. *Then we have*

(a) $I(f_0^\infty) \leqq I(f_1^\infty) \leqq \cdots \leqq I(f_n^\infty) \leqq I(f_{n+1}^\infty) \leqq \cdots \uparrow \sup_\pi I(\pi)$,
(b) *if for some* N *it occurs that* $f_N = f_{N+1}$, *then* f_N^∞ *is an optimal policy.*

REMARK. The existence of such f_{n+1} in (ii) is assured by Theorem 4.1.

PROOF. (a) Trivially we have

$$L_{f_{n+1}} I(f_n^\infty) = TI(f_n^\infty) \geqq L_{f_n} I(f_n^\infty) = I(f_n^\infty) .$$

Lemma 5.2 applies and gives for every $n \geqq 0$

$$I(f_{n+1}^\infty) \geqq I(f_n^\infty) .$$

Next we can obtain inductively

$$TI(f_n^\infty) \geqq T^n I(f_1^\infty) \qquad \text{for } n \geqq 1 . \tag{5.2}$$

It is obvious from the proof of Theorem 4.2 that

$$\lim_{n\to\infty} T^n I(f_1^\infty) = w^* = \sup_\pi I(\pi) ,$$

which together with (5.2) implies $\lim_{n\to\infty} TI(f_n^\infty) \geqq \sup_\pi I(\pi)$, and the converse inequality is trivally true. Hence

$$\lim_{n\to\infty} TI(f_n^\infty) = \sup_\pi I(\pi) . \tag{5.3}$$

Since on the other hand it holds that

$$I(f_{n+1}^{\infty}) = L_{f_{n+1}} I(f_{n+1}^{\infty}) \geqq L_{f_{n+1}} I(f_n^{\infty}) = TI(f_n^{\infty}),$$

by (5.3) we get $\lim_{n\to\infty} I(f_n^{\infty}) = \sup_{\pi} I(\pi)$.

(b) Suppose that $f_N = f_{N+1}$ and that there is a Borel measurable $\hat{f}$ satisfying $I(\hat{f}^{\infty}) \geqq I(f_{N+1}^{\infty})$. Then we have

$$L_{f_{N+1}} I(f_N^{\infty}) = TI(f_N^{\infty}) \geqq L_{\hat{f}} I(f_N^{\infty}),$$

which implies

$$I(f_{N+1}^{\infty}) \geqq L_{\hat{f}} I(f_{N+1}^{\infty})$$

because $f_N = f_{N+1}$. From Lemma 5.2 it follows that

$$I(\hat{f}^{\infty}) \leqq I(_{N+1}^{\infty}).$$

The converse inequality is true by hypothesis, consequently

$$I(\hat{f}^{\infty}) = I(f_N^{\infty}) = I(f_{N+1}^{\infty}),$$

which implies f_N^{∞} is optimal. This completes the proof.

Acknowledgment. The author wishes to express his hearty thanks to the referee for many useful comments.

REFERENCES

[1] BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.
[2] BLACKWELL, D. and RYLL-NARDZEWSKI, C. (1963). Non-existence of everywhere proper conditional distributions. *Ann. Math. Statist.* **34** 223–225.
[3] DUBINS, L. E. and SAVAGE, L. J. (1965). *How to Gamble if you Must.* McGraw-Hill, New York.
[4] HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes.* Wiley, New York,
[5] KURATOWSKI, K. (1958). *Topologie*, **1** (4th ed.). Warszawa, Poland.
[6] KURATOWSKI, K. (1961). *Topologie*, **2** (3rd ed.). Warszawa, Poland.
[7] KURATOWSKI, K. and RYLL-NARDZEWSKI, C. (1965). A general theorem on selectors. *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys.* **13** 397–403.
[8] MAITRA, A. (1968). Discounted dynamic programming on compact metric spaces. *Sankhyā Ser. A* **30** 211–216.
[9] OLECH, C. (1965). A note concerning set-valued measurable functions. *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys.* **13** 317–321.
[10] STRAUCH, R. E. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37** 871–890.
[11] STRAUCH, R. E. (1967). Measurable gambling houses. *Trans. Amer. Math. Soc.* **126** 64–72.
[12] SUDDERTH, W. D. (1969). On the existence of good stationary strategies. *Trans. Amer. Math. Soc.* **135** 399–414.

DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KYUSHU UNIVERSITY
FUKUOKA, JAPAN