九州大学学術情報リポジトリ
Kyushu University Institutional Repository

AVERAGE REWARD MARKOVIAN DECISION PROCESSES IN THE COMPLETELY ERGODIC CASE

Iwamoto, Seiichi
Department of Mathematics, Kyushu University

https://doi.org/10.5109/13072

出版情報：統計数理研究. 15 (3/4), pp.55-68, 1973-03. Research Association of Statistical Sciences
バージョン：
権利関係：

KYUSHU UNIVERSITY

AVERAGE REWARD MARKOVIAN DECISION PROCESSES IN THE COMPLETELY ERGODIC CASE

By

Seiichi IWAMOTO*

(Received July 15, 1972)

1. Introduction.

We are concerned with average reward Markovian decision processes on the arbitrary state and compact action spaces. In general, Markovian decision processes are determined by four objects; S, A, q and r. S and A are non-empty Borel sets, q is a regular conditional probability on S given SA, and r is a Baire function on SAS. We interpret S as the set of states of some system, and A as the set of actions available to us at each stage. The set of actions is assumed to be independent of the state. When the system is in state s and we take action a, the system moves to a new state s' according to the conditional distribution $q(\cdot \mid s, a)$, and we receive a reward $r(s, a, s')$. The process is then repeated from the new state s' and so on. We then wish to maximize the expected average reward per unit time.

The purpose of this paper is to make studies on the existence of an optimal stationary policy, the policy improvement procedure under Doeblin condition and the relation between the average reward problem and the discounted total reward problem. The works of Howard [4], Derman [2] and Ross [6] were made on the problems, but in their works at least one of the state space or the action space was rather restrictive in contrast with the one in ours.

In Section 2, we shall give the basic notation and definitions to be used throughout this paper. In Section 3, assuming our processes keep up a kind of complete ergodicity resulted from Doeblin condition, we shall give sufficient conditions for the existence of an optimal stationary policy and show that a policy improvement procedure can be effectively used under some additional assumptions. In connection with the existence of an optimal policy, a functional equation fundamental to our average case will be exposed. In Section 4, we shall give the reduction of the average reward problem to the discounted total problem, and the method of successive approximations to solve the fundamental functional equation treated in Section 3. The complete ergodicity is not assumed in the reduction of the average problem to the discounted problem, but in successive approximations it is the consequence of the assumptions just set up.

* Department of Mathematics, Kyushu University, Fukuoka.

2. Probabilistic definitions and the optimization problem.

First we give the probabilistic notation and definitions following closely those of [1]. A Borel set X is a Borel subset of a Polish space. Polish space is a complete separable metric space. Unless otherwise noted, "measurable" means measurable with respect to the σ-field of a Borel subset of X, which is expressed by $\mathscr{B}(X)$. A probability is a probability measure on the measurable space $(X, \mathscr{B}(X))$, and the set of all probabilities on $(X, \mathscr{B}(X))$ is denoted by $P(X)$. If X and Y are non-empty Borel sets, a conditional probability on Y given X is a function $q(\cdot\,|\,\cdot)$ such that for each $x \in X$, $q(\cdot\,|\,x)$ is a probability on $(Y, \mathscr{B}(Y))$ and for each Borel subset B in $\mathscr{B}(Y)$, $q(B\,|\,\cdot)$ is a Baire function on X. $Q(Y\,|\,X)$ denotes the class of all conditional probabilities $q(\cdot\,|\,\cdot)$ defined as above. $\mathscr{B}(X\,|\,Y)$ will denote the set of all measurable function from X to Y. It will be shown that $\mathscr{B}(X\,|\,Y) \subset Q(Y\,|\,X)$. We denote the Cartesian product of X and Y by XY. Furthermore, the Cartesian product space of $X_1, X_2, \cdots$ will be denoted by $X_1X_2, \cdots$, where each X_n is a Borel set. In this paper, the set of all bounded Baire functions on X is denoted by $M(X)$. $C_1(X)$ denotes the set of all bounded upper semi-continuous function of X. Obviously $C_1(X) \subset M(X)$. If $u, v \in M(X)$, $u \geqq v$ means $u(x) \geqq v(x)$ for all $x \in X$, and $\|u\| = \sup_{x \in X} |u(x)|$.

For any $u \in M(XY)$ and any $q \in Q(Y\,|\,X)$, qu denotes the element of $M(X)$ whose value at $x \in X$ is given by

$$qu(x) = \int_Y u(x, y)dq(y\,|\,x).$$

For any $p \in P(X)$, $q \in Q(Y\,|\,X)$, pq is the probability on XY such that for all $u \in M(XY)$, $pq(u) = p(qu)$. Any probability m on XY has a factorization $m = pq$; $p \in P(X)$ is unique and is just the marginal distribution of the first coordinate variable with respect to m, and $q \in Q(Y\,|\,X)$ is not unique and is a version of the conditional distribution of the second coordinate variable given the first.

The above notation may be extended in an obvious way to a finite or countable sequence of non-empty Borel sets $X_1, X_2, \cdots$. If $q_n \in Q(X_{n+1}\,|\,X_1X_2 \cdots X_n)$ for $n \geqq 1$ and $p \in P(X_1)$, $pq_1q_2 \cdots q_n$ is a probability on $X_1X_2 \cdots X_{n+1}$, $pq_1q_2 \cdots$ is a probability on the infinite product space $X_1X_2 \cdots$, $q_2q_3 \in Q(X_3X_4\,|\,X_1X_2)$, for any $u \in M(X_1X_2 \cdots X_{n+1})$, $n \geqq 1$ and any m, $1 \leqq m \leqq n$, $q_m \cdots q_n u \in M(X_1 \cdots X_m)$, 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 $y \in Y$. Thus, for example, if $q \in Q(Y\,|\,X)$ and $u \in M(Y)$, then $qu \in M(X)$, and $q \in Q(Y\,|\,X)$ will also denote the element $q' \in Q(Y\,|\,ZX)$ such that $q'(\cdot\,|\,z, \cdot) = q(\cdot\,|\,\cdot)$ for all $z \in Z$, etc.

A $p \in P(X)$ is degenerate if for some $x \in X$, $p\{x\} = 1$, and will sometimes be denoted by $\delta(x)$. A $q \in Q(Y\,|\,X)$ is degenerate if $q(\cdot\,|\,x)$ is degenerate for each $x \in X$. The degenerate q are exactly those which there is a measurable function f from X to Y such that $q(\cdot\,|\,x) = \delta(f(x))$ for each $x \in X$. That is to say the degenerate q is an element of $\mathscr{B}(X\,|\,Y)$. Any such $f \in \mathscr{B}(X\,|\,Y)$ will also denote its associated degenerate q, so that, for any $u \in M(XY)$, $qu(x) = u(x, f(x))$ for any $x \in X$.

Next we shall formulate our optimization problem in Markovian decision processes. Our *optimization problem* is defined by four tuple (S, A, q, r), where S, the *set of states* of some system, is any non-empty Borel set, A, the *set of actions* available to us, is also any non-empty Borel set, $q \in Q(S|SA)$ is the *law of the motion* of the system—when the system is in s and action a is chosen, the system moves to next state s' according to the conditional distribution $q(\cdot|s, a)$; and $r \in M(SAS)$ is a *reward function*—when the system is in state s, choose action a and s' is a new state, we will receive a reward $r(s, a, s')$. We wish to maximize the expected average reward per unit time, when the process starts on state s.

A *policy* π is a sequence $\{\pi_1, \pi_2, \cdots\}$, where each $\pi_n \in Q(A|H_n)$ and $H_n = SA \cdots S$ $(2n-1$ factors) is the set of possible histories of the system when the n-th act must be chosen. The set of all policies is denoted by Π. A policy π is (non-randomized) *Markov* if each π_n is a degenerate element of $Q(A|S)$, i.e. $\pi = \{f_1, f_2, \cdots\}$, where $f_i \in \mathscr{B}(S|A)$, and is (non-randomized) *stationary* if there is a $f \in \mathscr{B}(S|A)$ such that $\pi_n = f$ for all n. The stationary policy defined by f is denoted by $f^{(\infty)}$. The set of all stationary policies is denoted by Π_s. It is noted that $\Pi_s \subset \Pi$.

Any policy, together with law of the motion q of the system, defines for each initial state s a conditional probability distribution on the set $X = ASAS \cdots$ of the future of the system, i.e. it defines an element of $Q(X|S)$, namely $e_\pi = \pi_1 q \pi_2 q \cdots$. Denote the coordinate functions on SX by $s_1, a_1, s_2, a_2, \cdots$, so our reward on the n-th day, as a function of the history of the system, is $r(s_n, a_n, s_{n+1})$, and total reward by the n-th day also, as a function of the history, is $\sum_{k=1}^{n} r(s_k, a_k, s_{k+1})$. The limiting average expected reward $J(\pi)$ from the policy π, as a function of the initial state s_1, is then

$$J(\pi)(s_1) = \varliminf_{n\to\infty} \frac{1}{n} e_\pi\Big[\sum_{k=1}^{n} r(s_k, a_k, s_{k+1})\Big].$$

It should be noted that the ambiguous notations are used.

For any $v \in M(S)$ and $0 \leqq \beta \leqq 1$, we denote by $I_n^\beta(\pi, v)$ the expected β-discounted total reward if we terminate after the n-th stage and receive a terminal reward $v(s_{n+1})$ at the terminal state. Thus,

$$I_n^\beta(\pi, v)(s_1) = e_\pi\Big[\sum_{k=1}^{n} \beta^{k-1} r(s_k, a_k, s_{k+1}) + \beta^n v(s_{n+1})\Big].$$

Let $I_n^\beta(\pi) = I_n^\beta(\pi, 0)$ for $0 \leqq \beta \leqq 1$ and $I_n(\pi) = I_n^1(\pi)$. Then, it is clear that

$$J(\pi)(s_1) = \varliminf_{n\to\infty} \frac{1}{n} I_n(\pi)(s_1)$$

and that for all $\pi \in \Pi$

$$|J(\pi)| \leqq \|r\|.$$

Furthermore, let $I^\beta(\pi) = \lim_{n\to\infty} I_n^\beta(\pi)$ for $0 \leqq \beta < 1$. Note that this limit always exists.

3. Existence of the stationary policy and the policy improvement procedure.

In this section we are concerned with the sufficient conditions for the existence of an optimal stationary policy in the average reward Markovian decision processes on general state and action spaces. It will be shown that Howard's policy improvement can be generalized to the case of the general state and action spaces under some weak conditions introduced by Maitra [5].

Let $\{\mu_n;\ n\geqq 1\}$ be a sequence in $P(S)$ and $\mu\in P(S)$. If $\mu_n(E)$ converges to $\mu(E)$ uniformly in $E\in\mathscr{B}(S)$ as $n\to\infty$, then we shall say that μ_n converges to μ or $\mu_n\Rightarrow\mu$ in symbol.

LEMMA 1. *If $\mu_n\Rightarrow\mu$, then $\int_S g(s)d\mu_n(s)$ converges to $\int_S g(s)d\mu(s)$ as $n\to\infty$ for every $g\in M(S)$.*

PROOF. The proof of this lemma is derived by extending from indicator functions.

Let $p_f(s'|s)=q(s'|s,f(s))$ for any $f\in\mathscr{B}(S|A)$, where $q(s'|s,a)$ is a given Markov transition law of the system. For any $f\in\mathscr{B}(S|A)$ n-step transition probabilities $p_f^{(n)}\equiv p_f^{(n)}(s,E)$ are easily calculated in the following:

$$p_f^{(0)}\equiv p_f^{(0)}(s,E)=\begin{cases}1 & \text{for}\quad s\in E\ (E\in\mathscr{B}(S))\\ 0 & \text{otherwise},\end{cases}$$

$$p_f^{(1)}\equiv p_f^{(1)}(s,E)=p_f(s,E),$$

$$p_n^{(n+1)}\equiv p_f^{(n+1)}(s,E)=\int_S p_f^{(n)}(s',E)p_f^{(1)}(s,ds').$$

These are also stochastic transition laws.

The following Corollary is an immediate consequence of Lemma 1.

COROLLARY. *If for any fixed $f\in\mathscr{B}(S|A)$, $\frac{1}{n}\sum_{t=0}^{n-1}p_f^{(t)}(s,\cdot)\Rightarrow\pi(s,\cdot)$ for each $s\in S$ and if π is Markov transition law, then $\frac{1}{n}\sum_{t=0}^{n-1}\int_S p_f^{(t)}(s,ds')g(s')$ converges $\int_S\pi(s,ds')g(s')$ for each $s\in S$ and any $g\in M(S)$.*

Now we shall set up some hypotheses, called Doeblin condition, for any given Markov transition law $p(s,E)$ over general state space S. We shall say that the Markov transition law $p(s,E)$ satisfies Hypothesis (D) (or Doeblin condition) if there is a finite-valued measure φ of sets $A\in\mathscr{B}(S)$ with $\varphi(S)>0$, an integer $\nu\geqq 1$, and a positive ε such that

$$p^{(\nu)}(s,A)\leqq 1-\varepsilon\qquad\text{if}\quad\varphi(A)\leqq\varepsilon.$$

Hypothesis (D) will be strengthened to the following Hypothesis (D′). We shall say that the Markov transition law $p(s,E)$ satisfies Hypothesis (D′) if there is a measure φ on $\mathscr{B}(S)$ sets, with $0<\varphi(S)<\infty$, an integer $\nu\geqq 1$, a positive $\delta>0$, and a $\mathscr{B}(S)$ set C for which

$$\varphi(C)>0,$$

$$p_0^{(\nu)}(s,s')\geqq\delta,\qquad s\in S,\qquad s'\in C.$$

Here $p_0^{(\nu)}(s,\cdot)$ is the density of the absolutely continuous components of $p^{(\nu)}(s,\cdot)$

with respect to φ.

We shall use the following lemma, due to Doob [3], in the policy improvement procedures.

LEMMA 2. (Doob [3]) *If the Markov transition law $p(s, E)$ satisfies Hypothesis* (D′), *then there is a stationary absolute probability distribution $p(\cdot) \in P(S)$, with $p(C_1) \geqq \delta\varphi(C_1)$ when $C_1 \subset C$, for which*

$$|p^{(n)}(s, A) - p(A)| \leqq [1-\delta\varphi(C)]^{(n/\nu)-1}, \qquad n=1, 2, \cdots.$$

In the following we shall use Assumption (I): a Markov transition law resulted by any stationary policy satisfies Hypothesis (D′). That is, we shall say Assumption (I) is satisfied if for any $f \in \mathscr{B}(S|A)$ there exist, depending on f, a measure φ_f of $\mathscr{B}(S)$ sets, with $0 < \varphi_f(S) < \infty$, an integer $\nu = \nu(f) \geqq 1$, a $\delta = \delta(f) > 0$, and a $\mathscr{B}(S)$ set $C = C(f)$ such that

$$\varphi(C) > 0 \quad \text{and} \quad p_{0f}^{(\nu)}(s, s') \geqq \delta, \qquad s \in S, \quad s' \in C,$$

where $p_{0f}^{(\nu)}(s, \cdot)$ is the density of the absolutely continuous components of $p_f^{(\nu)}(s, \cdot)$ with respect to φ_f. Throughout the paper, the following three assumptions given in [5] will remain operative: (1) A is a compact metric space, (2) r is a bounded upper semi-continuous function on SA and (3) $q \in Q(S|SA)$ is weak continuous, that is, $(s_n, a_n) \to (s, a)$ implies that $q(\cdot|s_n, a_n)$ converges weakly to $q(\cdot|s, a)$.

LEMMA 3. *Under Assumption (I), for any $f \in \mathscr{B}(S|A)$*

$$J(f^{(\infty)})(s) = \int_S p_f(ds')r(s', f(s')).$$

Hence average reward from the stationary policy $f^{(\infty)}$ is independent of the initial state s.

PROOF. For any $f \in \mathscr{B}(S|A)$, $\pi = \{f, f, \cdots\} = f^{(\infty)} \in \Pi_s$, resulting stationary Markov transition law is $p_f(s, E) = q(E|s, f(s))$, and resulting reward is $r_f(s) = r(s, f(s))$. Hence, by definition

$$J(f^{(\infty)})(s) = \varliminf_{n\to\infty} \frac{1}{n} \sum_{\nu=0}^{n-1} \int_S p_f^{(\nu)}(s, ds')r_f(s').$$

But Assumption (I) establishes the existence of a stationary absolute probability distribution $p_f(\cdot)$ on S independent of initial state. Furthermore, Lemma 2 yields that

$$p_f^{(\nu)}(s, E) \to p_f(E) \qquad \text{(uniformly in } s \text{ and } E).$$

Hence, similarly,

$$\frac{1}{n} \sum_{\nu=0}^{n-1} p_f^{(\nu)}(s, E) \to p_f(E) \qquad \text{(uniformly in } s \text{ and } E).$$

By Lemma 1 $\frac{1}{n} \sum_{\nu=0}^{n-1} \int_S p_f^{(\nu)}(s, ds')r_f(s')$ converges to $\int_S p_f(ds')r_f(s')$ for each $s \in S$. We conclude that

$$J(f^{(\infty)})(s) = \int_S p_f(ds')r_f(s').$$

This completes the proof.

In other words, Assumption (I) implies that the setting up the criteria in terms of a probability p on S, e. g. p-optimal, $p(\cdot)$-average optimal or (p, ε)-optimal criteria,

is of no use. We remark that even if in Lemma 3 $r(s, a) \in C_1(SA)$ is extended to $r(s, a) \in M(SA)$, the same result remains valid.

Now we shall use the following lemmas and Selector Theorem, due to Maitra [5].

LEMMA 4 (Maitra [5]). *Let u be a bounded upper semi-conitnuous on SA. Define $u^*: S \to R$ by $u^*(s) = \max_{a \in A} u(s, a)$. Then u^* is upper semi-continuous on S.*

LEMMA 5 (Maitra [5]). *Let u be a bounded upper semi-continuous function on SA. Define $U: S \to 2^A$ by $U(s) = \{a \in A;\ u(s, a) = \max_{a' \in A} u(s, a')\}$, where 2^A is the set of all non-empty closed subset in A. Then U is a Borel map.*

SELECTOR THEOREM (Maitra [5]). *Let u be a bounded upper semi-continuous function on SA. Then there exists a Borel measurable map f from S into A such that $u(s, f(s)) = \max_{a \in A} u(s, a)$ for all $s \in S$.*

LEMMA 6 (Maitra [5]). *Let $w: S \to R$ be a bounded upper semi-continuous function. Then $g: SA \to R$ defined by $g(s, a) = \int w(\cdot) dq(\cdot \mid s, a)$ is upper semi-continuous on SA.*

We can now state and prove our theorem with the aid of above lemmas and Selector Theorem.

THEOREM 1. *If there exists a $\{g, v(\cdot)\}$ such that $v(\cdot) \in C_1(S)$, satisfying a functional equation*

$$g + v(s) = \max_{a \in A} \left\{ r(s, a) + \int_S q(ds' \mid s, a) v(s') \right\} \qquad s \in S, \tag{1}$$

then there exists an $f^* \in \mathscr{B}(S \mid A)$ such that for $s \in S$ and every $\pi \in \Pi$

$$g = J(f^{*(\infty)})(s) \geqq J(\pi)(s).$$

Here, $f^{*(\infty)}$ is the stationary policy which for each state $s \in S$ prescribes the action that maximize the right hand side of (1).

PROOF. The proof of this theorem is similar to that of Derman [2], except for the use of above lemmas and Selector Theorem. Let a_s^* $(s \in S)$ be the action which maximize the right hand side of (1). By Lemma 5 and Lemma 6, $U(s) = \{a_s^*\}$ is a Borel map from S to 2^A. Selector Theorem enables us to choose a Borel measurable $f^* \in \mathscr{B}(S \mid A)$ from S to A such that

$$f^*(s) \in U(s) \qquad \text{for every} \quad s \in S.$$

This reduces (1) to

$$g + v(s) = r_{f^*}(s) + \int_S p_{f^*}(s, ds') v(s'), \qquad s \in S, \tag{2}$$

where

$$r_{f^*}(s) = r(s, f^*(s)) \quad \text{and} \quad p_{f^*}(s, E) = q(E \mid s, f^*(s)).$$

On integrating (2) by $p_{f^*}^{(\nu)}$, ν-step transition probability, which is calculated from $p_{f^*}(s, E)$, we have

$$g + \int_S p_{f^*}^{(\nu)}(s_1, ds) v(s) = \int_S p_{f^*}^{(\nu)}(s_1, ds) r_{f^*}(s) + \int_S p_{f^*}^{(\nu+1)}(s_1, ds') v(s'). \tag{3}$$

An interchange of the order of the integrations in (3) is justified by Fubini's lemma.

By taking average over $\nu=0,1,\cdots,T-1$ in (3)

$$g=\frac{1}{T}\sum_{\nu=0}^{T-1}\int_S p_{f^*}^{(\nu)}(s_1,ds)r_{f^*}(s)+\frac{1}{T}\left[v(s_1)-\int_S p_{f^*}^{(T)}(s_1,ds')v(s')\right],$$

and canceling in the limit, we have

$$\begin{aligned} g&=\lim_{n\to\infty}\frac{1}{T}\sum_{\nu=0}^{T-1}\int_S p_{f^*}^{(\nu)}(s_1,ds)r_{f^*}(s) \\ &=J(f^{*(\infty)})(s_1). \end{aligned} \tag{4}$$

Thus, g is the expected average reward per unit time from the stationary policy $f^{*(\infty)}$.

Now we must show that $f^{*(\infty)}$ is optimal over Π. We shall successively define $g_n(s)$; $n=0,1,2,\cdots$ by

$$\begin{aligned} g_0(s)&=\max_{a\in A} r(s,a) \qquad s\in S, \\ g_{n+1}(s)&=\max_{a\in A}\left\{r(s,a)+\int_S q(ds'|s,a)g_n(s')\right\} \qquad s\in S. \end{aligned} \tag{5}$$

Thus $g_0(\cdot)$ is well defined because of compact action space and it is clear that $g_0(\cdot)\in C_1(S)$ by Lemma 4. Lemma 4 together with Lemma 6, implies that $g_n(\cdot)$, $n=1,2,\cdots$ are well defined and that $g_n(\cdot)\in C_1(S)$ for each n. By definition, $g_n(s)$ expresses the total expected reward obtained over the periods $0,1,\cdots,n$ operating optimally. We shall show that there exists a finite M such that for $n=0,1,2,\cdots$,

$$ng+v(s)-M\leqq g_n(s)\leqq ng+v(s)+M, \qquad s\in S. \tag{6}$$

For $n=0$, (6) holds since $v(\cdot)$ and $r(\cdot,\cdot)$ are bounded functions. Assume (6) holds for $n\leqq N$. Then, by (5), (6) and (1), we get

$$\begin{aligned} g_{N+1}(s)&=\max_{a\in A}\left\{r(s,a)+\int_S q(ds'|s,a)g_N(s')\right\} \\ &\leqq\max_{a\in A}\left\{r(s,a)+\int_S q(ds'|s,a)[Ng+v(s')+M]\right\} \\ &=Ng+M+\max_{a\in A}\left\{r(s,a)+\int_S q(s'|s,a)v(s')\right\} \\ &=Ng+M+g+v(s) \\ &=(N+1)g+v(s)+M, \end{aligned}$$

the right hand inequality of (6) for $N+1$. The left follows in the same way.

From the definition it follows that for any $\pi\in\Pi$, $I_n(\pi)(s)$ is the total expected reward over the periods $0,1,\cdots,n$ from π, started from state s. Since $g_n(s)$ is the results from optimal policy for those periods, we have

$$I_n(\pi)(s)\leqq g_n(s), \qquad n=1,2,\cdots.$$

Hence,

$$\frac{I_n(\pi)(s)}{n} \leqq \frac{g_n(s)}{n}, \qquad n=1, 2, \cdots.$$

Then (6) yields that

$$\liminf_{n\to\infty} \frac{I_n(\pi)(s)}{n} \leqq \lim_{n\to\infty} \frac{g_n(s)}{n}$$
$$= g, \qquad s \in S.$$

This states that

$$J(\pi)(s) \leqq g = J(f^{*(\infty)})(s), \qquad s \in S,$$

which completes the proof.

COROLLARY. *Under the same conditions as in Theorem* 1, *it holds that*

$$\|g_n(s) - ng\| \leqq 2M \qquad \textit{for} \quad n=1, 2, \cdots.$$

Next we shall consider a generalization of the policy improvement routine originally achieved by Howard [4]. Here, we set up the Assumption (II): for any $f \in \mathscr{B}(S|A)$, there exists a $\{g, v(s)\}$ such that $v(\cdot) \in C_1(S)$, satisfying the functional equation

$$g + v(s) = r(s, f(s)) + \int_S q(ds'|s, f(s))v(s'), \qquad s \in S \tag{7}$$

We now define our improvement procedure under Assumptions (I) and (II). For any given $f \in \mathscr{B}(S|A)$ we construct $f' \in \mathscr{B}(S|A)$ as follows: set $f'(s) = f(s)$ for each state $s \in S$ such that

$$r(s, f(s)) + \int_S q(ds'|s, f(s))v(s') = \max_{a \in A} \left\{ r(s, a) + \int_S q(ds'|s, a)v(s') \right\}, \tag{8}$$

and put $f'(s) = a'$ for the state s not satisfying (8), by using an action a' such that

$$r(s, a') + \int_S q(ds'|s, a')v(s') > r(s, f(s)) + \int_S q(ds'|s, f(s))v(s'). \tag{9}$$

That is,

$$f'(s) = \begin{cases} f(s) & \text{on} \quad \{s\,;\ u(s, f(s)) = \max_{a \in A} u(s, a)\} \\ & \qquad\qquad\qquad\qquad = \tilde{S}(f) \in \mathscr{B}(S) \\ a' & \text{on} \quad \tilde{S}(f)^c, \end{cases}$$

where

$$u(s, a) = r(s, a) + \int_S q(ds'|s, a)v(s') \in C_1(SA).$$

We remark that for example, Selector Theorem enables us to choose f' as Borel measurable, since there exists a measurable f' such that

$$u(s, f'(s)) = \max_{a \in A} u(s, a) \qquad \text{on} \quad \tilde{S}^c \in \mathscr{B}(S).$$

The following theorem shows that the policy can be properly improved in the case of general state.

THEOREM 2. *Assume that Assumptions (I) and (II) hold. If* $\tilde{S} = S$, *then* $f^{(\infty)}$ *is optimal over* Π. *If* $\tilde{S} \neq S$, *then*

$$J(f'^{(\infty)})(s) \geqq J(f^{(\infty)})(s) \qquad s \in S.$$

Furthermore if $\varphi_{f'}(\tilde{S}^c(f) \cap C(f')) > 0$, *then*

$$J(f'^{(\infty)})(s) > J(f^{(\infty)})(s) \qquad s \in S.$$

PROOF. Let $q(s' \mid s, f'(s)) = p_{f'}(s, ds')$. Let $\varepsilon(s)$ be the difference between the left hand side and right one in (9); that is,

$$\varepsilon(s) = \begin{cases} r(s, f(s)) + \int_S q(ds' \mid s, f(s))v(s') \\ \qquad -\max_{a \in A} \left\{ r(s, a) + \int_S q(ds' \mid s, a)v(s') \right\} = 0 \qquad \text{on } \tilde{S}, \\ r(s, a') + \int_S q(ds' \mid s, a')v(s') - r(s, f(s')) \\ \qquad -\int_S q(ds' \mid s, f(s))v(s') > 0 \qquad \text{on } \tilde{S}^c. \end{cases}$$

For any $s_1 \in S$ and integer ν we have, using Assumption (II), that

$$\begin{aligned} \int_S p_{f'}^{(\nu)}(s_1, ds)\varepsilon(s) = & \int_S p_{f'}^{(\nu)}(s_1, ds) r(s, f'(s)) \\ & + \int_S p_{f'}^{(\nu+1)}(s_1, ds')v(s') \\ & - g - \int_S p_{f'}^{(\nu)}(s_1, ds)v(s). \end{aligned} \tag{10}$$

On averaging over $\nu = 0, 1, \cdots, n-1$ in (10) we get

$$\begin{aligned} \int_S \varepsilon(s) \frac{1}{n} \sum_{\nu=0}^{n-1} p_{f'}^{(\nu)}(s_1, ds) = & \frac{1}{n} \sum_{\nu=0}^{n-1} \int_S p_{f'}^{(\nu)}(s_1, ds) r_{f'}(s) \\ & + \frac{1}{n} \left(\int_S p_{f'}^{(n)}(s_1, ds')v(s') - v(s_1) \right) - g. \end{aligned} \tag{11}$$

By Assumption (I), there exists the limit

$$\lim_{n \to \infty} \frac{1}{n} \sum_{\nu=0}^{n-1} p_{f'}^{(\nu)}(s_1, ds) = p_{f'}(ds).$$

Then on letting $n \to \infty$ in (11) we have, because of boundedness of $\varepsilon(s)$,

$$\int_S \varepsilon(s) p_{f'}(ds) = J(f'^{(\infty)})(s_1) - g.$$

On the other hand, it follows that

$$\int_S \varepsilon(s) p_{f'}(ds) = \int_{\tilde{S}^c} \varepsilon(s) p_{f'}(ds) \geqq 0.$$

In particular $\varphi_{f'}(\tilde{S}^c \cap C(f')) > 0$ implies $p_{f'}(\tilde{S}^c \cap C(f')) > 0$ by Lemma 2. Thus, if $\varphi_{f'}(\tilde{S}^c \cap C(f')) > 0$, then

$$\int_{\tilde{S}^c} \varepsilon(s) p_{f'}(ds) \geqq \int_{\tilde{S}^c \cap C(f')} \varepsilon(s) p_{f'}(ds) > 0,$$

namely

$$J(f'^{(\infty)})(s_1) > g = J(f^{(\infty)})(s_1), \qquad s_1 \in S.$$

This completes the proof of Theorem 2.

4. Reduction of average problem to discounted problem and the method of successive approximations.

We shall treat the Markovian decision processes satisfying Assumption (III): there are a state s_0 and $\alpha > 0$ such that

$$q(\{s_0\} \mid s, a) \geqq \alpha \qquad \text{for all} \quad s \in S, \quad a \in A.$$

For convenience we call the average reward Markovian decision process which has been stated as above the original M. D. P., and we shall define a modified M. P. D. with state space, action space and reward function not altered, but with transition probability q' altered.

In the modified M. D. P., q' is to be given by

$$q'(B \mid s, a) = \begin{cases} \dfrac{q(B \mid s, a)}{1-\alpha} & \text{for} \quad s_0 \notin B \\[2ex] \dfrac{q(B \mid s, a) - \alpha}{1-\alpha} & \text{for} \quad s_0 \in B \end{cases}$$

for any $B \in \mathscr{B}(S)$.

Note that q' is a Markov transition law.

We now consider the modified M. D. P. with β-discounted total expected reward over the infinite future and study its connection to the original M. D. P. with average criterion satisfying Assumption (III).

One of Maitra's results, which we shall use, is stated in

LEMMA 7 (Maitra [5]). *For any Markovian decision process* (S, A, q, r) *and* $0 \leqq \beta < 1$, *there exists a stationary policy which is* β*-optimal over* Π, *and the optimal reward* $(= \sup_{\pi \in \Pi} I^{\beta}(\pi)(s))$ *is upper semi-continuous on* S.

COROLLARY. *For our modified M. D. P.* (S, A, q', r) *and* $0 \leqq \beta < 1$, *there exists a* β*-optimal stationary policy over* Π.

It should be noted that the (s, a)-weak continuity of $q(\cdot \mid s, a)$ implies that of $q'(\cdot \mid s, a)$, since

$$\int_S g(s')dq'(s' \mid s, a) = \frac{1}{1-\alpha}\int_S g(s')dq(s' \mid s, a) - \frac{\alpha}{1-\alpha} g(s_0) \qquad \text{for} \quad g(\cdot) \in M(S).$$

Hence we may assume that for the modified M. D. P. (S, A, q', r) A is compact, r is bounded upper semi-continuous, and $q' \in Q(S \mid SA)$ is weak continuous.

Let π'_β be the β-optimal policy with respect to the modified M. D. P., and let $I'^{\beta}(\pi'_\beta)(s)$ be the β-discounted total expected reward with respect to the modified M. D. P., when the process starts in state s and policy π'_β is used. Note that any policy for the original M. D. P. can also be considered as a policy for the modified one and vice versa. The fundamental theorem on the reduction is the following:

THEOREM 3. *Under Assumption (III), there exists an optimal stationary policy for the original M. D. P., and the optimal average reward per unit time is* $\alpha I'^{1-\alpha}(\pi'_{1-\alpha})(s_0)$. *Furthermore, the optimal stationary policy is the one which takes the actions which maximize the right hand side of*

$$\alpha I'^{1-\alpha}(\pi'_{1-\alpha})(s_0)+f'(s)=\max_{a\in A}\left[r(s, a)+\int_S f'(s')dq(s'|s, a)\right], \tag{11}$$

where

$$f'(s)=I'^{1-\alpha}(\pi'_{1-\alpha})(s)-I'^{1-\alpha}(\pi'_{1-\alpha})(s_0). \tag{12}$$

PROOF. Blackwell [1] has shown that since $\pi'_{1-\alpha}$ is $(1-\alpha)$-optimal with respect to the modified M. D. P. (S, A, q', r), $I'^{1-\alpha}(\pi'_{1-\alpha})(s)$ satisfies the optimal equation, namely,

$$I'^{1-\alpha}(\pi'_{1-\alpha})(s)=\max_{a\in A}\left[r(s, a)+(1-\alpha)\int_S I'^{1-\alpha}(\pi'_{1-\alpha})(s')dq'(s'|s, a)\right], \qquad s\in S.$$

By (12) we have

$$f'(s)+I'^{1-\alpha}(\pi'_{1-\alpha})(s_0)=(1-\alpha)I'^{1-\alpha}(\pi'_{1-\alpha})(s_0)$$
$$+\max_{a\in A}\left[r(s, a)+(1-\alpha)\int_S f'(s')dq'(s'|s, a)\right].$$

This yields that

$$\alpha I'^{1-\alpha}(\pi'_{1-\alpha})(s_0)+f'(s)=\max_{a\in A}\left[r(s, a)+\int_S f'(s')dq(s'|s, a)\right]. \tag{13}$$

Thus, $\{\alpha I'^{1-\alpha}(\pi'_{1-\alpha})(s_0), f'(s)\}$ $(f'(\cdot)\in C_1(S))$ satisfies the optimal equation with respect to the average case, that is, condition of Theorem 1 is satisfied. It follows that there exists an optimal stationary policy $f^{*(\infty)}\in \Pi_s$ with respect to the original M. D. P. i. e.

$$\alpha I'^{1-\alpha}(\pi'_{1-\alpha})(s_0)=J(f^{*(\infty)})(s)\geqq J(\pi)(s) \qquad s\in S, \quad \pi\in \Pi,$$

and that $f^{*(\infty)}$ is the stationary one which takes the actions that maximize the right hand side of (11). This completes the proof of the Theorem 3.

We now show that the solution $\{g, v(\cdot)\}$ of the functional equation (1) can be solved by the method of successive approximations under Assumption (III)′: there exist an integer $\nu\geqq 0$, a quantity $0<\alpha\leqq 1$, and a state $s_0\in S$ such that for all $a_1, a_2, \cdots, a_{\nu+1}\in A$ and $s\in S$

$$q^{(\nu+1)}(\{s_0\}|s;\ a_1, a_2, \cdots, a_{\nu+1})\geqq \alpha$$

holds. This is a generalization of White's method. It should be noted that Assumption (III) implies Assumption (III)′.

THEOREM 4. *Under Assumption (III)′, the sequence* $\{g_n, v_n(\cdot), n\geqq 1\}$ *defined by*

$$\begin{aligned} V_n(s)&=\max_{a\in A}\left\{r(s, a)+\int_S q(ds'|s, a)v_{n-1}(s')\right\}, \qquad s\in S,\\ g_n&=V_n(s_0),\\ v_n(s)&=V_n(s)-g_n, \end{aligned} \tag{14}$$

converges uniformly to the solution $\{g, v(\cdot)\}$ *of the functional equation* (1),

$$g+v(s)=\max_{a\in A}\left\{r(s,a)+\int_S q(ds'\,|\,s,a)v(s')\right\} \qquad s\in S, \tag{1}$$

where

$$v_0(s)=\max_{a\in A} r(s,a)\,.$$

PROOF. By Lemma 4 and Lemma 6, it follows that $V_n(\cdot)$, $v_n(\cdot)$ are well-defined and that $V_n(\cdot)$, $v_n(\cdot)\in C_1(S)$. We need only prove the uniformity of convergence, since this implies that the limiting form is a solution of (1). For any sequence of upper semi-continuous functions $\{V_n(s)\}$ define

$$\nabla_n(V)=\inf_{s\in S}[V_n(s)-V_{n-1}(s)]$$

$$\Delta_n(V)=\sup_{s\in S}[V_n(s)-V_{n-1}(s)]\,,$$

and similarly,

$$\nabla_n(v)=\inf_{s\in S}[v_n(s)-v_{n-1}(s)]$$

$$\Delta_n(v)=\sup_{s\in S}[v_n(s)-v_{n-1}(s)]\,.$$

When $n\geqq\nu+3$, it can be easily shown that

$$\begin{aligned}V_n(s)-V_{n-1}(s)&\geqq\inf_{a\in A}\int_S q(ds'\,|\,s,a)[v_{n-1}(s')-v_{n-2}(s')]\\&=\inf_{a\in A}\left\{\int_S q(ds'\,|\,s,a)[V_{n-1}(s')-V_{n-2}(s')]-(V_{n-1}(s_0)-V_{n-2}(s_0))\right\}.\end{aligned}$$

Hence,

$$\begin{aligned}&V_n(s)-V_{n-1}(s)+\{V_{n-1}(s_0)-V_{n-2}(s_0)\}\\&\qquad\geqq\inf_{a\in A}\left\{\int_S q(ds'\,|\,s,a)\Big[\inf_{a'\in A}\int_S q(ds''\,|\,s',a')\Big][V_{n-2}(s'')-V_{n-3}(s'')]\right.\\&\qquad\qquad\left.-\{V_{n-2}(s_0)-V_{n-3}(s_0)\}\right\},\end{aligned}$$

namely,

$$\begin{aligned}&V_n(s)-V_{n-1}(s)+\{V_{n-1}(s_0)-V_{n-2}(s_0)\}+\{V_{n-2}(s_0)-V_{n-3}(s_0)\}\\&\qquad\geqq\inf_{a\in A}\left\{\int_S q(ds'\,|\,s,a)\Big[\inf_{a'\in A}\int_S q(ds''\,|\,s',a')\Big][V_{n-2}(s'')-V_{n-3}(s'')]\right\}\\&\qquad=\inf_{a,a'}\int_S q^{(2)}(ds''\,|\,s\,;\,a,a')[V_{n-2}(s'')-V_{n-3}(s'')]\,.\end{aligned}$$

Repeating this interation procedure $(\nu-1)$ times, we have

$$\begin{aligned}&V_n(s)-V_{n-1}(s)+\sum_{k=1}^{\nu}\{V_{n-k}(s_0)-V_{n-k-1}(s_0)\}\\&\qquad\geqq\inf_{a_1,a_2,\cdots,a_\nu}\int_S q^{(\nu)}(s'\,|\,s\,;\,a_1,a_2,\cdots,a_\nu)\{V_{n-\nu}(s')-V_{n-\nu-1}(s')\}\,,\qquad\text{i.e.}\\&V_n(s)-V_{n-1}(s)+\{V_{n-1}(s_0)-V_{n-\nu-1}(s_0)\}\\&\qquad\geqq\inf_{a_1,a_2,\cdots,a_{\nu+1}}\int_S q^{(\nu+1)}(s'\,|\,s\,;\,a_1,a_2,\cdots,a_{\nu+1})\{v_{n-\nu-1}(s')-v_{n-\nu-2}(s')\}\,.\end{aligned}$$

Now by hypothesis

$$v_{n-\nu-1}(s_0)=v_{n-\nu-2}(s_0)=0$$

and

$$q(\{s_0\}\,|\,s\,;\,a_1,\,a_2,\,\cdots,\,a_{\nu+1})\geqq\alpha\,,$$

then we have

$$\nabla_n(V)+\{V_{n-1}(s_0)-V_{n-\nu-1}(s_0)\}\geqq(1-\alpha)\nabla_{n-\nu-1}(v)\,. \tag{15}$$

Similarly,

$$\Delta_n(V)+\{V_{n-1}(s_0)-V_{n-\nu-1}(s_0)\}\leqq(1-\alpha)\Delta_{n-\nu-1}(v)\,. \tag{16}$$

(15) and (16) imply that

$$\Delta_n(V)-\nabla_n(V)\leqq(1-\alpha)(\Delta_{n-\nu-1}(v)-\nabla_{n-\nu-1}(v))\,. \tag{17}$$

By the way,

$$\nabla_n(v)=\nabla_n(V)-(g_n-g_{n-1})$$

and

$$\Delta_n(v)=\Delta_n(V)-(g_n-g_{n-1})$$

yield

$$\Delta_n(v)-\nabla_n(v)=\Delta_n(V)-\nabla_n(V)\,. \tag{18}$$

From (17), (18)

$$\Delta_n(v)-\nabla_n(v)\leqq(1-\alpha)\{\Delta_{n-\nu-1}(v)-\nabla_{n-\nu-1}(v)\}\,.$$

Let $D_n(v)=\Delta_n(v)-\nabla_n(v)$. Then we have for some finite C, and for $n=N(\nu+1)+r$ $(0\leqq r\leqq\nu)$,

$$D_n(v)\leqq(1-\alpha)^N D_r(v)\leqq(1-\alpha)^N C\,.$$

Since $v_n(s_0)-v_{n-1}(s_0)=0$, it follows that

$$\nabla_n(v)\leqq 0\leqq\Delta_n(v)\,.$$

If

$$U_n(V)=\sup_{s\in S}|V_n(s)-V_{n-1}(s)|$$

and

$$U_n(v)=\sup_{s\in S}|v_n(s)-v_{n-1}(s)|\,,$$

then

$$U_n(v)\leqq D_n(v)\leqq(1-\alpha)^N C\,.$$

Hence $\{v_n(s),\,n\geqq 1\}$ converges uniformly to the function $v(s)$.

On the other hand, $U_n(V)\leqq U_{n-1}(v)$ implies the uniform convergence of $\{V_n(s)\}$ to the function $V(s)$, so does the sequence $\{v_n(s_0)\}=\{g_n\}$. The interchange of the order of maximum and limit can be justified by the inequalities

$$|\max_{a\in A}f_n(s,\,a)-\max_{a\in A}f(s,\,a)|\leqq\sup_{a\in A}|f_n(s,\,a)-f(s,\,a)|$$

and

$$\|f_n-f\|\leqq\|v_{n-1}-v\|\,,$$

where

$$f_n(s, a) = r(s, a) + \int_S q(s' \mid s, a) v_{n-1}(s') \in C_1(SA)$$

and

$$f(s, a) = r(s, a) + \int_S q(s' \mid s, a) v(s') \in C_1(SA).$$

Acknowledgement: The author would like to express his deep appreciation to Doctor N. Furukawa whose advices and inspiration helped make his paper possible.

References

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

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

[3] J. L. DOOB, *Stochastic Processes,* Wiely, New York (1953).

[4] R. A. HOWARD, *Dynamic Programming and Markov Processes,* Wiley, New York (1960).

[5] A. MAITRA, *Discounted dynamic programming on compact metric spaces,* Sankhyā Ser A., **30** (1968), 211–216.

[6] S. M. ROSS, *Arbitrary state Markovian Decision Processes,* Ann. Math. Statist., **39** (1968), 2118–2122.

[7] D. J. WHITE, *Dynamic Programming, Markov chains, and the method of successive approximations,* J. Math. Anal. Appl., **6** (1963), 373–376.