Interval Methods for Uncertain Markov Decision Processes

M. Kurano*, Masami Yasuda† and J. Nakagami†

Abstract

In this paper, the average cases of Markov decision processes with uncertainty is considered. That is, a controlled Markov set-chain model with a finite state and action space is developed by an interval arithmetic analysis, and we will find a Pareto optimal policy which maximizes the average expected rewards over all stationary policies under a new partial order. The Pareto optimal policies is characterized by a maximal solution of an optimality equation which is derived from the model. Also, a maximin policy is obtained and shown to be Pareto-optimal. A numerical example is given.

Keywords: Controlled Markov set-chains, average reward criterion, Pareto optimal, interval arithmetic.

AMS 1991 subject classification. Primary: 90c40; Secondary: 90c39.

1 Introduction and notations

In a real application of Markov decision processes(MDP's in short, cf.[7, 10, 15]), the require data must be estimated. Thus, the mathematical model of MDPs can be only be viewed as approximations. It may be useful that the model is ameliorated so to be more "robust" in the sense that it's reasonably efficient in rough approximations. How can be modelized this situation? One of a realistic way to answer such a problem is to use certain intervals containing the required data.

Applying Hartfiel's[5, 6] interval method for Markov chains, Kurano et al[13] has introduced a decision model, called a controlled Markov set-chain, which is robust for rough approximation of the transition matrix in MDPs. The discounted reward problems was developed in [12, 13] and the non-discounted case had been discussed in [8]. However, the functional characterization of optimal policies for average reward problems is not given yet.

In this paper, applying an interval arithmetic analysis, the average case of MDPs with uncertainty is considered. That is, in a controlled Markov set-chain with finite state and action spaces, we will find a Pareto optimal policy which maximizes the average expected rewards over all stationary policies under some partial order.

*Dep of Mathematics, Faculty of Education, †Dep of Mathematics & Informatics, Faculty of Science, Chiba University, Yayoi-cho, Inage-ku, Chiba, 263-8255 Japan

Analyzing the behavior of the expected rewards over T-horizon as T approaches ∞ under some regularity condition, the average expected rewards corresponding to any stationary policy is given as a unique solution of an interval equation. Also, Pareto optimal policies are characterized by a maximal solution of an optimality equation determined by the decision model.

Concerning about a Pareto optimal policy, we will consider a maximin policy, which maximizes the expected average reward in the worst cases scenarios for the decision process. Takahashi[17, 18] has introduced a weak D-Markov chain in order to treat bounds for state probabilities of aggregated chains in the large scale Markov chains and applied to tandem queueing networks. The results are closely related to ours and we will clear it in the sequel.

In the remainder of this section, we shall give some notation referring to the works [4, 5, 6] on Markov set-chain and an interval arithmetic by [14] and formulate a controlled Markov set-chain which will be examined in the sequel.

Let R, R^n and $R^{n\times m}$ be the sets of real numbers, real n- dimensional column vectors and real $n \times m$ matrices, respectively. We shall identify $n \times 1$ matrices with vectors and 1×1 matrices with real numbers, so that $R = R^{1\times 1}$ and $R^n = R^{n\times 1}$. Also, we denote by R_+, R^n_+ and $R^{n\times m}_+$ the subsets of entrywise non-negative elements in R, R^n and $R^{n\times m}$, respectively.

We equip $R^{n\times m}$ with the component-wise relations $\leq, <, \geq, >$. For any $\underline{A} = (\underline{a}_{ij})$, $\overline{A} = (\overline{a}_{ij})$ in $R^{n\times m}_+$ with $\underline{A} \leq \overline{A}$, we define the set of stochastic matrices, $\langle \underline{A}, \overline{A}\rangle$, by

$$\langle \underline{A}, \overline{A}\rangle := \{A \mid A = (a_{ij}) \text{ isan } n \times m \text{ stochasticmatrixwith } \underline{A} \leq A \leq \overline{A}\},$$

Let

$$\mathcal{M}_n := \{\mathcal{A} = \langle \underline{A}, \overline{A}\rangle \mid \langle \underline{A}, \overline{A}\rangle \neq \emptyset, \underline{A} \leq \overline{A} \text{ and } \underline{A}, \overline{A} \in R^{n\times n}_+\}.$$

The product of $\mathcal{A}$ and $\mathcal{B} \in \mathcal{M}_n$ is defined by

$$\mathcal{AB} := \{AB \mid A \in \mathcal{A}, B \in \mathcal{B}\}.$$

For any sequence $\{\mathcal{A}_i\}^{\infty}_{i=1}$ with $\mathcal{A}_i \in \mathcal{M}_n$ $(i \geq 1)$, we define the multiproduct inductively by

$$\mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_k := (\mathcal{A}_1 \cdots \mathcal{A}_{k-1})\mathcal{A}_k \quad (k \geq 2).$$

Denote by $C(R_+)$ the set of all bounded and closed intervals in R_+. Let $C(R_+)^n$ be the set of all n-dimensional column vectors whose elements are in $C(R_+)$, i.e.,

$$C(R_+)^n := \{D = (D_1, D_2, \cdots, D_n)' \mid D_i \in C(R_+)(1 \leq i \leq n)\}.$$

where d' denotes the transpose of a vector d.

The following arithmetics are used in Section 2. For $D = (D_1, D_2, \cdots, D_n)', E = (E_1, E_2, \cdots, E_n)' \in C(R_+)^n, h \in R^n_+$ and $\lambda \in R_+$, $D + E = \{d + e \mid d \in D, e \in E\}, \lambda D = \{\lambda d \mid d \in D\}$ and $h + D = \{h + d \mid d \in D\}$. If $D = ([\underline{d}_1, \overline{d}_1], \cdots, [\underline{d}_n, \overline{d}_n])', D$ will be denoted by $D = [\,\underline{d}, \overline{d}\,]$, where $\underline{d} = (\underline{d}_1, \cdots, \underline{d}_n)', \quad \overline{d} = (\overline{d}_1, \cdots, \overline{d}_n)'$ and $[\,\underline{d}, \overline{d}\,] = \{d \in R^n_+ \mid \underline{d} \leq d \leq \overline{d}\}$.

For any $D = (D_1, D_2, \cdots, D_n)' \in C(R_+)^n$ and subset G of $R^{1\times n}_+$ the product of G and D is defined as

$$GD := \{gd \mid g = (g_1, \cdots, g_n) \in G, d = (d_1, \cdots, d_n)' \in D, d_i \in D_i(1 \leq i \leq n)\}.$$

The following results are used in the sequel.

Lemma 1.1 ([6, 13])

(i) Any $\mathcal{A} \in \mathcal{M}_n$ is a convex polytope in the vector space $R^{n\times n}$.

(ii) For any compact convex subset $G \subset R_+^{1\times n}$ and $D = (D_1, D_2, \cdots, D_n)' \in C(R_+)^n$, it holds $GD \in C(R_+)$.

We will give a partial order $\succ, \succeq$ on $C(R_+)$ by the definition : For $[c_1, c_2], [d_1, d_2] \in C(R_+)$,

$$[c_1, c_2] \succeq [d_1, d_2] \quad \text{if} \quad c_1 \geq d_1, c_2 \geq d_2,$$

and

$$[c_1, c_2] \succ [d_1, d_2] \quad \text{if} \quad [c_1, c_2] \succeq [d_1, d_2] \quad \text{and} \quad [c_1, c_2] \neq [d_1, d_2].$$

For $\boldsymbol{v} = (v_1, v_2, \cdots, v_n)'$ and $\boldsymbol{w} = (w_1, w_2, \cdots, w_n)' \in C(R_+)^n$, we write $\boldsymbol{v} \succeq \boldsymbol{w}$ (by abase of notation) if $v_i \succeq w_i, \quad 1 \leq i \leq n$ and $\boldsymbol{v} \succ \boldsymbol{w}$ if $\boldsymbol{v} \succeq \boldsymbol{w}$ and $\boldsymbol{v} \neq \boldsymbol{w}$. Define a metric Δ on $C(R_+)^n$ by

$$\Delta(\boldsymbol{v}, \boldsymbol{w}) := \max_{i\in S} \delta(v_i, w_i)$$

for $\boldsymbol{v} = (v_1, v_2, \cdots, v_n)', \boldsymbol{w} = (w_1, w_2, \cdots, w_n)' \in C(R_+)^n$, where δ is the Hausdorff metric on $C(R_+)$ and given by

$$\delta([a, b], [c, d]) := |a - c| \vee |b - d| \quad \text{for} \quad [a, b], [c, d] \in C(R_+),$$

where $x \vee y = \max\{x, y\}$. Obviously, $(C(R_+)^n, \Delta)$ is a complete metric space.

A controlled Markov set-chain consists of four object:

$$(S, A, \underline{q}, \overline{q}, r),$$

where $S = \{1, 2, \cdots, n\}$ and $A = \{1, 2, \cdots, k\}$ are finite sets and for each $(i, a) \in S \times A, \underline{q} = \underline{q}(\cdot|i, a) \in R_+^{1\times n}$, $\overline{q} = \overline{q}(\cdot|i, a) \in R_+^{1\times n}$ with $\underline{q} \leq \overline{q}$ and $\langle \underline{q}, \overline{q} \rangle \neq \emptyset$ and $r = r(i, a)$ a function on $S \times A$ with $r \geq 0$. Note that A is used as the set in this section, different from that in the previous section. We interpret S as the set of states of some system, and A as the set of actions available at each state.

When the system is in state $i \in S$ and we take action $a \in A$, we move to a new state $j \in S$ selected according to the probability distribution on S, $q(\cdot|i, a)$, and we receive an immediate return, $r(i, a)$, where we know only that $q(\cdot|i, a)$ is arbitrarily chosen from $\langle \underline{q}(\cdot|i, a), \overline{q}(\cdot|i, a) \rangle$. This process is then repeated from the new state j. Denote by F the set of functions from S to A.

A policy π is a sequence $(f_1, f_2, \cdots)$ of functions with $f_t \in F, (t \geq 1)$. Let Π denote the class of policies. We denote by f^∞ the policy $(h_1, h_2, \cdots)$ with $h_t = f$ for all $t \geq 1$ and some $f \in F$. Such a policy is called stationary, denoted simply by f, and the set of stationary policies is denoted by Π_F.

We associate with each $f \in F$ the n-dimensional column vector $r(f) \in R_+^n$ whose i-th element is $r(i, f(i))$ and the set of stochastic matrices $\mathcal{Q}(f) := \langle \underline{Q}(f), \overline{Q}(f) \rangle \in \mathcal{M}_n$, where the (i, j) elements of $\underline{Q}(f)$ and $\overline{Q}(f)$ are $\underline{q}(j|i, f(i))$ and $\overline{q}(j|i, f(i))$, respectively, and $\langle \underline{Q}(f), \overline{Q}(f) \rangle$ is already defined.

For any $\pi = (f_1, f_2, \cdots) \in \Pi$, let $\boldsymbol{v}_1(\pi) = r(f_1)$ and, by setting $Q_0 = \text{identity}$,

$$\boldsymbol{v}_T(\pi) = \{\Sigma_{i=1}^T Q_1 Q_2 \cdots Q_{t-1} r(f_t) \mid Q_i \in \mathcal{Q}(f_i), i = 1, 2, \cdots, T - 1\} \quad (T \geq 2). \tag{1.1}$$

We observe, for example, that

$$\boldsymbol{v}_3(\pi) = r(f_1) + \mathcal{Q}(f_1)(r(f_2) + \mathcal{Q}(f_2)r(f_3)),$$

so that by Lemma 1.1 (ii) it holds that $\boldsymbol{v}_T(\pi) \in C(R_+)^n$ for all $T \geq 1$. For any $\pi \in \Pi$, let

$$\boldsymbol{v}(\pi) := \liminf_{T\to\infty} \frac{1}{T}\boldsymbol{v}_T(\pi), \tag{1.2}$$

where, for a sequence $\{D_k\} \subset C(R_+)^n$,

$$\liminf_{k\to\infty} D_k := \left\{ x \in R^n \,\middle|\, \limsup_{k\to\infty} \delta_1(x, D_k) = 0 \right\},$$

and $\delta_1(x, D) := \inf_{y\in D} \delta_2(x, y)$, δ_2 is a metric in R^n. Since $\boldsymbol{v}(\pi) \in C(R_+)^n$, $\boldsymbol{v}(\pi)$ is written as $\boldsymbol{v}(\pi) = [\underline{v}(\pi), \overline{v}(\pi)]$.

Definition. A policy $f^* \in \Pi_F$ is called *Pareto optimal* if there does not exist $f \in \Pi_F$ such that $\boldsymbol{v}(f^*) \prec \boldsymbol{v}(f)$.

In the above definition, we confine ourselves to the stationary policies, which simplifies our discussion in the sequel. In Section 2, a regularity condition for the class of transition matrices is introduced, under which the interval equations concerning the average rewards are investigated. In Section 3, the asymptotic behavior of $\boldsymbol{v}_T(f)$ as T approaches ∞ is given. And in Section 4, Pareto optimal policies are characterized by maximal solutions of optimality equation. Also, a maximin policy which maximizes the expected reward earned in the worst cases scenarios is given and shown to be Pareto optimal.

2 Assumption and preliminary lemmas

Henceforth, the following assumption will remain operative.

Assumption A (Primitivity). For any $f \in F$, each $Q \in \mathcal{Q}(f)$ is primitive, i.e., $Q^t > 0$ for some $t \geq 1$.

Obviously, if $\underline{Q}(f)$ is primitive in the sense of non-negative matrix (cf.[16]), Assumption A holds.

The following facts on Markov matrices are well-known (cf.[2, 11]).

Lemma 2.1 For any $f \in F$, let Q be any matrix in $\mathcal{Q}(f)$.

(i) The sequence $(I + Q + \cdots + Q^t)/(t+1)$ converges as $t \to \infty$ to a stochastic matrix Q^* with $Q^*Q = Q^*$, $\quad Q^* > 0$ and $\mathrm{rank}(Q^*) = 1$.

(ii) The matrix Q^* in (i) is uniquely determined by $Q^*Q = Q$ and $\mathrm{rank}(Q^*) = 1$.

Associated with each $f \in F$ is a corresponding operator $L(f)$, mapping $C(R_+)^n$ into $C(R_+)^n$, defined as follows.

For $\boldsymbol{v} \in C(R_+)^n$,

$$L(f)\boldsymbol{v} := r(f) + \mathcal{Q}(f)\boldsymbol{v}. \tag{2.1}$$

Note that from Lemma 1.1, $L(f)\boldsymbol{v} \in C(R_+)^n$ for each $\boldsymbol{v} \in C(R_+)^n$. Putting $\boldsymbol{v} = [\underline{v}, \overline{v}]$ with $\underline{v} \leq \overline{v}$, $\underline{v}, \overline{v} \in R_+^n$, (2.1) can be written as

$$L(f)\boldsymbol{v} = [\underline{L}(f)\underline{v}, \overline{L}(f)\overline{v}], \tag{2.2}$$

where $\underline{L}$ and $\overline{L}$ are operators from R^n into itself, defined by :

$$\begin{cases} \underline{L}(f)v = r(f) + \min_{Q\in\mathcal{Q}(f)} Qv, \\ \overline{L}(f)v = r(f) + \max_{Q\in\mathcal{Q}(f)} Qv. \end{cases} \tag{2.3}$$

and min(max) represents component-wise minimization(maximizing).

Let $\mathbf{e} := (1, 1, \cdots, 1)'$. Here, for any $f \in F$, we consider the interval equation:

$$r(f) + \mathcal{Q}(f)\boldsymbol{h} = \boldsymbol{v} + \boldsymbol{h}, \tag{2.4}$$

where $\boldsymbol{v} := [\underline{v}\mathbf{e}, \overline{v}\mathbf{e}]$, $\boldsymbol{h} = [\underline{h}, \overline{h}] \in C(R)^n$, $\underline{v}, \overline{v} \in R, \underline{h}, \overline{h} \in R^n$ with $\underline{v} \leq \overline{v}, \underline{h} \leq \overline{h}$.

Obviously, the interval equation can be rewritten by their extremal points as

$$\begin{cases} r(f) + \min_{Q\in\mathcal{Q}(f)} Q\underline{h} = \underline{v}\mathbf{e} + \underline{h} \\ r(f) + \max_{Q\in\mathcal{Q}(f)} Q\overline{h} = \overline{v}\mathbf{e} + \overline{h} \end{cases} \tag{2.5}$$

with

$$\underline{v} \leq \overline{v}, \quad \underline{h} \leq \overline{h} \tag{2.6}$$

where $\underline{v}, \overline{v} \in R, \underline{h}, \overline{h} \in R^n$.

Takahashi[17, 18] has showed that upper or lower bounds of the average rewards for weak D-Markov chains satisfies both of the equations (2.5), and had calculated them with Howard's policy improvement ([10]).

We have the following lemma.

Lemma 2.2([1, 8]) For any $f \in F$, the interval equation (2.5) determines $\boldsymbol{v}$ uniquely and $\boldsymbol{h}$ up to an additive constant $[c_1\mathbf{e}, c_2\mathbf{e}]$ with $c_1, c_2 \in R(c_1 < c_2)$.

3 Asymptotic bounds for the finite horizon reward

In this section we give the asymptotic behavior of $\boldsymbol{v}_T(f)$ as $T \to \infty$ under Assumption A, whose proofs are given in [9]. Throughout this section, we assume that Assumption A holds.

For any $\boldsymbol{g} \in C(R)^n$ and $f \in F$, the sequence $\{\boldsymbol{v}_T(f, \boldsymbol{g}), T \geq 0\}$ is defined as follows:

$$\boldsymbol{v}_0(f, \boldsymbol{g}) := \boldsymbol{g}$$

and

$$\boldsymbol{v}_T(f, \boldsymbol{g}) := \left\{\Sigma_{i=1}^T Q_1 \cdots Q_{t-1} r(f) + Q_1 \cdots Q_T \boldsymbol{g} \mid Q_i \in \mathcal{Q}(f), i = 1, \cdots, T\right\} \quad (t \geq 1). \tag{3.1}$$

Lemma 3.1 For any $\boldsymbol{g} \in C(R_+)^n$ and $f \in F$, the sequence $\{\boldsymbol{v}_T(f, \boldsymbol{g})\}$ satisfies that

$$\boldsymbol{v}_T(f, \boldsymbol{g}) = L(f)\boldsymbol{v}_{T-1}(f, \boldsymbol{g}) \quad (T \geq 1). \tag{3.2}$$

Since the solutions $\underline{v}$, $\overline{v}$, $\boldsymbol{h}$ of (2.4)-(2.6) in Section 2 are depending on $f \in F$, we will denote them respectively by $\boldsymbol{v}(f) = [\underline{v}(f)\mathbf{e}, \overline{v}(f)\mathbf{e}]$ and $\boldsymbol{h}(f) = [\underline{h}(f), \overline{h}(f)]$.

Lemma 3.2 For any $f \in F$, it holds that

$$\boldsymbol{v}_T(f, \boldsymbol{h}(f)) = T\boldsymbol{v}(f) + \boldsymbol{h}(f) \quad \text{for all } \ T \geq 0. \tag{3.3}$$

The following theorem is concerned with the asymptotic properties of $\boldsymbol{v}_T(f)$ as $T \to \infty$.

Theorem 3.1 For any $f \in F$, there exists $c_1, c_2, c_1', c_2' \in R$ $(c_1' \leq c_1, c_2' \leq c_2)$ such that

$$[(T\underline{v}(f) + c_1)\mathbf{e}, (T\overline{v}(f) + c_2')\mathbf{e}] \subset \boldsymbol{v}_T(f) \subset [(T\underline{v}(f) + c_1')\mathbf{e}, (T\overline{v}(f) + c_2)\mathbf{e}]$$
$$\text{for all } \ T \geq 0, \tag{3.4}$$

where $[a, b] = \emptyset$, if $a > b$.

For simplicity of the notation, let, for any $d \in R^n$ and $f \in F$,

$$\underline{\mathcal{Q}}(f, d) := \left\{ Q \in \mathcal{Q}(f) \,\middle|\, Qd = \min_{Q \in \mathcal{Q}(f)} Qd \right\},$$

and

$$\overline{\mathcal{Q}}(f, d) := \left\{ Q \in \mathcal{Q}(f) \,\middle|\, Qd = \max_{Q \in \mathcal{Q}(f)} Qd \right\}.$$

Corollary 3.1 For any $f \in F$, it holds

(i) $$\boldsymbol{v}(f) = [\underline{v}(f)\mathbf{e}, \overline{v}(f)\mathbf{e}]$$

and

(ii) $$\underline{v}(f)\mathbf{e} = \underline{Q}^* r(f), \quad \overline{v}(f)\mathbf{e} = \overline{Q}^* r(f)$$

for any $\underline{Q} \in \underline{\mathcal{Q}}(f, \underline{h})$ and $\overline{Q} \in \overline{\mathcal{Q}}(f, \overline{h})$, where $Q^* = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} Q^t$ for $Q \in \mathcal{Q}(f)$.

4 Characterization of Pareto optimal policies

In this section, we derive the optimality equation by which all Pareto optimal policies are characterized.

Lemma 4.1 Let f be Pareto optimal and $\boldsymbol{v}(f) = [\underline{v}(f)\mathbf{e}, \overline{v}(f)\mathbf{e}]$. Then, it holds

(i) $$\underline{v}(f)\mathbf{e} = \min_{Q \in \mathcal{Q}(f)} Q^* r(f),$$

and

(ii) $$\overline{v}(f)\mathbf{e} = \max_{Q \in \mathcal{Q}(f)} Q^* r(f).$$

Proof. Let $\overline{v}(f), \underline{v}(f) \in R$ and $\overline{h}(f), \underline{h}(f) \in R^n$ be a solution of (2.5). Then, for any $Q \in \mathcal{Q}(f)$, (2.5) implies

$$r(f) + Q\underline{h}(f) \succeq \underline{v}(f)\mathbf{e} + \underline{h}(f) \tag{4.1}$$

By $Q^*Q = Q$, (4.1) derives that $\underline{v}(f)\mathbf{e} \preceq Q^*r(f)$ This fact implies (i) by Corollary 3.1(ii). Similarly, (ii) can be proved. Q.E.D.

Lemma 4.2 For any f, g in F, suppose that

$$\boldsymbol{v}(f) + \boldsymbol{h}(f) \left\{ \begin{array}{c} \preceq \\ \prec \end{array} \right\} r(g) + \mathcal{Q}(g)\boldsymbol{h}(f). \tag{4.2}$$

Then, it holds that

$$\boldsymbol{v}(f) \left\{ \begin{array}{c} \preceq \\ \prec \end{array} \right\} \boldsymbol{v}(g). \tag{4.3}$$

Proof. The left and right extremal equation of (4.2) are given as follows.

$$\underline{v}(f)\mathbf{e} + \underline{h}(f) \left\{ \begin{array}{c} \preceq \\ \prec \end{array} \right\} r(g) + \min_{Q \in \mathcal{Q}(f)} Q\underline{h}(f) \tag{4.4}$$

$$\overline{v}(f)\mathbf{e} + \overline{h}(f) \left\{ \begin{array}{c} \preceq \\ \prec \end{array} \right\} r(g) + \max_{Q \in \mathcal{Q}(f)} Q\overline{h}(f). \tag{4.5}$$

By Lemma 4.1, there exists $\underline{Q} \in \mathcal{Q}(g)$ with $\underline{v}(g)\mathbf{e} = \underline{Q}^*r(g)$. Multiplying the both sides of (4.4) by $\underline{Q}^*$, we get that from $\underline{Q}^*\underline{Q} = \underline{Q}^*$ and $\underline{Q}^* > 0$ that $\underline{v}(f)\mathbf{e} \left\{ \begin{array}{c} \preceq \\ \prec \end{array} \right\} \underline{Q}^*r(g)$. Thus $\underline{v}(f)\mathbf{e} \left\{ \begin{array}{c} \preceq \\ \prec \end{array} \right\} \underline{v}(g)\mathbf{e}$ follows. Similarly, we get $\overline{v}(f)\mathbf{e} \left\{ \begin{array}{c} \preceq \\ \prec \end{array} \right\} \overline{v}(g)\mathbf{e}$, which proves (4.3). Q.E.D.

Let $\mathcal{U}$ be an arbitrary subset of $\mathcal{C}(R)^n$. A point $\boldsymbol{u} \in \mathcal{U}$ is called an efficient element of $\mathcal{U}$ with respect to $\preceq$ on $\mathcal{C}(R)^n$ iff it holds that there does not exist $\boldsymbol{v} \in \mathcal{U}$ such that $\boldsymbol{u} \prec \boldsymbol{v}$. We denote by $\mathrm{eff}(\mathcal{U})$ the set of all elements of $\mathcal{U}$ efficient with respect to $\preceq$ on $\mathcal{C}(R)^n$. For any $\boldsymbol{u} \in \mathcal{C}(R)^n$, let

$$\mathcal{L}(\boldsymbol{u}) := \mathrm{eff}\left(\{L(f)\boldsymbol{u} \mid f \in F\}\right),$$

where $L(f)\boldsymbol{u} \in \mathcal{C}(R)^n$ is defined in (2.1). We note that $\mathcal{L}(\boldsymbol{u}) \subset \mathcal{C}(R)^n$ for any $\boldsymbol{u} \in \mathcal{C}(R)^n$.

Here, we consider the following interval equations inducing efficient set-function $\mathcal{L}(\cdot)$ on $\mathcal{C}(R)^n$.

$$\boldsymbol{v} + \boldsymbol{h} \in \mathcal{L}(\boldsymbol{h}), \tag{4.6}$$

where $\boldsymbol{v} = [\underline{v}\mathbf{e}, \overline{v}\mathbf{e}], \boldsymbol{h} = [\underline{h}, \overline{h}] \in \mathcal{C}(R)^n$ and $\underline{v} \leq \overline{v}, \underline{h} \leq \overline{h}, \underline{v}, \overline{v} \in R, \underline{h}, \overline{h} \in R^n$. The equation (4.6) is called an optimality equation, by which Pareto optimal policies are characterized. A solution $(\boldsymbol{v}, \boldsymbol{h})$ of the optimal equation (4.6) is called maximal if there does not exist any solution $(\boldsymbol{v}', \boldsymbol{h}')$ of (4.6) such that $\boldsymbol{v} \prec \boldsymbol{v}'$. In the following, Pareto optimal policies are characterized by maximal solutions of the optimality equation.

Theorem 4.1 A policy f^∞ is Pareto optimal if and only if the pair $(\boldsymbol{v}(f), \boldsymbol{h}(f))$ given by Lemma 2.2 is a maximal solution to the optimality equation (4.6).

Proof. The proof of "only if " part is easily obtained from Lemma 4.2. In order to prove "if" part, suppose that $(\boldsymbol{v}(f), \boldsymbol{h}(f))$ is a maximal solution of (4.6) but f^∞ is

not Pareto optimal. Then, there exists $g \in F$ with $\boldsymbol{v}(f) \prec \boldsymbol{v}(g)$. Since $\{\boldsymbol{v}(g')|\boldsymbol{v}(g) \preceq \boldsymbol{v}(g'), g' \in F\}$ is finite, it has a maximal element $\boldsymbol{v}(g^{(1)})$ with respect to the partial order $\preceq$. For this $g^{(1)}$, suppose that $\boldsymbol{v}(g^{(1)}) + \boldsymbol{h}(g^{(1)}) \notin \mathcal{L}(\boldsymbol{h}(g^{(1)}))$. Then, there exists an $i_0 \in S$ and $a_0 \in A$ such that

$$\boldsymbol{v}(g^{(1)})_{i_0} + \boldsymbol{h}(g^{(1)})_{i_0} \prec r(i_0, a_0) + \left(q(i_0, a_0)\boldsymbol{h}(g^{(1)})\right)_{i_0}, \tag{4.7}$$

where $q(\cdot|i_0, a_0) := \langle \underline{q}(\cdot|i_0, a_0), \overline{q}(i_0, a_0)\rangle$. Define $f^{(1)}$ by

$$f^{(1)} = \begin{cases} a_0, & \text{if} \quad i = i_0, \\ g^{(1)}(i), & \text{if} \quad i \neq i_0. \end{cases}$$

Then, from (4.6), it holds that

$$\boldsymbol{v}(g^{(1)}) + \boldsymbol{h}(g^{(1)}) \prec r(f^{(1)}) + Q(f^{(1)})\boldsymbol{h}(g^{(1)}). \tag{4.8}$$

Thus, by Lemma 4.2, $\boldsymbol{v}(g^{(1)}) \prec \boldsymbol{v}(f^{(1)})$. For this $f^{(1)}$, obviously $\boldsymbol{v}(f) \prec \boldsymbol{v}(f^{(1)})$. If $\boldsymbol{v}(f^{(1)}) + \boldsymbol{h}(f^{(1)}) \notin \mathcal{L}(\boldsymbol{h}(f^{(1)}))$, we can construct $f^{(2)}$ with $\boldsymbol{v}(f^{(1)}) \prec \boldsymbol{v}(f^{(2)})$, repeating the above discussion. Since F is a finite set, by repeating this method successively, we come to the conclusion that there exists $f^{(k)} \in F$ such that $\boldsymbol{v}(f) \prec \boldsymbol{v}(f^{(k)})$ and $(\boldsymbol{v}(f^{(k)}), \boldsymbol{h}(f^{(k)}))$ satisfies (4.6). However, this contradicts that $(\boldsymbol{v}(f), \boldsymbol{h}(f))$ is maximal. Q.E.D.

Remark. For vector-valued discounted MDPs, Furukawa[3] and White[19] had derived the optimal equation including efficient set-function on R^n, by which optimal policies are characterized. The form of the optimality equation (4.6) is corresponding to the average case of controlled Markov set-chains.

5 The maximim policy and an numerical example

In this section, a maximin policy which maximizes the expected average reward earned in the worst case scenarios of the decision process is given and shown to be Pareto optimal. Also a numerical example is given.

Let $\mathbf{q}(i, a) := \langle \underline{q}(\cdot|i, a), \overline{q}(\cdot|i, a)\rangle$ for each $i \in S$ and $a \in A$. For each $i \in S$ and $f \in F$, denote by $\underline{G}(i, f)$ the set of $a \in A$ for which

$$\underline{v}(f) + \underline{h}(f)_i < r(i, a) + \min_{q \in \mathbf{q}(i,a)} \sum_{j=1}^{n} q(j|i, a)\underline{h}(f)_j,$$

where $\underline{v}(f)$ and $\underline{h}(f) = (\underline{h}(f)_1, \cdots, \underline{h}(f)_n)$ is solution of (2.5). Let $g \in F$ be such that $g(i) \in \underline{G}(i, f)$ for any i with $\underline{G}(i, f) \neq \emptyset$ and $g(i) = f(i)$ for any i with $\underline{G}(i, f) = \emptyset$. Then, we have the following.

Lemma 5.1 For any f with $\underline{G}(i, f) \neq \emptyset$ for some $i \in S$, $\underline{v}(f) \prec \underline{v}(g)$.

The following lemma is proved from the idea of policy improvement (cf.[10]).

Lemma 5.2 The left side optimality equations (5.1) below determine $\underline{v}^*$ uniquely and $\underline{h} \in R^n$ up to an additive constant.

$$\underline{v}^* + \underline{h}_i = \max_{a \in A} \left(r(i, a) + \min_{q \in \mathbf{q}(i,a)} \sum_{j=1}^{n} q(j|i, a)\underline{h}_j \right) \quad (1 \leq i \leq n). \tag{5.1}$$

Let, for each i $(1 \leq i \leq n)$,

$$A_i := arg \max_{a \in A} \left(r(i,a) + \min_{q \in \mathbf{q}(i,a)} \sum_{j=1}^{n} q(j|i,a) \underline{h}_j \right).$$

For each $i \in S$ and $f \in F$ with $f(i) \in A_i$ for all $i \in S$, denote by $\overline{G}(i,f)$ the set of $a \in A_i$ for which

$$\overline{v}(f) + \overline{h}(f)_i < r(i,a) + \max_{q \in \mathbf{q}(i,a)} \sum_{j=1}^{n} q(j|i,a) \overline{h}(f)_j,$$

where $\overline{v}(f)$ and $\overline{h}(f) = (\overline{h}(f)_1, \cdots, \overline{h}(f)_n)$ is a solution of (2.5). Using $\overline{G}(i,f)$, we can prove the following.

Lemma 5.3 The right side optimality equations below determine $\overline{v}^*$ uniquely and $\overline{h} \in R^n$ up to an additive constant.

$$\overline{v}^* + \overline{h}_i = \max_{a \in A_i} \left(r(i,a) + \max_{q \in \mathbf{q}(i,a)} \sum_{j=1}^{n} q(j|i,a) \overline{h}_j \right) \quad (1 \leq i \leq n). \tag{5.2}$$

Let, for each $i (1 \leq i \leq n)$,

$$A_i^* := arg \max_{a \in A_i} \left(r(i,a) + \max_{q \in \mathbf{q}(i,a)} \sum_{j=1}^{n} q(j|i,a) \overline{h}_j \right).$$

A policy f^* with $f^*(i) \in A_i^*$ for all i in S is called a maximin policy. Then, $(\boldsymbol{v}(f^*), \boldsymbol{h}(f^*))$ corresponding to a maximin policy is obviously a maximal solution of the optimality equation (4.6). So, from Theorem 4.1, we have the following Theorem.

Theorem 5.1 A maximin policy f^* is Pareto optimal and $\boldsymbol{v}(f^*) = [\underline{v}^* \mathbf{e}, \overline{v}^* \mathbf{e}]$.

Here we shall give a numerical example which illustrates Theorem 5.1. For simplicity, let $\underline{q}_{ij}^a := \underline{q}(j|i,a)$, $\overline{q}_{ij}^a := \overline{q}(j|i,a)$ and $r(a) := (r(1,a), r(2,a))$. Consider the following controlled Markov set-chain model: $S = \{1,2\}, A = \{1,2\}$,

$$(\underline{q}_{ij}^1) = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix}, \quad (\overline{q}_{ij}^1) = \begin{pmatrix} \frac{2}{3} & \frac{1}{2} \\ \frac{1}{2} & \frac{2}{3} \end{pmatrix},$$

$$(\underline{q}_{ij}^2) = \begin{pmatrix} \frac{2}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{2}{5} \end{pmatrix}, \quad (\overline{q}_{ij}^2) = \begin{pmatrix} \frac{1}{2} & \frac{3}{5} \\ \frac{3}{5} & \frac{1}{2} \end{pmatrix},$$

and $r(1) = (1,2), \quad r(2) = (1, 2.1)$. Then, the equation (5.1) which $\underline{v}^*$ and $\underline{h} = (\underline{h}_1, \underline{h}_2)'$ is given as follows :

$$\underline{v}^* + \underline{h}_1 = \max \begin{cases} 1 + \min\{(2\underline{h}_1 + \underline{h}_2)/3, (\underline{h}_1 + \underline{h}_2)/2, \\ 1 + \min\{(2\underline{h}_1 + 3\underline{h}_2)/5, (\underline{h}_1 + \underline{h}_2)/2 \end{cases}$$

and

$$\underline{v}^* + \underline{h}_2 = \max \begin{cases} 2 + \min\{(\underline{h}_1 + 2\underline{h}_2)/3, (\underline{h}_1 + \underline{h}_2)/2, \\ 2.1 + \min\{(3\underline{h}_1 + 2\underline{h}_2)/5, (\underline{h}_1 + \underline{h}_2)/2. \end{cases}$$

After a simple calculation, the solution of the above with $\underline{h}_1 = 0$ becomes that $\underline{v}^* = 1.5$ and $\underline{h} = (0, 1)'$. Also, we easily find $A_1 = \{2\}$ and $A_2 = \{1, 2\}$. Similarly, by solving the equation $\overline{h}_1 = 0$, we get $\overline{v}^* = 23/14, \overline{h} = (0, 15/14)', A_1^* = \{2\}$ and $A_2^* = \{1\}$. So, by Theorem 5.1, f^* with $f^*(1) = 2$ and $f^*(2) = 1$ is Pareto optimal and $\boldsymbol{v}(f^*) = [(3/2)\mathbf{e}, (23/14)\mathbf{e}]$.

References

[1] Bather, J.; (1973) Optimal decision procedures for finite Markov chains, Part II : Communicating systems, *Adv. Appl. Prob.*, **5**, pp.521-540.

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

[3] Furukawa,N.; (1980) Characterization of Optimal Policies in Vector-valued Markovian Decision Process, Math.Oper.Res. vol.5, pp.271-279.

[4] Hartfiel, D. J.; (1993) Cyclic Markov set-chain. *J.Stat.Comp.Simul.*, **46**, pp.145-167.

[5] Hartfiel, D. J. and Seneta, E.; (1994) On the theory of Markov Set-chains, *Adv. Appl. Prob.* 26, pp.947-964.

[6] Hartfiel, D. J.; (1998) Markov Set-chains, Springer-Verlag, Berlin.

[7] Hinderer, K.; (1970) *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter.* Springer-Verlag, New York.

[8] Hosaka, M. and Kurano, M.; Non-discounted Optimal policy in controlled Markov Set-chains, *Submitted to Journal of Opern. Res. of Japan.*

[9] Hosaka, M., Horiguchi, M. and Kurano, M.; Controlled Markov Set-chains under Average Criteria, *Submitted to the 7th Bellman Continuum.*

[10] Howard, R.; (1960) *Dynamic Programming and Markov processes,* MIT Press, Cambridge MA.

[11] Kemeny, J. G. and Snell, J. L.; (1960) *Finite Markov-Chains,* Van Nostrand, New York.

[12] Kurano, M., Nakagami, J. and Horiguchi, M.; (1998) Controlled Markov Set-Chains with Set-valued rewards, *To appear in the Proceeding of International Conference on Nonlinear Analysis and Convex Analysis(NACA98).*

[13] Kurano, M. , Song, J. , Hosaka, M. and Huang, Y.;(1998) Controlled Markov Set-Chains with Discounting, *J.Appl.Prob.*, pp.293-302.

[14] Nenmaier, A.; (1984) New techniques for the analyses of linear interval equations, *Linear Algebra Appli.* 58, pp.273-325.

[15] Puterman, M. L.; (1994) *Markov decision processes: Discrete Stochastic Dynamic Programming,* John Wiley & Sons, INC.

[16] Seneta, E. (1981) *Nonnegative Matrices and Markov Chains,* Springer-Verlag, New York.

[17] Takahashi,Y.; (1984) Weak D-markov Chain and Its Application to a Queueing Network, in G.Iazeolla, P.J.Courtois, and A.Hordijk eds. "Mathematical Computer Performance and Reliability", North-Holland, Amsterdam, pp.153-165.

[18] Takahashi,Y.; (1988) A Weak D-markov Chain approach to Tandem Queueing Network, in G.Iazeolla, P.J.Courtois, and A.Hordijk eds. "Mathematical Computer Performance and Reliability", North-Holland, Amsterdam, pp.151-159.

[19] White,D.J.; (1982) Multi-objective infinite-horizon discounted Markov Decision Processes, J.Math.Anal.Appl., vol.89, pp.639-647.