数理解析研究所講究録
358巻 1979年 125-133

Partially Observable Markov Decision Problems with Vector-valued Criteria

Akira Ichikawa
Faculty of Engineering
Shizuoka University

1. Introduction. Optimal control of discrete time Markov processes with partial observation has been studied by many authors, for example, [1], [5], [6], [7]. Smallwood and Sondik [6] in particular considered a Markov chain with finite states, signals and actions. They have formulated an optimal control problem over a finite horizon and presented an algorithm for an optimal policy and the minimum cost. Sondik [7] has then developed a further study on the infinite horizon problem with discounting. He introduces a new concept of finite transient policies and proposes an algorithm. We [3], [4] have studied the same problem from a different angle and examined the relation between these two methods.

Recently the theory of Markov decision problems has been extended to the case of vector-valued criteria. Furukawa [2] has studied vector-valued Markov decision problems with countable states and established a policy improvement algorithm as well as the characterization of optimal policies. In this paper we take the model in [3], [4] and establish main results in [2] for our Markov process.

2. The model. Let $T = \{o,1,2,\cdots\}$, $Y = \{1,2,\cdots N\}$, $S = \{1,2,\cdots M\}$ and $U = \{1,2,\cdots K\}$ be the index set, the state space, the signal space and the control space respectively. Our basic stochastic process is a Markov chain $y_t \in Y$, $t \in T$ which is not directly observable. The system dynamics is described as follows. At time $t \in T$ we know that y_t has a probability distribution $x_t = x = (x_i) \in R^N$ (row vector), i.e., $x_i = P_r\{y_t=i\}$, $i=i,2,\cdots N$. If we choose a control $u_t = u$

then the process makes a transition according to the transition matrix $P^u = (p^u_{ij}) \in R^{N\times N}$ ($N\times N$-matrix). From the new state y_{t+1} we receive a signal $s_t \in S$. We assume that the conditional probability of observing s, given that the current state is i and the control u is selected, is r^u_{is}. Let $R^u_s = \text{diag}\,\{r^u_{is}\} \in R^{N\times N}$ (a diagonal matrix) and $e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in R^N$. Then the probability of observing s, given that the current probability distribution is x and the control u is selected, is given by $\{s|x,u\} = xP^uR^u_s$. By the Bayes' rule the distribution of x_{t+1} of y_{t+1} is then [6]

$$x_{t+1} = T(x|s,u) \triangleq \frac{xP^uR^u_s}{\{s|x,u\}} \tag{2.1}$$

The process is repeated with the new distribution x_{t+1}. It is convenient to regard x_t as the state of our system. In fact x_t is a Markov process with values in R^N[7].

To introduce an optimization problem we need some preliminary definitions. Let $X \in R^N$ be the set of probability vectors i.e., $X = \{x=(x_i):\ x_i \geq 0,\ \sum_{i=1}^{N} x_i = 1\}$. Let Δ be the set of mappings $\delta : X \to U$ and define $\Pi = \{\delta_t, t\in T:\ \delta_t \in \Delta\}$. Each element of Π is called a policy. A stationary policy is a policy which is independent of t i.e., $\delta_t = \delta$ for all $t \in T$. Hence we may identify Δ with the set of stationary policies. Now we introduce an R^p-valued cost function

$$C_\delta(x_0) = E_{x_0} \sum_{t=0}^{\infty} \beta^t x_t Q^{\delta(x_t)}\ ,\quad \delta \in \Delta \tag{2.2}$$

where $0 < \beta < 1$ and $Q^u \in R^{N\times p}$, $u \in U$. We wish to minimize $C_\delta(x_0)$ over Δ in the sense of Definition 2.1.

<u>Definition 2.1.</u> A policy δ_* is optimal if

$$C_\delta(x_0) \leq C_{\delta*}(x_0),\ \forall x_0 \in X \Rightarrow C_{\delta*}(x_0) = C_\delta(x_0),$$

where $\leqq$ means componentwise inequality.

Definition 2.2 [2]. Let $\Omega \in R^p$ be nonempty. A point $\xi \in \Omega$ is minimal if $\eta \leq \xi$, $\eta \in \Omega \Longrightarrow \eta = \xi$. The set of all minimal points in Ω is denoted $e(\Omega)$. Let $B^p(\mathcal{X})$ be the space of p-vector valued bounded functions with sup norm, where we may take any norm in R^p. Define on $B^p(\mathcal{X})$ mappings

$$\begin{aligned}(L_u f)(x) &= xQ^u + \beta \sum_s \{s|x,u\} f(T(x|s,u)), u \in U \\ (L_\delta f)(x) &= xQ^{\delta(x)} + \beta \sum_s \{s|x,\delta(x)\} f(T(x|s,\delta(x))), \ \delta \in \Delta\end{aligned} \tag{2.3}$$

and a multi-valued mapping

$$(L_* f)(x) = e(\bigcup_{u \in U} (L_u f)(x)). \tag{2.4}$$

Remark: Since $\bigcup_{u \in U}(L_u f)(x)$ has only a finite number of points, $(L_* f)(x)$ is nonempty and well-defined.

One can easily show that L_u, L_δ are contractions on $B^p(X)$ and that the unique fixed point of L_δ is the cost C_δ corresponding to the policy $\delta \in \Delta$.

Lemma 2.1. L_u and L_δ are monotone.

Proof. They are monotone componentwise.

Definition 2.3 [2]. A function $f_* \in B^p(X)$ is said to be a fixed point of L_* if $f_*(x) \in (L_* f_*)(x)$, $\forall_x \in X$. It is said to be minimal if $f(x) \leq f_*(x)$, $f \in L_* f \Longrightarrow f_* = f$.

We are interested in finding fixed points of L_* and in characterizing an optimal policy. We present two useful lemmas.

Lemma 2.2. Let $\{\delta_n\} \in \Delta$ be arbitrary. Then there exists a subsequence $\{\delta_{n_j}\}$ which is convergent to some $\delta \in \Delta$ pointwise i.e.,

$\delta_{n_j}(x) \to \delta(x)$, $\forall_x \in X$.

Proof. Let $V^n = \{V_i^n\}$ be the partition of X given by

$$V_i^n = \{x | \delta_n(x) = i\},$$

where we omit V_i^n whenever it is empty. Let $V^\infty = \prod_{n=1}^{\infty} V^n$ be the partition given by the product of all V^n. We assume $V^\infty = \{W^m\}$, $m=1,2,\cdots$.

Then each δ_n takes a single value on any W^m. Hence there exists a subsequence $\delta_{n_{1_j}}$ such that $\delta_{n_{1_j}}(x) = i_1$, $x \in W^1$. Similarly there exists a subsequence $\delta_{n_{2_j}}$ of $\delta_{n_{1_j}}$ such that $\delta_{n_{2_j}}(x) = i_2$, $x \in W^2$.

In general there exists a subsequence $\delta_{n_{m_j}}$ such that $\delta_{n_{m_j}}(x) = i_m, x \in W^m$.

Now take the diagonal sequence $\delta_{n_{j_j}}$, $j=1,2,\cdots$. Then except possibly first finite numbers of j $\delta_{n_{j_j}}(x) = i_m$ on W^m for any m. Therefore $\delta_{n_{j_j}}(x) \to \delta(x)$, where $\delta(x) = i_m$ on W^m.

Lemma 2.3. If $\delta_n(x) \to \delta(x)$ and $f_n(x) \to f(x)$, then

$$(L_{\delta_n} f_n)(x) \to (L_\delta f)(x) \ , \ \forall_x \in X \ .$$

Proof. $(L_{\delta_n} f_n)(x) - (L_\delta f)(x)$

$$= x(Q^{\delta_n(x)} - Q^{\delta(x)}) + \beta \sum_s [\{s|x,\delta_n(x)\} f_n(T(x|s,\delta_n(x)))$$

$$- \{s|x,\delta(x)\} f(T(x|s,\delta(x)))] \ .$$

For fixed $x \in X$, there exists an integer $N > 0$ such that $n \geq N \Rightarrow \delta_n(x) = \delta(x)$.

Hence L.H.S. $= \beta\{s|x,\delta(x)\}[f_n(T(x|s,\delta(x)) - f(T(x|s,\delta(x)))]$

$$\to 0 \text{ as } n \geq N \to \infty \ .$$

Policy improvement. We shall show that policy improvement is valid for our problem.

Theorem 2.1. For any $\delta_0 \in \Delta$ given there exists a sequence $\{\delta_n\} \in \Delta$ such that $L_{\delta_{n+1}} C_{\delta_n} \in L_* C_{\delta_n}$, $L_{\delta_{n+1}} C_{\delta_n} \leq C_{\delta_n}$ and $C_{\delta_{n+1}} \leq C_{\delta_n}$.

Proof. Note that $(L_* C_{\delta_n})(x)$ is nonempty and $L_{\delta_n} C_{\delta_n} = C_{\delta_n}$. Since $(L_{\delta_n} C_{\delta_n})(x) \in \bigcup_{u \in U} (L_u C_{\delta_n})(x)$, we can choose $u = u(x)$ such that $(L_u C_{\delta_n})(x) \leq (L_{\delta_n} C_{\delta_n})(x)$. Hence there exists $\hat{u} = \hat{u}(x)$ such that $(L_{\hat{u}} C_{\delta_n})(x) \in (L_* C_{\delta_n})(x)$ and $(L_{\hat{u}} C_{\delta_n})(x) \leq (L_{\delta_n} C_{\delta_n})(x)$ for any $x \in X$.

Now define $\delta_{n+1}(x) = \hat{u}(x)$. Then

$(L_{\delta_{n+1}} C_{\delta_n})(x) \le (L_{\delta_n} C_{\delta_n})(x) = C_{\delta_n}(x)$ and $(L_{\delta_{n+1}} C_{\delta_n})(x) \in (L_* C_{\delta_n})(x)$.

But $L_{\delta_{n+1}}$ is monotone, so

$$C_{\delta_{n+1}} \leftarrow L^m_{\delta_{n+1}} C_{\delta_n} \le \cdots \le L^2_{\delta_{n+1}} C_{\delta_n} \le L_{\delta_{n+1}} C_{\delta_n} \le C_{\delta_n} .$$

Lemma 2.4. Let δ_n be given as in Theorem 2.1. Then $C\delta_n \to C_\infty \in B^p(\mathbf{x})$ and there exists a subsequence δ_{n_j} of δ_n such that $\delta_{n_j}(x) \to \delta_\infty(x), \delta_\infty \in \Delta$. Furthermore, $C_\infty = L_{\delta_\infty} C_{\delta_\infty} = C_{\delta_\infty}$.

Proof. Since C_{δ_n} is monotone decreasing and bounded below, there exists a limit C_∞. By Lemma 2.2 there exists a subsequence δ_{n_j} such that $\delta_{n_j} \to \delta_\infty \in \Delta$ pointwise. By Theorem 2.1

$$C_{\delta_{n_j}} \le L_{\delta_{n_j}} C_{\delta_{n_j-1}} \le C_{\delta_{n_j-1}} .$$

Now we can pass to the limit $n_j \to \infty$ to obtain

$$C_\infty \le L_{\delta_\infty} C_\infty \le C_\infty .$$

But L_{δ_∞} has a unique fixed point C_{δ_∞}, so $C_\infty = C_{\delta_\infty} = L_{\delta_\infty} C_\infty$.

Theorem 2.2. There always exists a fixed point of L_*. In fact C_∞ given in Lemma 2.4. is a fixed point of L_*.

Proof. Since $L_{\delta_\infty} C_\infty = C_\infty$, $C_\infty \in \bigcup_{u \in U} L_u C_\infty$. Suppose there exists $\xi \in (L_* C_\infty)(x)$ such that $\xi \le C_\infty(x)$ strictly. Then there exists at least one component, say k^{th} one, such that $(\xi)_k < C_\infty(x)|_k$.

So there exists $\varepsilon > 0$ such that

$$(\xi)_k \le C_\infty(x)|_k - \varepsilon \quad . \tag{2.5}$$

Note that there exists $\delta \in \Delta$ such that $L_\delta C_\infty = \xi$ by definition.

Hence $(L_\delta C_\infty)(x)|_k \le C_\infty(x)|_k - \varepsilon$. Now define

$$\hat{\delta}(y) = \begin{cases} \delta_\infty(y), & y \neq x \\ \delta(x), & y = x . \end{cases}$$

Then $(L_{\hat{\delta}}C_\infty)(x) \le C_\infty(x)$ and

(2.6) $$(L_{\hat{\delta}}C_\infty)(x)|_k \le C_\infty(x)|_k - \varepsilon \ .$$

Now take n_j large enough and define

$$\hat{\delta}_{n_j}(y) = \begin{cases} \delta_{n_j}(y) \ , & y \neq x \\ \delta(x) \ , & y = x \end{cases}$$

then $\delta_{n_j}(y) \to \hat{\delta}(y)$ and

(2.7) $$(L_{\delta_{n_j}} C_{\delta_{n_j-1}})(x)|_{k'} = (L_{\delta_{n_j}} C_{\delta_{n_j-1}})(x)|_{k'} \ , \quad k' \neq k$$

(2.8) $$(L_{\delta_{n_j}} C_{\delta_{n_j-1}})(x)|_k - \frac{1}{3}\varepsilon \le (L_{\hat{\delta}}C_\infty)(x)|_k$$

(2.9) $$C_\infty(x)|_k = (L_{\delta_\infty} C_\infty)(x)|_k \le (L_{\delta_{n_j}} C_{\delta_{n_j-1}})(x)|_k + \frac{1}{3}\varepsilon \ ,$$

Now adding (2.6), (2.8), (2.9) we obtain

(2.10) $$(L_{\hat{\delta}_{n_j}} C_{\delta_{n_j-1}})(x)|_k \le (L_{\delta_{n_j}} C_{\delta_{n_j-1}})(x)|_k - \frac{1}{3}\varepsilon \ .$$

Combining (2.7) and (2.10) we obtain

$$L_{\hat{\delta}_{n_j}} C_{\delta_{n_j-1}} \le L_{\delta_{n_j}} C_{\delta_{n_j-1}} \qquad \text{strictly,}$$

which is a contradiction to the fact $L_{\delta_{n_j}} C_{\delta_{n_j-1}} \in L_* C_{\delta_{n_j-1}}$.

Hence $\xi \in (L_*C_\infty)(x)$, $\xi \le C_\infty(x) \Rightarrow \xi = C_\infty(x)$. Thus we have shown $C_\infty \in L_*C_\infty$.

Characterization of an optimal policy. When C_δ is real-valued, it is known that there always exists an optimal policy and that it is a unique fixed point of L_*[4]. Next we shall present a necessary and sufficient condition of an optimal policy.

Theorem 2.3. A stationary policy δ_* is optimal iff C_{δ_*} is a minimal fixed point of L_* .

Proof. Let δ_* be optimal. First we show $C_{\delta_*}(x) \in (L_*C_{\delta_*})(x)$, $\forall_x \in X$.

Note that $C_{\delta_*} = L_{\delta_*} C_{\delta_*}$ and $(L_{\delta_*} C_{\delta_*})(x) \varepsilon \bigcup_{u \varepsilon U} L_u C_{\delta_*}$. Suppose there exists $\xi \varepsilon (L_* C_{\delta_*})(x)$ such that $\xi \leq C_{\delta_*}(x)$ strictly. Then for some $\bar{u} = \bar{u}(x) \varepsilon U$, $\xi = (L_{\bar{u}} C_{\delta_*})(x)$. Define

$$\bar{\delta}(y) = \begin{cases} \delta_*(y), & y \neq x \\ \bar{u}, & y = x . \end{cases}$$

Then $(L_{\bar{\delta}} C_{\delta_*})(y) \leq C_{\delta_*}(y)$. By monotonicity of $L_{\bar{\delta}}$ we have

$$C_{\bar{\delta}} \leftarrow L^m_{\bar{\delta}} C_{\delta_*} \leq \cdots \leq L^2_{\bar{\delta}} C_{\delta_*} \leq L_{\bar{\delta}} C_{\delta_*} \leq C_{\delta_*} .$$

But C_{δ_*} is optimal, so $C_{\bar{\delta}} = C_{\delta_*}$. In particular

$$\xi = (L_{\bar{\delta}} C_{\delta_*})(x) = C_{\delta_*}(x),$$

which implies $C_{\delta_*} \varepsilon L_* C_{\delta_*}$.

Now we show that C_{δ_*} is minimal. Suppose there exists a fixed point f of L_*, then there exists $\delta \varepsilon \Delta$ such that $f = L_\delta f$. But L_δ has a unique fixed point C_δ, so $f = C_\delta$. Then $C_{\delta_*} \leqq f$.

Conversely, suppose C_{δ_*} is a minimal fixed point of L_*. Suppose for some $\delta \varepsilon \Delta$, $C_\delta \leq C_{\delta_*}$. Then we can constract a sequence δ_n as in Theorem 2.1 with $\delta_0 = \delta$. Then

$$C_{\delta_{n+1}} \leq L_{\delta_{n+1}} C_{\delta_n} \leq \cdots \leq C_\delta < C_{\delta_*} .$$

By Lemma 2.4 there exists a limit C_∞ of C_{δ_n} and δ_∞ of δ_{n_j}, a subsequence and $C_\infty = C_{\delta_\infty} = L_{\delta_\infty} C_\infty \leq C_\delta \leq C_{\delta_*}$. By Theorem 2.2 C_∞ is a fixed point of L_*. Now minimality of C_{δ_*} implies $C_{\delta_\infty} = C_\infty$, which necessarily yield $C_\delta = C_\infty = C_{\delta_*}$.

<u>Final remarks.</u> In the case of real-valued C_δ's we have presented an algorithm for an optimal policy and the minimal cost. The main problem in numerical computation is that X is uncountably infinite. But our algorithm involves only a finite number of vectors at each step. In the case of vector-valued C_δ's we cannot establish the existence of an optimal policy, but we may seek for an algorithm for fixed points of L_*.

We cannot directly extend our algorithm in [4] to the new situation and each step to find δ_n, C_{δ_n} is more complicated. So we shall discuss computational aspects elsewhere.

References

[1] E. B. Dynkin, Controlled random sequences, Theory Prob. Appl., X, 1965, pp.1-14.

[2] N. Furukawa, Vector-valued Markovian decision processes with countable state space, 1978.

[3] A. Ichikawa, A note on partially observable Markov decision problems, Reports Fac. Eng., Shizuoka University, 1978, pp.71-76.

[4] K. Sawaki and A. Ichikawa, Optimal control for partially observable Markov decision processes over an infinite horizon, J. Opns. Res. Soc. Japan, 21, 1978, pp.1-16.

[5] Y. Sawaragi and T. Yoshikawa, Discrete-time Markovian decision processes with incomplete state observation, Ann. Math. Stat., 41, 1970, pp.78-96.

[6] R. D. Smallwood and E. J. Sondik, Optimal control of partially observable processes over the finite horizon, Opns. Res., 21, 1973, pp.1071-1088.

[7] E. J. Sondik, The optimal control of partially observable Markov processes over the infinite horizon: Discounted Costs, Opns. Res., 26, 1978, pp.282-304.