Nihonkai Math. J.
Vol.2(1991), 71-91

Average-time Criterion for Vector-valued Markovian Decision Systems

Hang-Chin Lai and Kensuke Tanaka

Abstract. The optimization problem of a Markovian decision model for vector-valued loss function is investigated under the discrete average-time criterion. For a convex cone D, a D-optimal policy is defined as a policy which takes minimal point, with respect to the ordering induced from D, among the set of all cluster points of the expected average vector losses. By using a numerical modification, we prove that a D-optimal policy exists in the vector-valued decision system for the average criterion. Conversely, under some additional conditions, a D-optimal policy is also an optimal policy of the modified decision system for numerical loss function constructed by a weighted vector.

1. Introduction

In a dynamic programming problem, the Markovian models on

AMS 1980 subject classification: 90C47,90D35

infinite horizon have been studied by many authors. Much of the earlier works in this area were done by Blackwell [2],[3] and Strauch [17]. Hinderer [11] gave an extensive account of Markovian decision process with discrete time parameter. Previously, average criterion for Markovian decision processes were investigated by Ross [14], [15], [16], Tijims [19], etc. However, these papers are restricted in the Markovian models with real-valued loss function. Recently, under the influence of multiobjective optimization theory (see Yu [21] and Tanino and Sawaragi [18]), many authors studied vector-valued Markovian models (cf. Furukawa [5], Hartley [7], Henig [8], [9] and White [20], etc.). However, they restricted themselves in the case of discounted model. Up till now the vector-valued Markovian model under average criterion in general state and action spaces has not yet been formulated.

In this paper, we will study the optimization problem of vector-valued Markovian decision model under the time average criterion as opposed to the criterion of numerical Markovian decision models. We will show the existence of a D-optomal policy which minimizes the set of all cluster points of the expected time average vector losses. We show that the D-optimal policy is indeed more general than the optimal solution of the usual optimization problem. To this end, we introduce a vector as a weighted factor in the positive polar cone of a convex cone D, and modify the vector-valued Markovian decision model to be a new decision system with numerical loss function. It follows that an optimal policy of modified decision model is also a

D-optimal policy of the vector decision one. Further, from the convexity of the set of all cluster points for all policies, and under approprate conditions, the converse version of the above result is also true, that is, a D-optimal policy of the vector-valued decision model is an optimal policy of the modified decision system with numerical loss function.

This paper is organized in the following way. In Section 2, we formulate the average time criterion of vector-valued Markovian decision model. In Section 3, we present some notations and definitions for D-optimal policy. Section 4 is the main part of this paper, and we show that the D-optimal policy exists in our decision system. We also establish the relation between the vector decision system and the modified decision system.

2. Formulation of vector-valued Markovian decision model with an expected average reward criterion

A vector-valued decision system of Markovian model is specified by a set of five elements

$$(S, A, F, q, r). \qquad (2.1)$$

where

(i) S is a non-empty Borel subset of a Polish space (that is, complete separable metric space), the state space of the decision system.

(ii) A is a non-empty Borel subset of a Polish space, the action space.

(iii) F is a Borel measurable multifunction which associates each state $s \in S$, a non-empty feasible set $F(s) \subset A$ of actions.

(iv) q is a transition probability measure $q(\cdot|s,a)$ on the Borel subsets of S for any $(s,a) \in GrF=\{(s,a)|a \in F(s)\}$. The graph of multifunction F, is a Borel subset of $S\times A$. For a Borel subset $B \subset S$, the mapping $q(B|\cdot,\cdot)$: $GrF \to R$ is a Borel function in $(s,a) \in GrF$. This function $q(B|s,a)$ plays the law of motion in the decision system.

(v) $r(\cdot,\cdot) = (r_1, r_2, \cdots , r_m)(\cdot,\cdot)$: $GrF \to R^m$ is an m-dimensional vector-valued function, it is a one step vector loss function.

Note that the feasible set F(s) of actions depends only on the state $s \in S$, and $q(\cdot|s,a)$ is independent of the time. A policy π is defined as an infinite sequence $(\pi_1, \pi_2, \cdots , \pi_t, \cdots)$, where each element π_t is a conditional probability on A under the known histories $H_1 = S$, $H_t = (GrF)H_{t-1}$, $t \geq 2$, the set of possible histories up to the t-th stage. Let s_t and a_t denote the t-th state and the t-th action, respectively. Assume that π_t satisfies the constraint $\pi_t(F(s_t)|h_t) = 1$ for any given history $h_t = (s_1,a_1,s_2,a_2, \cdots ,s_t)$ in the decision system. A policy π is said to be stationary if there exists a Borel measurable mapping $f:S \to A$ such that $f(s) \in F(s)$ for all $s \in S$, and $\pi_t(f(s_t)|h_t) = 1$ for any given history $h_t = (s_1,a_1,s_2,a_2, \cdots s_t)$.

Throughout this paper, we let Π denote the set of all

policies. Let R^m be the range space of the vector loss function r which is an ordered vector space ordering by a pointed convex cone D, that is, a convex cone D such that $D \cap (-D) = \{\theta\}$, where θ denotes the zero vector.

Now, we interpret the decision process as follows. If a policy $\pi = (\pi_1, \pi_2, \cdots\cdots, \pi_t, \cdots\cdots)$ is applied for a successive discrete time $t = 1, 2, 3, \cdots\cdots$, we observe the subsequent variant states $s_t \in S$ of the decision system, and through proper analysis, then choose an action $a_t \in F(s_t)$ under the conditional probability π_t for the past history h_t up to the time t. Such an action will incur one step vector loss function $r(s_t, a_t)$. Then, the decision process moves to a new state s_{t+1} according to the transition probability measure $q(\cdot | s_t, a_t)$, and the process of the decision system is then developed from the state s_{t+1}. So, given an initial distribution $p(\cdot)$ on S, any policy π together with a transition probability q, we define a probability measure p_t^{π} on the set $(S \times A)^t = S \times A \times S \times A \times \cdots\cdots S \times A$ up to time t (cf. Hinderer [11,p.80]), that is, $p_t^{\pi} = p\pi_1 q\pi_2 \cdots\cdots \pi_{t-1} q\pi_t$. Whence, if we use the policy $\pi = (\pi_1, \pi_2, \cdots)$, then, at the time t, the expected vector loss is given by

$$E_{\pi}[r(s_t, a_t)] = \int_{(S \times A)^t} r(s_t, a_t) dp_t^{\pi}(h_t) \qquad (2.2)$$

$$= (\cdots, \int_{(S \times A)^t} r_i(s_t, a_t) dp_t^{\pi}(h_t), \cdots)_{i=1}^{m}.$$

Therefore the total expected vector loss, up to the time n, is given by

$$\Phi^n(\pi) = \sum_{t=1}^{n} E_\pi[r(s_t, a_t)] \qquad (2.3)$$

$$= (\cdots, \sum_{t=1}^{n} E_\pi[r_i(s_t, a_t)], \cdots)_{i=1}^{m}.$$

In this decision system, we wish to find a policy π^* which minimizes the set of all cluster points of $\{ \Phi^n(\pi)/n \mid n = 1, 2, 3, \cdots \text{ for } \pi \in \Pi \}$ in R^m with respect to an order cone. This means that no other policy yields a smaller cluster point under the ordered structure.

For a weighted vector $\hat{d} = (\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_m)$, we have

$$\langle \hat{d}, \Phi^n(\pi)/n \rangle = \sum_{t=1}^{n} \langle \hat{d}, E_\pi[r(s_t, a_t)]/n \rangle$$

$$= \sum_{t=1}^{n} \sum_{i=1}^{m} \hat{d}_i E_\pi[r_i(s_t, a_t)]/n. \qquad (2.4)$$

We will find a π^* which minimizes (2.4) over $\pi \in \Pi$. For this purpose, we will show that, for any $\pi \in \Pi$,

$$\langle \hat{d}, \Phi^n(\pi^*)/n \rangle \leq \langle \hat{d}, \Phi^n(\pi)/n \rangle \quad \text{for } n = 1,2, \cdots, \qquad (2.5)$$

Let $C(\pi) \subset R^m$ be the set of all cluster points of $\{\Phi^n(\pi)/n, n = 1,2, \cdots\}$. We shall prove that, as $n \to \infty$ in (2.5), there exists

$\pi^* \in \Pi$ with $\Phi(\pi^*) \in C(\pi^*)$ such that

$$\langle \hat{d}, \Phi(\pi^*) \rangle \leq \langle \hat{d}, \Phi(\pi) \rangle \quad \text{for all} \quad \Phi(\pi) \subset C = \bigcup_{\pi \in \Pi} C(\pi).$$

This policy π^* is a D-optimal policy in our decision system, and $\Phi(\pi^*)$ is a minimal point of the set C in R^m with respect to the cone D. To show the existence of a D-optimal policy in our decision system, we proceed to the next section.

3. D-optimal policy in the decision system

Let clE and intE be respectively the closure and the interior of a subset E in R^m. For any subset E in R^m, the positive polar cone of E is given by

$$E^* = \{ y \in R^m \mid \langle x, y \rangle \geq 0 \text{ for all } x \in E \}, \qquad (3.1)$$

where $\langle x, y \rangle$ is the inner product of x and y in R^m. A cone generated by a subset E in R^m is defined by the set :

$$[E] = \{ y \in R^m \mid y = \lambda x, \ x \in E, \ \lambda \in R_+ \}, \qquad (3.2)$$

where R_+ is the set of all nonnegative real numbers.

Now, consider a subset $L \subset R^m$ such that

(i) $\theta = (0, 0, \cdots, 0) \notin L$ and $e = (1, 1, \cdots, 1) \in L$.

(ii) $L^+ = \{ y \in R^m \mid \langle x,y \rangle > 0 \text{ for all } x \in L \} \neq \phi$.

(iii) $L \cup \{\theta\} = D$ is a convex cone with vertex at the origin θ.

Note that this L is a convex cone without vertex θ in R^m and D denotes a convex cone which determine a partial order in R^m. We will use the sets L and D throughout the paper. Further, we introduce a set of weighted vectors by

$$L_1^+ = \{ \hat{d} = (\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_m) \in L^+ \mid \sum_{i=1}^{m} \hat{d}_i = 1 \}. \quad (3.3)$$

Then, $L_1^+ \neq \theta$ since $e \in L$. Let π be any policy in Π,

$$E(\pi) = \{ \Phi^n(\pi)/n, \quad n=1,2,\cdots \} \quad (3.4)$$

be the set of all time average vector expected rewards up to $n = 1, 2, \cdots$, where

$$\Phi^n(\pi) = \sum_{t=1}^{n} E_\pi[r(s_t, a_t)],$$

and $E_\pi[r(s_t,a_t)]$ is given in (2.2). Denote by $C(\pi)$ the set of all cluster points of the set $E(\pi)$, that is, for each $\Phi(\pi) \in C(\pi)$, there exists a subsequence $\{n_k\}$ of $\{n\}$ such that

$$\Phi^{n_k}(\pi)/n_k \to \Phi(\pi) \quad \text{as } k \to \infty.$$

The basic problem is to find an optimal policy $\pi^* \in \Pi$ such that $\Phi(\pi^*) \in C(\pi^*)$ minimizes $C = \bigcup_{\pi} C(\pi)$ for our decision system

with respect to the convex cone D in R^m.

Definition 3.1 A policy π^* is said to be D-optimal policy for time average criterion of the decision system (2.1) if there is no other policy π such that

$$\Phi(\pi^*) \in \Phi(\pi) + L \quad \text{for some } \Phi(\pi) \in C. \tag{3.5}$$

Remark 3.1 For a closed convex cone E, if L = intE(resp. L = E - {θ}), the policy π^* in (3.5) is usually called a E-weak (resp.E-strong) optimal policy (see Aubin [1,p.295]).

Note that the D-optimal policy π^* need not be unique.

Let Ext[C|D] be the set of all minimal cluster points in C for all D-optimal policies. This is similar to the terminology of all D-extreme points of C given by Yu (see p.336,Definition 4.1 in Yu[21]).

The following lemma is essential for our later works.

Lemma 3.1 Let $\hat{d} = (\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_m) \in L_1^+$. If π^* is a policy such that $\Phi(\pi^*) \in C(\pi^*)$ and

$$< \hat{d}, \Phi(\pi^*) > \leq < \hat{d}, \Phi(\pi) > \quad \text{for all } \Phi(\pi) \in C.$$

Then, π^* is a D-optimal policy of the decision system (2.1).

Proof. Assume to the contrary that π^* is not a D-optimal policy. Then there exists a policy π with $\Phi(\pi) \in C(\pi)$ such that

$$\Phi(\pi^*) \in \Phi(\pi) + L.$$

Thus, there is a $d \in L$ such that $\Phi(\pi^*) = \Phi(\pi) + d$. Then, for

$\hat{d} \in L_1^+$,

$$< \hat{d}, \Phi(\pi^*) > = < \hat{d}, \Phi(\pi) > + < \hat{d}, d >.$$

Since $< \hat{d}, d > > 0$ by the definition of L_1^+, it follows that

$$< d, \Phi(\pi^*) > > < \hat{d}, \Phi(\pi) >.$$

This contradicts our hypothesis.

Observing $\hat{d} = (\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_m) \in L_1^+$ in Lemma 3.1, we get the existence of a D-optimal policy for the system (2.1). Since

$$< \hat{d}, \Phi(\pi)/n > = \sum_{i=1}^{m} \sum_{t=1}^{n} \hat{d}_i E_\pi [r_i(s_t, a_t)]/n$$

$$= \sum_{t=1}^{n} E_\pi [\sum_{i=1}^{m} \hat{d}_i r_i(s_t, a_t)]/n$$

$$= \sum_{t=1}^{n} E_\pi [< \hat{d}, r(s_t, a_t) >]/n, \qquad (3.6)$$

we see that $< \hat{d}, r(s_t, a_t) >$ is essential in the vector-valued decision system, but the numerical number $< \hat{d}, \Phi^n(\pi)/n >$ is different from the usual optimization problem. Whence, we consider a modified decision system as the following form

$$(S, A, F, q, < \hat{d}, r >). \qquad (3.7)$$

In this modified decision system, it is only $< \hat{d}, r >$ in place of r in the vector decision system (2.1). As we have known, the real-valued Markovian decision process with expected average reward criterion has been developed by some authors (cf. Ross

[14], [15], [16] and Tijms [19]). Employing their arguments, we see that an optimal policy exists in the modified decision system. Then, Lemma 3.1 is applicable to derive the existence of a D-optimal policy in the original decision system (2.1).

4. The existence of a D-optimal policy in the dynamic decision system

Let B(S) be the set of all bounded Borel measurable real functions on S. We need some additional assumptions on F, r and q in the decision system (2.1). Let $P_k(A)$ be the collection of compact subsets of the Polish space A which is an action space A in (2.1). We impose the following assumptions.

(A1) $F: S \to P_k(A)$ is a Borel measurable multifunction.

(A2) Each component $r_i(s,a)$, $i=1,2,3 \cdots,m$, of the vector-valued loss function $r(s,a)$ is bounded on GrF, and is continuous in $a \in F(s)$ for each $s \in S$.

(A3) For any $u \in B(S)$, the integral functional

$$(s,a) \in GrF \to \int_S u(x)dq(x|s,a)$$

is lower semicontinuous(l.s.c.) in $a \in F(s)$ for each $s \in S$.

Now, we consider the decision system (3.7) in place of (2.1): for any $\hat{d} \in L_1^+$ (see (3.3)), use the scalar $\langle \hat{d}, r \rangle$

instead of the vector r in (2.1), and write the new decision system as follows:

$$(S, A, F, q, \langle \hat{d}, r \rangle). \qquad (4.1)$$

From the assumptions (A1) and (A2), the loss function $\langle \hat{d}, r \rangle$ in the system (4.1) is bounded on GrF, and is continuous in $a \in F(s)$ for each $s \in S$. Thus, for a weighted vector $\hat{d} \in L_1^+$, we define an operator T on B(S) by

$$Tu(s) = \min_{a \in F(s)} [\langle \hat{d}, r(s,a) \rangle + \int_S u(x) dq(x|s,a)]. \qquad (4.2)$$

Evidently, $Tu \in B(S)$ whenever $u \in B(S)$. For simplicity, we let

$$L(a)u(s) = \langle \hat{d}, r(s,a) \rangle + \int_S u(x) dq(x|s,a). \qquad (4.3)$$

Then, the Eq.(4.2) is simply rewritten as

$$Tu(s) = \min_{a \in F(s)} L(a)u(s).$$

Under the above preparation, we proceed to prove the existence theorem for the D-optimal policy of average criterion on the system (2.1).

Theorem 4.1 If there exists a function u in B(S) and a constant α such that

$$\alpha + u(s) = Tu(s) \qquad \text{for all } s \in S, \qquad (4.4)$$

then, there exists a stationary optimal policy f of average

criterion for the decision system (2.1) such that for each $\Phi(f) \in C(f)$,

$$\Phi(f) \in \mathrm{Ext}[C|D].$$

Furthermore, f is s Borel mapping from S into A such that

$$\alpha + u(s) = L(f(s))u(s), \qquad (4.5)$$

where C(f) denotes the set of all cluster points of E(f) (see (3.4)) for the policy f and $C = \bigcup_{\pi} C(\pi)$.

Proof. From (A2) and (A3), L(a)u(s) in (4.3) is a l.s.c. function in $a \in F(s)$, the minimum in (4.2) is attained by Borel measurable selector f for F since F satisfies (A1), that is (4.5), and so

$$\begin{aligned} \alpha + u(s) &= L(f(s))u(s) \\ &\le L(a)u(s) \qquad \text{for all } (s,a) \in GrF. \end{aligned} \qquad (4.6)$$

This fact can be proved by an argument similar to that of Theorem 2 in Himmelberg et al.[10]. So, using the property of a weighted vector $\hat{d} \in L_1^+$ and (4.3), the expression (4.6) can be rewritten as

$$\begin{aligned} \alpha + \langle \hat{d}, (u(s)) \rangle &= \langle \hat{d}, r(s,f(s)) \rangle + \int_S \langle \hat{d}, (u(x)) \rangle dq(x|s,f(s)) \\ &\le \langle \hat{d}, r(s,a) \rangle + \int_S \langle \hat{d}, (u(x)) \rangle dq(x|s,a), \end{aligned} \qquad (4.7)$$

for all $(s,a) \in GrF$, where $(u(\cdot)) = (u(\cdot), u(\cdot), \cdots, u(\cdot)) \in R^m$.

Let $h_t = (s_1, a_1, s_2, a_2, \cdots\cdots, s_t)$ denote the history of the decision process up to time t, and let $h_t' = (h_t, a_t)$. For any

policy π,

$$E_\pi[\sum_{t=1}^{n}\{((u(s_{t+1})) - E_\pi[(u(s_{t+1}))|h_t^i]\}] = \theta,$$

where $E_\pi[(u(\cdot))|h_t^i] = (\cdots\cdot, E_\pi[u(\cdot)|h_t^i], \cdots\cdot) \in R^m$.
Consequently, for $\hat{d} \in L_1^+$, we get

$$\begin{aligned}
&< \hat{d}, E_\pi[\sum_{t=1}^{n}\{((u(s_{t+1})) - E_\pi[(u(s_{t+1}))|h_t^i]\}] > \\
&= \sum_{t=1}^{n} E_\pi[<\hat{d}, (u(s_{t+1}))> - <\hat{d}, E_\pi[(u(s_{t+1}))|h_t^i]>] \qquad (4.8) \\
&= 0.
\end{aligned}$$

But,

$$\begin{aligned}
&< \hat{d}, E_\pi[(u(s_{t+1}))|h_t^i\} > \\
&= < \hat{d}, \int_S (u(x))dq(x|s_t, a_t) > \\
&= \int_S < \hat{d}, (u(x)) > dq(x|s_t, a_t) \\
&= L(a_t)< \hat{d}, (u(s_t)) > - < \hat{d}, r(s_t, a_t) > \quad \text{(by using (4.3))} \\
&\geq T< \hat{d}, (u(s_t)) > - < \hat{d}, r(s_t, a_t) > \quad \text{(by using (4.2))} \\
&= \alpha + < \hat{d}, (u(s_t)) > - < \hat{d}, r(s_t, a_t) > \quad \text{(by using (4.4)),} \qquad (4.9)
\end{aligned}$$

the equality in (4.9) holds for the policy f since f is determined as the minimal action. Thus, inserting (4.9) into (4.8), we obtain

$$\sum_{t=1}^{n} E_{\pi}[\langle \hat{d}, (u(s_{t+1}))\rangle - \alpha - \langle \hat{d}, (u(s_t))\rangle + \langle \hat{d}, r(s_t, a_t)\rangle] \geq 0,$$

or

$$\alpha \leq \langle \hat{d}, \frac{E_{\pi}[(u(s_{n+1}))-(u(s_1))]}{n} + \frac{\Phi^n(\pi)}{n} \rangle \qquad (4.10)$$

for all n. The equality in (4.10) holds when the policy f is chosen. Note that

$$\Phi^n(\pi) = \sum_{t=1}^{n} E_{\pi}[r(s_t, a_t)].$$

Here, the set of all cluster points of the set

$$\{ \frac{E_{\pi}[(u(s_{n+1}))-(u(s_1))}{n} + \frac{\Phi^n(\pi)}{n}, \ n=1,2,3,\cdots\cdots \}$$

is equal to $C(\pi)$ since, for any $u \in B(S)$,

$$\lim_{n \to \infty} \frac{E_{\pi}[(u(s_{n+1}))-(u(s_1))]}{n} = \theta.$$

Consequently, since $H_+ = \{ z \in R^m \mid \langle \hat{d}, z \rangle \geq \alpha \}$ is a closed positive half space with a support hyperplane $H = \{ z \in R^m \mid \langle$

$\hat{d}, z > = \alpha \}$ as its boundary, it follows from (4.10) that

$$H_+ \supset C(\pi) \qquad \text{for all } \pi \in \Pi. \tag{4.11}$$

Similary, since the equality in (4.10) holds for the stationary policy f, we have

$$H \supset C(f). \tag{4.12}$$

Therefore, (4.11) and (4.12) imply that, for each $\Phi(f) \in C(f)$,

$$< \hat{d}, \Phi(f) > \leq < \hat{d}, \Phi(\pi) > \qquad \text{for all } \Phi(\pi) \in C,$$

where $C = \bigcup_{\pi} C(\pi)$. This shows that f is an optimal policy for the average criterion of the system (4.1). Hence, applying Lemma 3.1, we see that f is a D-optimal policy of time average criterion of the vector-valued decision system (2.1).

Remark 4.1 In Ross [14-16], he treats

$$\limsup_{n \to \infty} \Phi^n(\pi)/n \quad \text{or} \quad \liminf_{n \to \infty} \Phi^n(\pi)/n$$

in the case of real-valued loss function. It can be considered as an element of $C(\pi)$ in one dimensional case. Hence, Theorem 4.1 with $D = [0, \infty) \subset R$ and $\hat{d} = 1 \in L_1^+ \subset R$ gives the results in the case of the real-valued loss function.

In order to give a converse version of Theorem 4.1 in the modified decision system, we introduce the cone convexity for a set E.

Definition 4.1 Let $D = L \cup \{\theta\}$ be a conex cone. A subset E

in R^m is said to be D-convex if $E + D$ is convex in R^m.

The following theorem is a partial converse of Theorem 4.1.

Theorem 4.2 Let π^* be a D-optimal policy of average criterion for the decision system (2.1), and let C (see Section 3) be a D-convex set. Suppose that there is a cluster point $\Phi(\pi^*)$ in $C(\pi^*)$ such that

$$cl[W] \cap cl[-D] = \{\theta\}, \tag{4.13}$$

where $W = C + D - \Phi(\pi^*)$ and $[W]$ is the cone generated by W. Then, π^* is an optimal policy for the average criterion of the modified decision system (4.1) with a numerical loss function $\langle \hat{d}, r \rangle$ on Π for some $\hat{d} \in L_1^+$.

Proof. From the expression (3.6) of $\Phi^n(\pi)/n$ for $\hat{d} \in L_1^+$, it is sufficient to show that there exists $\hat{d} \in L_1^+$ such that

$$\langle \hat{d}, \Phi(\pi^*) \rangle \leq \langle \hat{d}, \Phi(\pi) \rangle \quad \text{for all } \Phi(\pi) \in C. \tag{4.14}$$

Thus we wish to have a non-zero vector $\hat{d} \in L^+ \cap W^*$, where W^* is the positive polar cone of W. We will prove it by contradiction. Suppose to the contrary that $L^+ \cap W^* = \phi$. Since L^+ and W^* are convex and $\text{int}L^+ \neq \phi$, by the separation theorem for two disjoint convex sets (see for example, Bazarra & Shetty [4], or Luenberger [13]), we see that there exists a non-zero vector $\bar{d} \in (R^m)^* = R^m$ such that

$$\inf_{y \in W^*} \langle \bar{d}, y \rangle \geq \sup_{x \in L^+} \langle \bar{d}, x \rangle.$$

As $0 \in W^*$, $\langle \bar{d}, x \rangle \leq 0$ for all $x \in L^+$. We have

$$\bar{d} \in (-L^+)^* = \mathrm{cl}(-D). \tag{4.15}$$

Since, the continuous linear functional $\bar{d}$ acting on elements near zero in L^+ will take values near zero, it follows that

$$\sup_{x \in L^+} \langle \bar{d}, x \rangle \leq 0 \quad \text{and} \quad \inf_{y \in W^*} \langle \bar{d}, y \rangle \geq 0.$$

Hence,

$$\bar{d} \in W^{**} = \mathrm{cl}[W]. \tag{4.16}$$

From (4.15) and (4.16), we see that $\mathrm{cl}[W] \cap \mathrm{cl}(-D)$ contains a non-zero $\bar{d}$ which contradicts the assumption (4.13), and it contains only the zero vector. Therefore,

$$W^* \cap L^+ \neq \phi.$$

Let $\tilde{d}$ be a non-zero vector such that $\langle \tilde{d}, y \rangle \geq 0$ for all $y \in W$, that is,

$$\langle \tilde{d}, \Phi(\pi) + d - \Phi(\pi^*) \rangle \geq 0 \qquad \text{for all } \Phi(\pi) \in C.$$

Here d may be chosen zero in D. Hence,

$$\langle \tilde{d}, \Phi(\pi) \rangle \geq \langle \tilde{d}, \Phi(\pi^*) \rangle \quad \text{for all } \Phi(\pi) \in C. \tag{4.17}$$

Further, since $\tilde{d} \in L^+$ and $e = (1, 1, \cdots, 1) \in L$, there exists $\alpha > 0$ such that $\alpha = \langle \tilde{d}, e \rangle$. So, dividing both sides of (4.17) by the positive number α and letting $\hat{d} = \tilde{d}/\alpha$, we then have $\hat{d} \in L_1^+$, and hence

$$\langle \hat{d}, \Phi(\pi) \rangle \geq \langle \hat{d}, \Phi(\pi^*) \rangle \qquad \text{for all } \Phi(\pi) \in C.$$

This shows that π^* is an optimal policy of the average criterion of the modified decision system (4.1).

References

[1] Aubin,J.P.(1979), Mathematical Methods of Game and Economic Theory, North-Holland, Amsterdam.

[2] Blackwell,D.(1962), Discrete dynamic programming, Ann. Math. Statist., 33, 719-726.

[3] Blackwell,D.(1965), Discounted dynamic programming, Ann. Math. Statist., 36, 226-235.

[4] Bazarra,M.S.and Shetty,C.M.(1979), Nonlinear Programming, John Wiley & Sons, New York.

[5] Furukawa,N.(1980), Characterization of optimal policies in vector-valued Markovian decision processes, Math.Oper. Res., 5, 271-279.

[6] Hartley,R.(1978), On cone-efficiency, cone-convexity and cone-compactness, SIAM J. Appl. Math., 34, 211-222.

[7] Hartley,R.(1979), Finite, Discounted, Vector Markov Decision Processes, Note in Decision Theory 85, University of Manchester.

[8] Henig,M.(1983), Vector-valued dynamic programming, SIAM J. Control Optim., 21, 490-499.

[9] Henig,M.(1985), The principle of optimality in dynamic programming with returns in partially ordered sets, Math. Oper. Res., 10, 462-470.

[10] Himmelberg,C.J.,Parthasarathy,T., and Vleck,F.S.(1976), Optimal plans for dynamic programming problems, Math. Oper. Res., 1, 390-394.

[11] Hinderer,K.(1970), Foundations of Non-Stationary Dynamic

Programming with Discrete-Time Parameter, Springer-Verlag, Berlin.

[12] Lai,H.C., and Tanaka,K.(1986), On a D-solution of a cooperative m-person discounted Markov game, J. Math. Anal. Appl., 115, 578-591.

[13] Luenberger,D.G.(1969), Optimization by Vector Space Methods, Wiley-Interscience, New York.

[14] Ross,S.M.(1968), Non-discounted denumerable Markovian decision models, Ann. Math. Statist., 39, 412-423.

[15] Ross,S.M.(1968), Arbitrary state Markovian decision processes, Ann. Math. Statist., 39, 2118-2122.

[16] Ross,S.M.(1970), Applied Probability Models with Optimization Applications, Holden-Day, San Francisco.

[17] Strauch,R.(1966), Negative dynamic programming, Ann. Math. Statist., 37, 871-890.

[18] Tanino,T.and Sawaragi,Y.(1979), Duality theory in multiobjective programming, J. Optim. Theory Appl., 27, 509-529.

[19] Tijms,H.C.(1975), On dynamic programming with arbitrary state space, compact action space and the average return as criterion, Report BW 55/75, Math. Centre, Amsterdam.

[20] White,D.J.(1980), Multi-Objective, Infinite Horizon, Discounted Markov Decision Processes, Notes of Decision Theory 95, University of Manchester.

[21] Yu,P.L.(1974), Cone convexity, cone extreme point, and nondominated solutions in decision problems with multiobjectives, J. Optim. Theory Appl., 14, 319-377.

Hang-Chin Lai
Department of Mathematics,
University of Cape Town,
South Africa.

Kensuke Tanaka
Department of Mathematics,
Faculty of Science,
Niigata University,
Japan

Received January 20, 1991