A MARKOV DECISION PROCESS WITH NON-STATIONARY TRANSITION LAWS

Furukawa, Nagata

https://doi.org/10.5109/13030

出版情報：統計数理研究. 13 (1/2), pp.41-52, 1968-03. 統計科学研究会
バージョン：
権利関係：

A MARKOV DECISION PROCESS WITH NON-STATIONARY TRANSITION LAWS

By

Nagata FURUKAWA

(Received Dec. 15, 1967)

§ 1. Introduction.

In most of works on Markov decision process ([1], [2], [3] etc.) Markov transition laws of states are assumed to be stationary in the sense that the transition law at any time depends on the state and the action at present time but not on the time. This stationarity of motion law yields some elegant properties on optimal strategies in the discounted case: In [2] Blackwell has proved that in the discounted case there was a (p, ε)-optimal stationary strategy for any initial distribution p and any $\varepsilon < 0$, especially in the case of a countable action space there was an ε-optimal stationary strategy for any $\varepsilon > 0$, and in the case of a finite action space an optimal stationary strategy. In [3] Strauch has cleared the stationary property of a strong (p, ε)-optimal strategy in the discounted case.

In this paper we shall be concerned with a Markov decision process with non-stationary Markov transition laws, and we shall study the existence and the properties of an optimal strategy, the existence of a strong (p, ε)-optimal strategy, and the strategy improvements.

Our decision problem is based on four objects S, A, $q = \{q_1, q_2, \cdots\}$ and r. S and A are non-empty Borel sets, each q_j is a conditional probability distribution on S given $S \times A$, and r is a bounded Baire function on $S \times A \times S$. Here S is the set of states, A the set of feasible actions, q the sequence of Markov transition laws of states, and r a reward function. When the system is at a time j and in a state s and an action a is taken, the system moves to a new state according to the conditional probability distribution $q_j(\cdot \mid s, a)$, and if the system moves to a new state s', then we shall receive a reward $r(s, a, s')$. In this situation we wish to maximize the total discounted expected reward over the infinite future.

A *strategy* π is a sequence $\pi_1, \pi_2, \cdots$, where π_j is a conditional probability distribution on A given $(s_1, a_1, \cdots, a_{j-1}, s_j)$ for each j, and π is denoted by $\{\pi_1, \pi_2, \cdots\}$. A *Markov strategy* is a sequence $f_1, f_2, \cdots$, where each f_j is a measurable mapping from S to A, and a *l-stationary strategy* is a Markov strategy such that $\pi = \{\bar{f}, \bar{f}, \bar{f}, \cdots\}$ where $\bar{f} = \{f_1, f_2, \cdots, f_l\}$. q is called *l-stationary* if $q = \{\bar{q}, \bar{q}, \bar{q}, \cdots\}$ with $\bar{q} = \{q_1, q_2, \cdots, q_l\}$.

Our main results are the following: For any initial probability distribution p, any $\varepsilon > 0$ and any q there is a (p, ε, q)-optimal Markov strategy, i.e. there is a Markov strategy π^* for which $p\{I(\pi^*, q) \geqq I(\pi, q) - \varepsilon\}$ for every strategy π, where $I(\pi, q)$ denotes the total discounted expected reward from a strategy π (Section 4). If for each $j \geqq 0$

there is a $(\varepsilon, {}^{j}q)$-optimal strategy, there is a $(\varepsilon/(1-\beta), q)$-optimal Markov strategy, where $\varepsilon \geqq 0$ and ${}^{j}q = \{q_{j+1}, q_{j+2}, \cdots\}$ (Section 4). For each $j \geqq 0$ ${}^{j}\pi^*$ is a ${}^{j}q$-optimal if and only if the total discounted expected reward from π^* satisfies the system of optimality equations, where ${}^{j}\pi^* = \{\pi^*_{j+1}, \pi^*_{j+2}, \cdots\}$ for $\pi^* = \{\pi^*_1, \pi^*_2, \cdots\}$ (Section 4). If there exists a ${}^{j}q$-optimal strategy for each $j \geqq 0$, then there exists a q-optimal Markov strategy. If A is essentially countable, there is a (ε, q)-optimal Markov strategy for every $\varepsilon > 0$ and every q, and if A is essentially finite, then there is a q-optimal Markov strategy for every q (Section 4). Eespecially in the case when q is l-stationary, all of the results in Section 4 stated above hold by putting a l-stationary strategy in place of a Markov strategy (Section 5). Every sequence of Markov strategies is strongly ε-dominated by a Markov strategy for every $\varepsilon > 0$ (Section 6). For any p, $\varepsilon > 0$ and any q, there is a strong (p, ε, q)-optimal Markov strategy (Section 6). In Section 7 there are given several theorems on the improvement of strategies.

In [3] Strauch has derived all of his results on the strong optimality, i.e. the strong (p, ε)-domination, the existence of the strong (p, ε)-optimal Markov strategy and of the strong (p, ε)-optimal stationary strategy etc., with the help of a "conservation". But our proofs of results on the strong optimality are more direct without appealing to the "conservation" (Section 6).

§ 2. Probabilistic definitions.

By a Borel set we mean a Borel subset of some complete separable metric space. The class of all probability distributions on X is denoted by $P(X)$. For any nonempty Borel sets X, Y a conditional probability distribution 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 distribution on Y and for each Borel set $B \subset Y$, $q(B \mid \cdot)$ is a Baire function on X. The class of all conditional probability distributions on Y given X is denoted by $Q(Y \mid X)$. The product space of X and Y will be denoted by XY. The class of bounded Baire functions on X is denoted by $M(X)$. For any $u \in M(XY)$ and any $q \in Q(Y \mid X)$, qu denotes the element of $M(X)$ whose value at $x_0 \in X$ is $qu(x_0) = \int u(x_0, y) dq(y \mid x_0)$. For any $p \in P(X)$ and any $u \in M(X)$, pu denotes the integral of u with respect to p.

For any $p \in P(X)$, $q \in Q(Y \mid X)$, pq is the probability distribution on XY such that, for every $u \in M(XY)$, $pq(u) = p(qu)$. Every probability distribution m on XY has a factorization $m = pq$; p is unique and is just the marginal distribution of the first coordinate variable with respect to m; q is not quite unique; it is a version of the conditional distributiontion of the second coordinate variable given the first. These facts are given in [4].

We extend the above notation in an obvious way to a finite or countable sequence of non-empty Borel sets $X_1, X_2, \cdots$. If $q_n \in Q(X_{n+1} \mid X_1 \cdots X_n)$ for $n \geqq 1$ and $p \in P(X_1)$, $pq_1q_2 \cdots q_n$ is a probability distribution on $X_1X_2 \cdots X_{n+1}$, $pq_1q_2 \cdots$ is a probability distribution on the infinite product space $X_1X_2 \cdots$, $q_2q_3 \in Q(X_3X_4 \mid X_1X_2)$, for any $u \in M(X_1X_2 \cdots X_{n+1})$, $n \geqq 1$, and any m, $1 \leqq m \leqq n$, $q_m \cdots q_nu \in M(X_1 \cdots X_m)$, etc.

For the sake of simplicity, we introduce the following ambiguity: 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, for any $q\in Q(Y|X)$, $u\in M(Y)$, $qu\in M(X)$; any $q\in Q(Y|X)$ will also denote the element q' of $Q(Y|ZX)$ defined by $q'(\cdot|z, \cdot)=q(\cdot|\cdot)$, etc.

A $p\in P(X)$ is *degenerate* if it is concentrated at some one point $x\in X$; a $q\in Q(Y|X)$ is *depenerate* if each $q(\cdot|x)$ is degenerate. The degenerate q are exactly those for which there is a Baire function f mapping X into Y for which $q(\{f(x)\}|x)=1$ for all $x\in X$. Any such f will also denote its associated degenerate q, so that, for any $u\in M(XY)$, $fu(x)=u(x, f(x))$ for all $x\in X$.

We shall use the following.

LEMMA 2.1 (Blackwell [2]). *For any $q\in Q(Y|X)$, $u\in M(XY)$, there is a degenerate $f\in Q(Y|X)$ such that*

$$fu\geqq qu \quad \textit{for all} \quad x\in X.$$

§ 3. Decision problem definitions.

Our dynamic programming problem is defined by S, A, q, r where S, A are any non-empty Borel sets, $q=\{q_1, q_2, q_3, \cdots\}$, $q_i\in Q(S|SA)$ for $i=1, 2, 3, \cdots$, $r\in M(SAS)$, and $0<\beta<1$. A *strategy* π is a sequence $\{\pi_1, \pi_2, \pi_3, \cdots\}$, where $\pi_n\in Q(A|H_n)$ and $H_n=SASA\cdots S$ ($2n-1$ factors) is the set of possible histories of the system when the n-th act must be chosen. A strategy π is *Markov* if each π_n is a degenerate element of $Q(A|S)$, i.e. $\pi=\{f_1, f_2, f_3, \cdots\}$, where each f_n is a Baire function from S into A, and is *l-stationary* if it is Markov and $\pi=\{\bar{f}, \bar{f}, \cdots\}$ with $\bar{f}=\{f_1, f_2, \cdots, f_l\}$. The l-stationary strategy decided by $\bar{f}$ is denoted by $\bar{f}^{(\infty)}$.

For any strategy π, let ${}^n\pi=\{\pi_{n+1}, \pi_{n+2}, \cdots\}$ denote the strategy which π defines from the $(n+1)$-th stage onward. In particular, ${}^0\pi=\pi$. And let ${}^nq=\{q_{n+1}, q_{n+2}, \cdots\}$.

Any strategy π, together with the law of motion q, defines a conditional probability distribution on the set $X=ASAS\cdots$ of future of the system given the initial states s; i.e. it defines

$$e_\pi=\pi_1 q_1 \pi_2 q_2 \cdots \in Q(X|S).$$

Any reward function r defines an expected reward function on S gixen by

$$I(\pi, q)=e_\pi \sum_{j=1}^{\infty} \beta^{j-1} r(s_j, a_j, s_{j+1}).$$

For any $v\in M(S)$, let

$$I_n(\pi, q, v)=e_\pi\Big[\sum_{j=1}^{n} \beta^{j-1} r(s_j, a_j, s_{j+1})+\beta^n v\Big].$$

We shall denote $I_n(\pi, q, 0)$ by $I_n(\pi, q)$. Let Q^* denote the class of all sequences $\{q_1, q_2, \cdots\}$ such that $q_n\in Q(S|SA)$ for $n=1, 2, \cdots$.

It is clear that

LEMMA 3.1. *$I_n(\pi, q, v)\to I(\pi, q)$ as $n\to\infty$ for any $v\in M(S)$ and any $q\in Q^*$.*

For any $p\in P(S)$, any $\varepsilon>0$, and any $q\in Q^*$, π^* is called *(p, ε, q)-optimal* if $p\{I(\pi^*, q)\geqq I(\pi, q)-\varepsilon\}=1$ for every π. π^* is called *(ε, q)-optimal* if it is (p, ε, q)-optimal for every $p\in P(S)$, or, equivalently if $I(\pi^*, q)\geqq I(\pi, q)-\varepsilon$ for all π, and is called *q-optimal* if it is (ε, q)-optimal for every $\varepsilon>0$, or, equivalently if $I(\pi^*, q)\geqq I(\pi, q)$ for all π.

§ 4. Optimality.

LEMMA 4.1. *For any $p \in P(S)$, $\varepsilon > 0$, and $q \in Q^*$, there is a (p, ε, q)-optimal Markov strategy.*

The proof of this lemma is straightforward by replacing q in Theorem 2 of [2] by q_i.

With any measurable f_n from S to A and any $q_j \in Q(S|SA)$ we associate the operator T_{nj} from $M(S)$ to $M(S)$ defined by

$$T_{nj}u(s) = \int [r(s, f_n(s), t) + \beta u(t)] dq_j(t|s, f_n(s)) .$$

We shall call T_{nj} the *operator associated with* (f_n, q_j). With any Markov strategy $\pi = \{f_1, f_2, \cdots\}$ and any $q_j \in Q(S|SA)$ we associate the operator U_j from $M(S)$ to $M(S)$ defined by

$$U_j u(s) = \sup_n T_{nj}u(s) ,$$

where T_{nj} is the operator associated with (f_n, q_j). We shall call U_j the *operator associated with* (π, q_j). Let $(f_n, q_j) \sim T_{nj}$ mean that T_{nj} is the operator associated with (f_n, q_j), and let $(\pi, q_j) \sim U_j$ mean that U_j associated with (π, q_j). The following properties of T_{nj} are immediate from the definition.

THEOREM 4.1. (a) *T_{nj} is monotone; i.e. $u \leqq v$ implies $T_{nj}u \leqq T_{nj}v$.*

(b) *$T_{nj}(u+c) = T_{nj}u + \beta c$ for any constant c.*

(c) *If $\pi = \{f_1, f_2, \cdots\}$ is a Markov strategy and $(f_n, q_n) \sim T_{nn}$, then $T_{11}T_{22} \cdots T_{nn}v = I_n(\pi, q, v)$ for each n.*

(d) *If $\pi = \{f_1, f_2, \cdots\}$ is a Markov strategy and $(f_n, q_n) \sim T_{nn}$, then $T_{nn}I({}^n\pi, {}^nq) = I({}^{n-1}\pi, {}^{n-1}q)$ for each n.*

For any Markov $\pi = \{f_1, f_2, \cdots\}$ we shall say that measurable f mapping S into A is *π-generated* if there is a partition of S into Borel sets $S_1, S_2, \cdots$ such that $f = f_n$ on S_n and we shall say that a Markov $\pi' = \{g_1, g_2, \cdots\}$ is *π-generated* if each measurable g_n is π-generated. Let $F(\pi)$ denote the class of all π-generated measurable functions, and let $G(\pi)$ the class of all π-generated Markov strategies.

THEOREM 4.2. (a) *Let π be any Markov strategy. Then for every $q_j \in Q(S|SA)$, $\hat{T}_j u \leqq U_j u$ for any $\hat{f} \in F(\pi)$ where $(\hat{f}, q_j) \sim \hat{T}_j$.*

(b) *For any Markov π, any $\varepsilon > 0$ and any $q_j \in Q(S|SA)$, there exists $\hat{f}_j \in F(\pi)$ such that $\hat{T}_{jj}u \geqq U_j u - \varepsilon$ where $(\hat{f}_j, q_j) \sim \hat{T}_{jj}$.*

PROOF. (a) Let π be any Markov strategy. For any $\hat{f} \in F(\pi)$, let $(\hat{f}, q_j) \sim \hat{T}_j$. By the definition of $F(\pi)$, there exists a partition of S into Borel sets $S_1, S_2, \cdots$ such that $\hat{f} = f_n$ on S_n. Then for $s \in S_n$, $\hat{T}_j u(s) = T_{nj}u(s) \leqq U_j u(s)$. This holds for each n, which implies that $T_j u(s) \leqq U_j u(s)$ for all s.

(b) Let $\pi = \{f_1, f_2, \cdots\}$ be any Markov strategy, and let $(f_n, q_j) \sim T_{nj}$. We let $S_{nj} = \{s | T_{mj}u < U_j u - \varepsilon$ for $m \leqq n-1$, $T_{nj}u \geqq U_j u - \varepsilon\}$, and define $\hat{f}_j$ by $\hat{f}_j = f_n$ on S_{nj}. Then for $s \in S_{nj}$, $\hat{T}_{jj}u = T_{nj}u \geqq U_j u - \varepsilon$, where $(\hat{f}_j, q_j) \sim \hat{T}_{jj}$. So $\hat{T}_{jj}u \geqq U_j u - \varepsilon$ everywhere. Obviously $\hat{f}_j \in F(\pi)$. Hence the theorem is proved.

For any Markov π and any $q_j \in Q(S|SA)$, let $(\pi, q_j) \sim U_j$. We shall call $u^*_{j-1} = \lim_{n\to\infty} U_j U_{j+1} \cdots U_n u$ the *limit point associated with* $(\pi, {}^{j-1}q)$, where $u \in M(S)$. Let

$(\pi, {}^{j-1}q) \sim u^*_{j-1}$ mean that u^*_{j-1} is the limit point associated with $(\pi, {}^{j-1}q)$. In particular let $u^* = u^*_0$.

THEOREM 4.3. (a) *For any Markov π and any $q_j \in Q(S|SA)$, let $(\pi, q) \sim u^*$. Then $I(\hat{\pi}, q) \leqq u^*$ for every $\hat{\pi} \in G(\pi)$.*

(b) *For any Markov π, any $\varepsilon > 0$ and any $q_j \in Q(S|SA)$, there is a $\hat{\pi} \in G(\pi)$ such that $I(\hat{\pi}, q) \geqq u^* - \varepsilon$, where $(\pi, q) \sim u^*$.*

(c) *If for each $j \geqq 0$ there is a $(\varepsilon, {}^jq)$-optimal strategy, there is a $(\varepsilon/(1-\beta), q)$-optimal Markov strategy, where $\varepsilon \geqq 0$.*

(d) *Let $(f \equiv a, q_j) \sim T_{aj}$. Then if for avery $\varepsilon > 0$ and for every $j \geqq 0$ there is a $(\varepsilon, {}^jq)$-optimal strategy, there is a Markov strategy $\hat{\pi}$ such that the limit point $\hat{u}^*_j$ associated with $(\hat{\pi}, {}^jq)$ is a Baire function for each j and it satisfies the equations $\hat{u}^*_{j-1} = \sup_{a \in A} T_{aj}\hat{u}^*_j$ for $j = 1, 2, \cdots$.*

(e) *For each $j \geqq 0$ ${}^j\pi^*$ is a ${}^jq$-optimal if and only if the reward of π^* satisfies the optimality equations; $I({}^{j-1}\pi^*, {}^{j-1}q) = \sup_{a \in A} T_{aj} I({}^j\pi^*, {}^jq)$ for $j = 1, 2, \cdots$.*

PROOF. (a) Let π be any Markov strategy. Let $\hat{\pi} = \{g_1, g_2, \cdots\} \in G(\pi)$, and let $(g_j, q_j) \sim \hat{T}_j$. By Theorem 4.1 (d) we have $\hat{T}_{nn} I({}^n\hat{\pi}, {}^nq) = I({}^{n-1}\hat{\pi}, {}^{n-1}q)$ for each n, and backward inductively $\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{nn} I({}^n\hat{\pi}, {}^nq) = I(\hat{\pi}, q)$, i.e. $\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{nn} u_n = I(\hat{\pi}, q)$, where $u_n = I({}^n\hat{\pi}, {}^nq)$.

Since for any $u \in M(S)$

$$\|\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{nn}u_n - \hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{nn}u\| \leqq \beta^n \|u_n - u\| \leqq \beta^n(\|r\|/(1-\beta) + \|u\|),$$

we have

$$\|I(\hat{\pi}, q) - \hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{nn}u\| \leqq \beta^n(\|r\|/(1-\beta) + \|u\|).$$

Therefore $\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{nn}u \to I(\hat{\pi}, q)$ as $n \to \infty$. In virtue of Theorem 4.2 (a) $\hat{T}_{jj}u \leqq U_j u$ for each j, so that backward inductively $\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{jj}u \leqq U_1 U_2 \cdots U_j u$ for each j, which implies that $I(\hat{\pi}, q) \leqq u^*$ letting $j \to \infty$.

(b) Let $\varepsilon' = \varepsilon(1-\beta)$. Then, from Theorem 4.2 (b), there exists an $\hat{f}_j \in F(\pi)$ corresponding to (q_j, u) for which $\hat{T}_{jj}u \geqq U_j u - \varepsilon'$ where $(\hat{f}_j, q_j) \sim \hat{T}_{jj}$. Similary there is an $\hat{f}_{j-1} \in F(\pi)$ corresponding to $(q_{j-1}, U_j u - \varepsilon')$ for which $\hat{T}_{j-1,j-1}(U_j u - \varepsilon') \geqq U_{j-1}(U_j u - \varepsilon')$. Thus we verify inductively that

$$\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{jj}u \geqq U_1 U_2 \cdots U_j u - \varepsilon'(1 + \beta + \cdots + \beta^{j-1}) \qquad \text{for all } j \geqq 1.$$

We conclude that $I(\hat{\pi}, q) \geqq u^* - \varepsilon'/(1-\beta) = u^* - \varepsilon$, letting $j \to \infty$.

(c) Assume that $\pi^{*j} = \{\pi_{j_1}, \pi_{j_2}, \cdots\}$ is a $(\varepsilon, {}^jq)$-optimal strategy for each $j \geqq 0$. From Lemma 2.1, for each j there exists a degenerate f_j such that

$$I(\pi^{*j-1}, {}^{j-1}q) = \pi_{j-1}, {}_1q_j[r + \beta I({}^1\pi^{*j-1}, {}^jq)] \leqq \pi_{j-1}, {}_1q_j[r + \beta\{I(\pi^{*j}, {}^jq) + \varepsilon\}]$$
$$\leqq f_j q_j[r + \beta\{I(\pi^{*j}, {}^jq) + \varepsilon\}] = T_{jj} I(\pi^{*j}, {}^jq) + \beta\varepsilon,$$

where $(f_j, q_j) \sim T_{jj}$. We have inductively degenerate $f_1, f_2, \cdots, f_{j-1}$ for which

$$T_{11}T_{22} \cdots T_{jj} I(\pi^{*j}, {}^jq) \geqq I(\pi^{*0}, q) - \varepsilon(\beta + \beta^2 + \cdots + \beta^j) \qquad \text{for all } j \geqq 1.$$

Letting $j \to \infty$ yields $I(\hat{\pi}, q) \geqq I(\pi^{*0}, q) - \beta\varepsilon/(1-\beta)$, where $\hat{\pi} = \{f_1, f_2, \cdots\}$.

Since π^{*0} is a (ε, q)-optimal strategy, $I(\pi^{*0}, q) \geqq I(\pi, q) - \varepsilon$ for all π. Therefore $I(\hat{\pi}, q) \geqq I(\pi, q) - \varepsilon/(1-\beta)$ for all π, which implies that $\hat{\pi}$ is a $(\varepsilon/(1-\beta), q)$-optimal Markov

strategy.

(d) From (c), the hypothesis implies that there is a $(1/n, {}^jq)$-optimal Markov strategy π^{jn}, say, for each n, j. Let $\hat{\pi}$ be a Markov strategy for which $\pi^{jn} \in G(\hat{\pi})$ for all j, n and let $(\hat{\pi}, {}^jq) \rightsquigarrow \hat{u}_j^*$.

From (a), we have $I(\pi, {}^jq) \leqq \hat{u}_j^*$ for all $\pi \in G(\hat{\pi})$. Since $\pi^{jn} \in G(\hat{\pi})$ for all n, j, $I(\pi^{jn}, {}^jq) \leqq \hat{u}_j^*$ for all n, j. But from the definition of π^{jn}, $I(\pi^{jn}, {}^jq) \geqq I(\pi, {}^jq) - 1/n$ for all π. Therefore $\hat{u}_j^* \geqq I(\pi, {}^jq) - 1/n$ for all π. Letting $n \to \infty$ yields $\hat{u}_j^* \geqq I(\pi, {}^jq)$ for all π and all j. Since from (b) there exists a $\tilde{\pi}^{nj} \in G(\pi)$ such that $I(\tilde{\pi}^{nj}, {}^jq) \geqq \hat{u}_j^* - 1/n$ for each n, j, we conclude that

$$T_{aj}\hat{u}_j^* \leqq T_{aj}[I(\tilde{\pi}^{nj}, {}^jq) + 1/n] = I((a, \tilde{\pi}^{nj}), {}^{j-1}q) + \beta/n \leqq \hat{u}_{j-1}^* + \beta/n ,$$

which implies $\sup_{a \in A} T_{aj}\hat{u}_j^* \leqq \hat{u}_{j-1}^*$. On the other hand it holds that $\sup_{a \in A} T_{aj}\hat{u}_j^* \geqq U_j\hat{u}_j^* = \hat{u}_{j-1}^*$. Thus we have $\sup_{a \in A} T_{aj}\hat{u}_j^* = \hat{u}_{j-1}^*$ for each j.

(e) Assume that ${}^j\pi^*$ is a ${}^jq$-optimal strategy for each j. Then from (c) we may assume $\pi^* = \{f_1^*, f_2^*, \cdots\}$ is Markov without loss of generality. Hence $I({}^{j-1}\pi^*, {}^{j-1}q)_{s_0} = T_{f_j^*(s_0)}, {}_jI({}^j\pi^*, {}^jq)_{s_0}$ for each j, so that $I({}^{j-1}\pi^*, {}^{j-1}q)_{s_0} \leqq \sup_{a \in A} T_{aj}I({}^j\pi^*, {}^jq)_{s_0}$ for each j. Since this holds for any $s_0 \in S$, it follows that $I({}^{j-1}\pi^*, {}^{j-1}q) \leqq \sup_{a \in A} T_{aj}I({}^j\pi^*, {}^jq)$ everywhere for each j. But $I({}^{j-1}\pi^*, {}^{j-1}q) \geqq I((a, {}^j\pi^*), {}^{j-1}q) = T_{aj}I({}^j\pi^*, {}^jq)$ for every j and for all $a \in A$, which yields that $I({}^{j-1}\pi^*, {}^{j-1}q) \geqq \sup_{a \in A} T_{aj}I({}^j\pi^*, {}^jq)$ for every j. Finally we obtain optimality equations for π^*; $I({}^{j-1}\pi^*, {}^{j-1}q) = \sup_{a \in A} T_{aj}I({}^j\pi^*, {}^jq)$ for every j.

Conversely assume that π^* satisfies the optimality equations. In virtue of Lemma 4.1, for any $p \in P(S)$, $\varepsilon > 0$, $q \in Q^*$, there is a (p, ε, q)-optimal Markov strategy $\hat{\pi} = (\hat{f}_1, \hat{f}_2, \cdots)$, say, i.e. $p\{I(\hat{\pi}, q) \geqq I(\pi, q) - \varepsilon\} = 1$ for all π. So in particular it holds that $I(\hat{\pi}, q)_{s_0} \geqq I(\pi, q)_{s_0} - \varepsilon$ for all π. Let $(\hat{f}_j, q_j) \rightsquigarrow \hat{T}_{jj}$, then from the assumption $\hat{T}_{jj}I({}^j\pi^*, {}^jq) \leqq I({}^{j-1}\pi^*, {}^{j-1}q)$ for every j, and backward inductively $\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{jj}I({}^j\pi^*, {}^jq) \leqq I(\pi^*, q)$ for every j. Letting $j \to \infty$ yields $I(\hat{\pi}, q) \leqq I(\pi^*, q)$.

Therefore we have

$$I(\pi, q)_{s_0} \leqq I(\hat{\pi}, q)_{s_0} + \varepsilon \leqq I(\pi^*, q)_{s_0} + \varepsilon \qquad \text{for all } \pi ,$$

and we have $I(\pi, q)_{s_0} \leqq I(\pi^*, q)_{s_0}$ for all π by letting $\varepsilon \to 0$. This hold for any $s_0 \in S$, hence $I(\pi, q) \leqq I(\pi^*, q)$ for all π, which implies π^* is a q-optimal strategy. Similary we get ${}^j\pi^*$ is a ${}^jq$-optimal strategy for each j. This completes the proof.

The following Corollary is an immediate consequence of Theorem 4.3 (c).

COROLLARY. *If there is a ${}^jq$-optimal strategy for each $j \geqq 0$, then there is a q-optimal Markov strategy.*

We shall say that actions a and b are *equivalent at* (s, q_j) if $r(s, a, \cdot) = r(s, b, \cdot)$ and $q_j(\cdot \mid s, a) = q_j(\cdot \mid s, a)$, i.e. if $T_{aj}u(s) = T_{bj}u(s)$ for all $u \in M(S)$. We shall say that actions a and b are *equivalent at* (s, q) if $r(s, a, \cdot) = r(s, b, \cdot)$ and $q_j(\cdot \mid s, a) = q_j(\cdot \mid s, a)$ for all q_j in q, i.e. if $T_{aj}u(s) = T_{bj}u(s)$ for all $u \in M(S)$ and for all q_j in q. For any Markov $\pi = \{f_1, f_2, \cdots\}$ A will be called *essentially countable by* π if for every (s, a) there is an n for which $f_n(s)$ is equivalent to a at (s, q). A will be called *essentially finite by* π if there is a partition of S into Borel sets $S_1, S_2, \cdots$ such that for every (s, a) with $s \in S_n$, at least one of the actions $f_1(s), f_2(s), \cdots, f_n(s)$ is equivalent to a at

(s, q).

LEMMA 4.2. *If A is essentially finite by $\pi=\{f_1, f_2, \cdots\}$, then for any $q_j \in Q(S|SA)$ and any $u \in M(S)$ there exists $\hat{f}_j \in F(\pi)$ for which $\hat{T}_{jj}u = U_j u$ where $(\hat{f}_j, q_j) \sim \hat{T}_{jj}$ and $(\pi, q_j) \sim U_j$.*

PROOF. We let $S_{nj}=\{s \mid T_{mj}u < U_j u$ for $m \leqq n-1$, $T_{nj}u = U_j u\}$. Then $\{s_{nj}, n=1, 2, \cdots\}$ comes to be a partition of S, since in this case $U_j u = \max_n T_{nj}u$. We set $\hat{f}_j = f_n$ on S_{nj}. Then for $s \in S_{nj}$, $\hat{T}_{jj}u = T_{nj}u = U_j u$, which completes the proof.

THEOREM 4.4. (a) *If A is essentially countable by $\pi=\{f_1, f_2, \cdots\}$, there is a (ε, q)-optimal Markov strategy for every $\varepsilon > 0$ and every $q \in Q^*$.*

(b) *If A is essentially finite by π, there is a q-optimal Markov strategy for every $q \in Q^*$.*

PROOF. (a) From the assumption, for any $u \in M(S)$ $\sup_n T_{nj}u = U_j u = \sup_{a \in A} T_{aj}u$ for all j, where $(f_n, q_j) \sim T_{nj}$ and $(\pi, q_j) \sim U_j$. Thus $T_{aj}u \leqq U_j u$ for any $u \in M(S)$, all $a \in A$ and all j.

On the other hand, from Lemma 4.1, for any $p \in P(S)$, $\varepsilon > 0$ and $q \in Q^*$ there is a (p, ε, q)-optimal Markov strategy, say $\hat{\pi}=\{\hat{f}_1, \hat{f}_2, \cdots\}$, which yields that

$$I(\hat{\pi}, q)_{s_0} \geqq I(\pi, q)_{s_0} - \varepsilon \qquad \text{for all } \pi .$$

We let $(\hat{f}_j, q_j) \sim \hat{T}_{jj}$, then we have $\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{jj}u \leqq U_1 U_2 \cdots U_j u$ for all j. Letting $j \to \infty$ yields $I(\hat{\pi}, q) \leqq u^*$, where $(\pi, q) \sim u^*$.

Thus we have

$$I(\pi, q)_{s_0} \leqq I(\hat{\pi}, q)_{s_0} + \varepsilon \leqq u^*(s_0) + \varepsilon \qquad \text{for all } \pi .$$

Letting $\varepsilon \to 0$ yields that $I(\pi, q)_{s_0} \leqq u^*(s_0)$ for all π. This holds for all $s_0 \in S$, so that $I(\pi, q) \leqq u^*$ for all π.

In virtue of Theorem 4.3 (b), there is a $\tilde{\pi} \in G(\pi)$ for which $I(\tilde{\pi}, q) \geqq u^* - \varepsilon$. Therefore we have $I(\tilde{\pi}, q) \geqq I(\pi, q) - \varepsilon$ for all π. This $\tilde{\pi}$ is (ε, q)-optimal.

(b) From Lemma 4.2, there is a Markov strategy $\hat{\pi}=\{\hat{f}_1, \hat{f}_2, \cdots\}$ with $\hat{f}_j \in F(\pi)$ for $j=1, 2, \cdots$, such that

$$\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{jj}u = U_1 U_2 \cdots U_j u .$$

Letting $j \to \infty$ yields $I(\hat{\pi}, q) = u^*$, where $(\pi, q) \sim u^*$.

But in the proof of (a) we see that $u^* \geqq I(\pi)$ for all π. Thus we have $I(\hat{\pi}, q) \geqq I(\pi)$ for all π, which completes the proof.

§ 5. *l*-stationary strategy.

In this section we shall be concerned with the case when it is known to us that q_j varies in a cyclic manner.

We shall say that q is *l-stationary* if there exist q_j's $\in Q(S|SA)$, $(j=1, 2, \cdots, l)$, such that $q=\{\bar{q}, \bar{q}, \cdots\}$ where $\bar{q}=\{q_1, q_2, \cdots, q_l\}$. An l-stationary q is denoted by $\bar{q}^{(\infty)}$. We shall say that a Markov strategy π is *l-stationary* if there exist degenerate f_j's, $(j=1, 2, \cdots, l)$, such that $\pi=\{\bar{f}, \bar{f}, \cdots\}$ where $\bar{f}=\{f_1, f_2, \cdots, f_l\}$. An l-stationary strategy is denoted by $\bar{f}^{(\infty)}$.

For any Markov $\pi=\{f_1, f_2, \cdots\}$ we let U_j the operator associated with (π, q_j) for

$j=1, 2, \cdots, l$. We shall call $\bar{U}_l \equiv U_1 U_2 \cdots U_l$ the operator associated with $(\pi, \bar{q})$ where $\bar{q}=\{q_1, q_2, \cdots, q_l\}$. $(\pi, \bar{q}) \sim \bar{U}_l$ means that $\bar{U}_l$ is the operator associated with $(\pi, \bar{q})$.

The following Lemma is immediate from the definition of U_j.

LEMMA 5.1. (a) *U_j is monotone for each j.*

(b) *For any constant c, $U_j(u+c)=U_j u+\beta c$.*

THEOREM 5.1. *$\bar{U}_l$, the operator associated with $(\pi, \bar{q})$, is a contraction with a contraction coefficient β^l, i.e. $\|\bar{U}_l u-\bar{U}_l v\| \leqq \beta^l \|u-v\|$, where $\|u\|=\sup_s |u(s)|$.*

PROOF. From the difinition of a norm we have $v \leqq u+\|u-v\|$. In virtue of Lemma 5.1 it follows that

$$U_l v \leqq U_l u+\beta\|u-v\| ,$$

and so

$$U_l v-U_l u \leqq \beta\|u-v\| .$$

Similarly we have

$$U_{l-1}U_l v-U_{l-1}U_l u \leqq \beta^2\|u-v\| ,$$

and inductively we have

$$U_1 U_2 \cdots U_l v-U_1 U_2 \cdots U_l u \leqq \beta^l\|u-v\| ,$$

which shows that $\bar{U}_l v-\bar{U}_l u \leqq \beta^l\|u-v\|$. Interchange u and v to obtain $\bar{U}_l u-\bar{U}_l v \leqq \beta^l\|u-v\|$. Thus we have $\|\bar{U}_l u-\bar{U}_l v\| \leqq \beta^l\|u-v\|$, completing the proof.

The general properties of optimal plans for l-stationary q are contained in the following theorem.

THEOREM 5.2. (a) *For any Markov π and any l-stationary $q=\bar{q}^{(\infty)}$, let $(\pi, \bar{q}) \sim \bar{U}_l$. And let $\bar{u}_l^*$ be a fixed point of $\bar{U}_l$. Then $I(\hat{\pi}, \bar{q}^{(\infty)}) \leqq \bar{u}_l^*$ for every $\hat{\pi} \in G(\pi)$.*

(b) *For any Markov π and any l-stationary $q=\bar{q}^{(\infty)}$, let $(\pi, \bar{q}) \sim \bar{U}_l$. And let $\bar{u}_l^*$ be a fixed point of $\bar{U}_l$. Then for every $\varepsilon>0$ there is a l-stationary strategy $\bar{f}^{(\infty)} \in G(\pi)$ for which $I(\bar{f}^{(\infty)}, \bar{q}^{(\infty)}) \geqq \bar{u}_l^*-\varepsilon$.*

(c) *For any $p \in P(S)$, $\varepsilon>0$ and l-stationary $q=\bar{q}^{(\infty)}$, there is a $(p, \varepsilon, \bar{q}^{(\infty)})$-optimal l-stationary strategy.*

(d) *For any $\varepsilon \geqq 0$, if there is a $(\varepsilon, (q_j, q_{j+1}, \cdots, q_l, \bar{q}^{(\infty)}))$-optimal strategy for $j=1, 2, \cdots, l$, then there is a $(\varepsilon/(1-\beta), \bar{q}^{(\infty)})$-optimal l-stationary strategy, where $\bar{q}=\{q_1, q_2, \cdots, q_l\}$.*

(e) *Let q be l-stationary. Then ${}^j\pi^*$ is ${}^jq$-optimal for $j=1, 2, \cdots, l$ if and only if the expected reward of π^* statisfies the optimallity equations; $I({}^{j-1}\pi^*, {}^{j-1}q)=\sup_{a \in A} T_{aj} I({}^j\pi^*, {}^jq)$ for $j=1, 2, \cdots, l$.*

PROOF. (e) is immediate from (e) of Theorem 4.3. We shall prove (a)$\sim$(d) only.

(a) Let $\hat{\pi}=\{q_1, q_2, \cdots\}$ be any π-generated strategy. Let $(g_j, q_j) \sim \hat{T}_{jj}$. Then, as stated in the proof of Theorem 4.3 (a), if holds that $\hat{T}_{11}\hat{T}_{22} \cdots \hat{T}_{jj} u \to I(\pi, \bar{q}^{(\infty)})$ as $j \to \infty$ for all $u \in M(S)$.

Since $g_i \in F(\pi)$ for $i=1, 2, \cdots$, from Theorem 4.2 (a), it follows that

$$\hat{T}_{ml+1,1}\hat{T}_{ml+2,2} \cdots \hat{T}_{(m+1)l,l} u \leqq U_1 U_2 \cdots U_l u=\bar{U}_l u$$

for each m and for any $u \in M(S)$. Thus in particular

$$\hat{T}_{ml+1,1}\hat{T}_{ml+2,2} \cdots \hat{T}_{(m+1)l,l}\bar{u}_l^* \leqq \bar{U}_l \bar{u}_l^*=\bar{u}_l^*$$

for each m. Inductively we have

$$\prod_{m=0}^{n}(\hat{T}_{ml+1,1}\hat{T}_{ml+2,2}\cdots\hat{T}_{(m+1)l,l})\bar{u}_l^* \leqq \bar{u}_l^* .$$

Letting $n\to\infty$ yields that $I(\hat{\pi}, \bar{q}^{(\infty)})\leqq\bar{u}_l^*$. This hold for all $\hat{\pi}\in G(\pi)$.

(b) Let $\varepsilon'=\varepsilon(1-\beta)$. In accordance with Theorem 4.2 (b) there is an $\hat{f}_l\in F(\pi)$ corresponding to $(q_l, \bar{u}_l^*)$ for which

$$\hat{T}_{ll}\bar{u}_l^* \geqq U_l\bar{u}_l^*-\varepsilon' ,$$

where $(\hat{f}_l, q_l)\sim\hat{T}_{ll}$ and $(\pi, q_l)\sim U_l$. Similarly there is an $\hat{f}_{l-1}\in F(\pi)$ corresponding to $(q_{l-1}, U_l\bar{u}_l^*-\varepsilon')$ for which

$$\hat{T}_{l-1,l-1}(U_l\bar{u}_l^*-\varepsilon')\geqq U_{l-1}(U_l\bar{u}_l^*-\varepsilon')-\varepsilon' ,$$

where $(\hat{f}_{l-1}, q_{l-1})\sim\hat{T}_{l-1,l-1}$ and $(\pi, q_{l-1})\sim U_{l-1}$. Thus we have

$$\hat{T}_{l-1,l-1}\hat{T}_{ll}\bar{u}_l^* \geqq U_{l-1}U_l\bar{u}_l^*-\varepsilon'(1+\beta)$$

by Lemma 5.1 (b).

By backward induction we obtain

$$\hat{T}_{11}\hat{T}_{22}\cdots\hat{T}_{ll}\bar{u}_l^* \geqq U_1U_2\cdots U_l\bar{u}_l^*-\varepsilon'(1+\beta+\cdots+\beta^{l-1})=\bar{u}_l^*-\varepsilon'(1+\beta+\cdots+\beta^{l-1}) .$$

Again inductively we have

$$(\hat{T}_{11}\hat{T}_{22}\cdots\hat{T}_{ll})^n\bar{u}_l^* \geqq \bar{u}_l^*-\varepsilon'(1+\beta+\cdots+\beta^{nl-1}) \qquad \text{for all} \quad n\geqq 1 ,$$

from which it follows that $I(\bar{f}^{(\infty)}, \bar{q}^{(\infty)})\geqq\bar{u}_l^*-\varepsilon'/(1-\beta)=\bar{u}_l^*-\varepsilon$. This $\bar{f}^{(\infty)}$ is obviously l-stationary and π-generated.

(c) In virtue of Lemma 4.1 there is a $(p, \varepsilon/2, \bar{q}^{(\infty)})$-optimal Markov $\pi=\{f_1, f_2, \cdots\}$ for any $p\in P(S)$, $\varepsilon>0$ and any l-stationary $\bar{q}^{(\infty)}$. Let $(\pi, \bar{q})\sim\bar{U}_l$ and let $\bar{u}_l^*$ be a fixed point of $\bar{U}_l$. Then from (b) there is a π-generated l-stationary $\bar{f}^{(\infty)}$ for which

$$I(\bar{f}^{(\infty)}, \bar{q}^{(\infty)})\geqq\bar{u}_l^*-\varepsilon/2 .$$

From (a) it follows that

$$I(\pi, \bar{q}^{(\infty)})\leqq\bar{u}_l^* .$$

Thus we have

$$I(\bar{f}^{(\infty)}, \bar{q}^{(\infty)})\geqq\bar{u}_l^*-\varepsilon/2\geqq I(\pi, \bar{q}^{(\infty)})-\varepsilon/2 .$$

On the other hand, since π is $(p, \varepsilon/2, \bar{q}^{(\infty)})$-optimal, we have

$$P\{I(\pi, \bar{q}^{(\infty)})\geqq I(\pi', \bar{q}^{(\infty)})-\varepsilon/2\}=1 \qquad \text{for all} \quad \pi' .$$

Hence it follows that

$$P\{I(\bar{f}^{(\infty)}, \bar{q}^{(\infty)})\geqq I(\pi', \bar{q}^{(\infty)})-\varepsilon\}=1 \qquad \text{for all} \quad \pi' .$$

This $\bar{f}^{(\infty)}$ is $(p, \varepsilon, \bar{q}^{(\infty)})$-optimal l-stationary.

(d) Let $\tilde{q}_j=(q_j, q_{j+1}, \cdots, q_l)$. Assume that there is a $(\varepsilon, \tilde{q}_j, \bar{q}^{(\infty)}))$-optimal strategy, say $\pi^{*j}=\{\pi_{j1}, \pi_{j2}, \cdots\}$, for each $j=1, 2, \cdots, l$.

In accordance with Lemma 2.1, for each $j=1, 2, \cdots$, , there exists a degenerate f_j for which

$$\begin{aligned} I(\pi^{*j}, (\tilde{q}_j, \bar{q}^{(\infty)})) &= \pi_{j1}q_j[r+\beta I({}^1\pi^{*j}, (\tilde{q}_{j+1}, \bar{q}^{(\infty)}))] \\ &\leqq \pi_{j1}q_j[r+\beta\{I(\pi^{*j+1}, (\tilde{q}_{j+1}, \bar{q}^{(\infty)}))+\varepsilon\}] \end{aligned}$$

$$\leqq f_j q_j[r+\beta\{I(\pi^{*j+1}, (\tilde{q}_{j+1}, \bar{q}^{(\infty)}))+\varepsilon\}]$$
$$= T_{jj} I(\pi^{*j+1}, (\tilde{q}_{j+1}, \bar{q}^{(\infty)}))+\beta\varepsilon,$$

where $(f_j, q_j) \curvearrowright T_{jj}$. Since $(\tilde{q}_{l+1}, \bar{q}^{(\infty)}) = \bar{q}^{(\infty)}$ and so $\pi^{*l+1} = \pi^{*1}$, inductively we have

$$T_{11}T_{22} \cdots T_{ll} I(\pi^{*1}, \bar{q}^{(\infty)}) \geqq I(\pi^{*1}, \bar{q}^{(\infty)}) - \varepsilon(\beta+\beta^2+ \cdots +\beta^l).$$

Again inductively we have

$$(T_{11}T_{22} \cdots T_{ll})^n I(\pi^{*1}, \bar{q}^{(\infty)}) \geqq I(\pi^{*1}, \bar{q}^{(\infty)}) - \varepsilon(\beta+\beta^2+ \cdots +\beta^{nl}) \quad \text{for all } n.$$

Letting $n \to \infty$ yiels that

$$I(\bar{f}^{(\infty)}, \bar{q}^{(\infty)}) \geqq I(\pi^{*1}, \bar{q}^{(\infty)}) - \varepsilon\beta/(1-\beta),$$

where $\bar{f} = \{f_1, f_2, \cdots, f_l\}$.

But, since π^{*1} is $(\varepsilon, \bar{q}^{(\infty)})$-optimal, it holds that

$$I(\pi, \bar{q}^{(\infty)}) \leqq I(\pi^{*1}, \bar{q}^{(\infty)}) + \varepsilon \quad \text{for all } \pi.$$

Thus we obtain

$$I(\pi, \bar{q}^{(\infty)}) \leqq I(\bar{f}^{(\infty)}, \bar{q}^{(\infty)}) + \varepsilon/(1-\beta) \quad \text{for all } \pi,$$

which implies that $\bar{f}^{(\infty)}$ is a $(\varepsilon/(1-\beta), \bar{q}^{(\infty)})$-optimal l-stationary strategy.

§ 6. Strong optimality.

In this section we shall define another optimality more strengthened than that of preceding sections, and we shall investigate an existence and some properties of an optimal strategy in this sence.

We let $v_q^* = \sup_\pi I(\pi, q)$, π^* will be called a *strong* (p, ε, q)*-optimal* strategy if $p\{I(\pi^*, q) \geqq v_q^* - \varepsilon\} = 1$.

The following two theorems are immediate respectively from Theorem 4.3 and from the proof of Theorem 8.1 in [3] by replacing $(q, q, \cdots)$ by $(q_1, q_2, \cdots)$.

THEOREM 6.1. *For any $p \in P(S)$, any $q \in Q^*$ and any strategy π, there is a Markov strategy $\tilde{\pi}$ for which $pI(\tilde{\pi}, q) \geqq pI(\pi, q)$.*

THEOREM 6.2. *For any $p \in P(S)$, $\varepsilon > 0$ and any $q \in Q^*$ there is a strong (p, ε, q)-optimal strategy.*

The following theorem is an alalogy of Theorem 6.2 in Strauch [3]. We shall prove it directly from (a) and (b) of our Theorem 4.3, whereas Strauch did by using the property of the "conservation".

THEOREM 6.3. *For any sequence of Markov strategies $\{\pi^j, j=1, 2, \cdots\}$, $\varepsilon > 0$ and any $q \in Q^*$, there is a Markov $\hat{\pi}$ for which $I(\hat{\pi}, q) \geqq \sup_j I(\pi^j, q) - \varepsilon$.*

PROOF. We can find a Markov strategy π such that $\pi^j \in G(\pi)$ for $j=1, 2, \cdots$. Let $(\pi, q) \curvearrowright u^*$ for this π.

Then, from (b) of Theorem 4.3, there is a Markov strategy $\hat{\pi} \in G(\pi)$ such that

$$I(\hat{\pi}, q) \geqq u^* - \varepsilon.$$

Since $\pi^j \in G(\pi)$ for $j=1, 2, \cdots$, from (a) of Theorem 4.3, it holds that

$$u^* \geqq I(\pi^j, q) \quad \text{for } j=1, 2, \cdots,$$

which yield that

$$u^* \geqq \sup_j I(\pi^j, q) .$$

Thus we have

$$I(\hat{\pi}, q) \geqq u^* - \varepsilon \geqq \sup_j I(\pi^j, q) - \varepsilon ,$$

which completes the proof.

Now let us show that a strong (p, ε, q)-optimal strategy exists, in fact, among the Markov strategies, by making use of the above theorems.

THEOREM 6.4. *For any $p \in P(S)$, $\varepsilon > 0$ and any $q \in Q^*$, there is a strong (p, ε, q)-optimal Markov strategy.*

PROOF. We let $v_q^* = \sup_\pi I(\pi, q)$. Then, in virtue of Theorem 6.2, there exists a strategy $\hat{\pi}$ (not necessarily Markov) for which $p\{I(\hat{\pi}, q) \geqq v_q^* - \varepsilon\} = 1$. Hence we have

$$pI(\hat{\pi}, q) \geqq pv_q^* - \varepsilon .$$

From Theorem 6.1 there is a Markov strategy π^* for this $\hat{\pi}$ such that

$$pI(\pi^*, q) \geqq pI(\hat{\pi}, q) .$$

Therefore it follows that there exists a Markov strategy π^* for every $\varepsilon > 0$ such that

$$pI(\pi^*, q) \geqq pv_q^* - \varepsilon .$$

Thus for each integer $m \geqq 1$ we can find a Markov strategy π^m such that

$$pI(\pi^m, q) \geqq pv_q^* - 1/m .$$

Let $v_q' = \sup_m I(\pi^m, q)$, then it folows that $pv_q' \geqq pv_q^* - 1/m$. Letting $m \to \infty$ yields that $pv_q' \geqq pv_q^*$. But $pv_q^* \geqq pv_q'$, since $v_q^* \geqq v_q'$ from the definitions of v_q^* and v_q'. Consequently $pv_q^* = pv_q'$. Again from that $v_q^* \geqq v_q'$, we have

$$p\{v_q^* = v_q'\} = 1 .$$

In virtue of Theorem 6.3, for $\{\pi^m\}$ there is a Markov $\tilde{\pi}$ such that

$$I(\tilde{\pi}, q) \geqq v_q' - \varepsilon .$$

Finally we have

$$p\{I(\tilde{\pi}, q) \geqq v_q^* - \varepsilon\} = 1 ,$$

which say that this $\tilde{\pi}$ is a strong (p, ε, q)-optimal Markov strategy.

§ 7. Improvements ef strategies.

In this section we shall be concerned with several methods of strategy improvements.

THEOREM 7.1. *If for a Markov strategy $\hat{\pi} = \{\hat{f}_1, \hat{f}_2, \cdots\}$*

$$I((\hat{f}_j, {}^j\pi), {}^{j-1}q) \geqq I({}^{j-1}\pi, {}^{j-1}q) \qquad \textit{for} \quad j = 1, 2, \cdots ,$$

then $I(\hat{\pi}, q) \geqq I(\pi, q)$.

PROOF. We let $(\hat{f}_j, q_j) \sim \hat{T}_{jj}$, then from the assumption we get

$$\hat{T}_{jj} I({}^j\pi, {}^jq) \geqq I({}^{j-1}\pi, {}^{j-1}q) \qquad \text{for} \quad j = 1, 2, \cdots .$$

Consequently

$$\hat{T}_{j-1,j-1}\hat{T}_{jj} I({}^j\pi, {}^jq) \geqq \hat{T}_{j-1,j-1} I({}^{j-1}\pi, {}^{j-1}q) \geqq I({}^{j-2}\pi, {}^{j-2}q) \qquad \text{for} \quad j = 1, 2, \cdots ,$$

and inductively

$$\hat{T}_{11}\hat{T}_{22}\cdots\hat{T}_{jj}I({}^{j}\pi, {}^{j}q)\geqq I({}^{0}\pi, {}^{0}q)=I(\pi, q) \qquad \text{for} \quad j=1, 2, \cdots .$$

Letting $j\to\infty$ yields $I(\hat{\pi}, q)\geqq I(\pi, q)$.

THEOREM 7.2. *If* $I({}^{n-1}\pi^*, {}^{n-1}q)\geqq I((f, {}^{n}\pi^*), {}^{n-1}q)$ *for all degenerate* f *and for every* n, *then* π^* *is* q*-optimal.*

PROOF. Let $\pi=\{f_1, f_2, \cdots\}$ be any Markov strategy. By the assumption we have

$$I({}^{n-1}\pi^*, {}^{n-1}q)\geqq I((f_n, {}^{n}\pi^*), {}^{n-1}q) \qquad \text{for} \quad n=1, 2, \cdots .$$

Let $(f_j, q_j)\sim T_{jj}$, then plainly

$$T_{11}T_{22}\cdots T_{n-1,n-1}I({}^{n-1}\pi^*, {}^{n-1}q)\geqq T_{11}T_{22}\cdots T_{n-1,n-1}I((f_n, {}^{n}\pi^*), {}^{n-1}q) \qquad \text{for} \quad n=1, 2, \cdots ,$$

which implies that

$$I((f_1, f_2, \cdots, f_{n-1}, {}^{n-1}\pi^*), q)\geqq I((f_1, f_2, \cdots, f_n, {}^{n}\pi^*), q) \qquad \text{for} \quad n=1, 2, \cdots .$$

And inductively

$$I(\pi^*, q)\geqq I((f_1, f_2, \cdots, f_n, {}^{n}\pi^*), q) \qquad \text{for} \quad n=1, 2, \cdots .$$

Hence we have $I(\pi^*, q)\geqq I(\pi, q)$ by letting $n\to\infty$. This holds for any Markov strategy π, which completes the proof.

Now we shall define $G_j(s, \pi)$. Let $(f\equiv a, q_j)\sim T_{aj}$, then $G_j(s, \pi)$ is defined by

$$G_j(s, \pi)=\{a \mid T_{aj}I({}^{j}\pi, {}^{j}q)>I({}^{j-1}\pi, {}^{j-1}q)\} \qquad \text{for} \quad j=1, 2, \cdots , \text{ and for } s\in S .$$

We shall conclude with the following theorem.

THEOREM 7.3. (a) *If for every* $s\in S$ *and for every* j $G_j(s, \pi)$ *is an emply set, then* π *is a* q*-optimal strategy.*

(b) *Let* $\pi=\{f_1, f_2, \cdots\}$ *be any Markov strategy. If for each* j $g_j(s_0)\in G_j(s_0, \pi)$ *for some* s_0 *and* $g_j(s)=f_j(s)$ *for* s *such that* $g_j(s)\notin G_j(s, \pi)$, *then* $I(\hat{\pi}, q)\geqq I(\pi, q)$ *where* $\hat{\pi}=\{q_1, q_2, \cdots\}$.

PROOF. (a) From the assumption it follows that

$$I({}^{j-1}\pi, {}^{j-1}q)\geqq I((g_j, {}^{j}\pi), {}^{j-1}q) \qquad \text{for all } g_j \text{ and every } j .$$

Hence in virtue of Theorem 7.2 π is q-optimal.

(b) By the assumption, for each j we have

$$I((g_j, {}^{j}\pi), {}^{j-1}q)_{s_0}>I({}^{j-1}\pi, {}^{j-1}q)_{s_0}$$

and

$$I((g_j, {}^{j}\pi), {}^{j-1}q)_{s}=I({}^{j-1}\pi, {}^{j-1}q)_{s}$$

for s such that $g_j(s)\notin G_j(s, \pi)$, which implies that

$$I((g_j, {}^{j}\pi), {}^{j-1}q)\geqq I({}^{j-1}\pi, {}^{j-1}q) \qquad \text{for} \quad j=1, 2, \cdots .$$

Thus from Theorem 7.1 $I(\hat{\pi}, q)\geqq I(\pi, q)$.

References

[1] R. BELLMAN, *Dynamic Programming*, Princeton Univ. Press, (1957).
[2] D. BLACKWELL, *Discounted dynamic programming*, Ann. Math. Statist., **36** (1965), 226-235.
[3] R. E. STRAUCH, *Negative dynamic programming*, Ann. Math. Statist., **37** (1966), 871-890.
[4] M. LOÈVE, *Probability Theory*, Van Nostrand, Princeton, (1960).