ON THE EXISTENCE OF AN OPTIMAL STATIONARY I-POLICY IN NON-DISCOUNTED MARKOVIAN DECISION PROCESSES WITH INCOMPLETE STATE INFORMATION

Kurano, Masami
Department of Mathematics, Faculty of Education, Chiba University

https://doi.org/10.5109/13116

出版情報：統計数理研究. 17 (3/4), pp.75-81, 1977-03. Research Association of Statistical Sciences
バージョン：
権利関係：

ON THE EXISTENCE OF AN OPTIMAL STATIONARY I-POLICY IN NON-DISCOUNTED MARKOVIAN DECISION PROCESSES WITH INCOMPLETE STATE INFORMATION

By

Masami KURANO*

(Received October 20, 1976)

1. Introduction

Discrete-time Markovian decisin processes (MDP's) with incomplete state information have been investigated by many authors (for example [1], [4], [7]) and it was shown that MDP's with incomplete state information could be transformed to MDP's with complete state information by turning the state space to the new one which is the set of all probabilities on the state space and by turning the transition probability to new one which was constructed by the Bayes theorem.

Sawaragi, Yoshikawa [7] defined the Information policy (I-policy) and showed that there exists an optimal stationary I-policy in the discounted case when the action space is finite by making use of the results of Blackwel [2]. In this paper, we treat MDP's with incomplete state information, average cost criterion and an infinite planning horizon. We give sufficient conditions for existence of an optimal statinary I-policy. We make use of a technique used by Taylor [8] and Ross [5] for the proof. Also, we discuss the Howard's policy improvement.

2. Definitions and notations

In this section we develop the definitions and notations on a class of MDP's with incomplete state information in a similiar way to [7]. A *Borel* set X is a Borel subset of a complete separable metric space and the set of all probabilities on X is denoted by $P(X)$. If X and Y are non-empty Borel sets, the set of all conditional probabilities on Y given X is denoted by $Q(Y/X)$.

We denote the Cartesian product of X and Y by XY. $F(X)$ denotes the set of all bounded Baire functions on X. MDP's with incomplete state information is defined by S, A, M, q^s, q^m and c. S is the set of *states*, M is the set of *observation signals* and A is the set of *actions*. These sets are all finite and $S=\{1,\cdots,N\}$, $M=\{1,\cdots,L\}$ and $A=\{1,\cdots,K\}$. The *law of motion* q^s is an element of $Q(S/SA)$,

* Department of Mathematics, Faculty of Education, Chiba University, Chiba 280, Japan.

the characteristic of the *measuring system* $q^m \in Q(M/S)$, the *cost function* $c \in F(SA)$. Assume that the state of the system at time $n, n=0,1,2,\cdots$, is $s_n \in S$, and that we choose an action $a_n \in A$, then the system moves to a new state s_{n+1}, selected according to $q^s(s_{n+1}/s_n, a_n)$ and a cost $c(s_n, a_n)$ is incurred. We cannot observe the state s_{n+1} directly. We can only obtain an observation signal $m_{n+1} \in M$ generated according to $q^m(m_{n+1}/s_{n+1})$. We define $\mathcal{X} = P(S) = \{(x_1, x_2 \cdots, x_N)/x_i \geqq 0, \sum_1^N x_i = 1\}$. A *policy* ω is a sequence $\{\omega_1, \omega_2, \cdots\}$ where $\omega_n \in Q(A/D_n)$ and $D_n = \mathcal{X}AMA\cdots M$ ($2n+1$ factors) is the set of possible data concerning the history of the system up to the n-th stage. Given that we have obtained data $d_n = (x_0, a_0, m_1, a_1, \cdots, m_n) \in D_n$, we choose the n-th action a_n according to $\omega_n(a_n/d_n)$. Ω is the set of all policies. Let $\{S_t: t=0,1,2,\cdots\}$ and $\{\Delta_t: t=0,1,2,\cdots\}$ denote the sequence of states and of actions. For any policy $\omega \in \Omega$, $\beta \in (0,1)$ and any initial information $x \in \mathcal{X}$, let $g_n(x, \beta, \omega) = \sum_{t=0}^{n-1} \beta^t \sum_{s=1}^{N} E_\omega[c(S_t, \Delta_t)/S_0 = s] x_s$ and let $g(x, \beta, \omega) = \lim_{n\to\infty} g_n(x, \beta, \omega)$, where x_s is the s-th coordinate of x and β is a discounted factor. We define $g_n(x, \beta) = \inf_{\omega\in\Omega} g_n(x, \beta, \omega)$ and $g(x, \beta) = \inf_{\omega\in\Omega} g(x, \beta, \omega)$. Let $H_n = \mathcal{X}A\mathcal{X}A\cdots\mathcal{X}$ ($2n+1$ factors). An *I-policy* π is a sequence $\{\pi_1, \pi_2, \cdots\}$, where each π_n is an element of $Q(A/H_n)$. With any $a \in A$ and $m \in M$, we associate the operater T_{am} from $\mathcal{X}$ to $\mathcal{X}$ defined by

$$[T_{am}x]_{s'} = \frac{\sum_{s\in S} q^s(s'/s, a) q^m(m/s') x_s}{\sum_{s,s'\in S} q^s(s'/s, a) q^m(m/s') x_s}$$

where $[T_{am}x]_{s'}$ is the s'-th coordinate $T_{am}x \in \mathcal{X}$. By $x_t = T_{a_t m_t} x_{t-1}$, corresponding to any $d_n \in D_n$ an element $h_n = (x_0, a_0, x_1, \cdots, x_n) \in H_n$ is determined. Therefore, any I-policy is thought to be an element of Ω. An I-policy π is called *stationary* if each π_n is an element of $Q(A/\mathcal{X})$ and if there is a measurable fuction g from $\mathcal{X}$ to A such that $\pi_n(\{g(x)\}/x) = 1$ for all $x \in \mathcal{X}$. We need the following lemma.

LEMMA 1. ([1] [7])

(1) $$\lim_{n\to\infty} g_n(x, \beta) = g(x, \beta),$$

$\{g_n(x, \beta)\}$ and $\{g(x, \beta)\}$ satisfy the following functional equation:

(2) $$g_n(x, \beta) = \min_{a\in A} \{c(x, a) + \beta \sum_{m\in M} g_{n-1}(T_{am}x, \beta) p(m/x, a)\} \quad \textit{for all } x \in \mathcal{X} \textit{ and}$$

(3) $$g(x, \beta) = \min_{a\in A} \{c(x, a) + \beta \sum_{m\in M} g(T_{am}x, \beta) p(m/x, a)\}$$

where $p(m/x, a) = \sum_{s',s\in S} q^m(m/s') q^s(s'/s, a) x_s$ *and* $c(x, a) = \sum_{s\in S} c(s, a) x_s$.

3. The existence of an optimal stationary I-policy in the average cost criterion

For any $\omega \in \Omega$, let

$$G(x, \omega) = \limsup_{n\to\infty} (n+1)^{-1} \sum_{t=0}^{n} \sum_{s=1}^{N} E_\omega[c(S_t, \Delta_t)/S_0 = s] x_s .$$

Thus $G(x, \omega)$ is the expected average cost per unit time when the process starts in the initial information x on S and policy ω is used. The problem is to minimize

$G(x,\omega)$ with respect ω for all $x\in\mathfrak{X}$. We need the following results given by Taylor [3] and Ross [5].

THEOREM 1. *If there exist an* $f(x,y)\in F(\mathfrak{X}\mathfrak{X})$ *and a* $\gamma(y)\in F(\mathfrak{X})$ *such that*

$$f(x,y)+\gamma(y)=\min_{a\in A}[c(x,a)+\sum_{m\in M}f(T_{am}x,y)p(m/a,x)] \tag{4}$$

$$\textit{for all } x\in\mathfrak{X} \textit{ and some } y\in\mathfrak{X},$$

then there exists a statinary I-policy π^* *such that* $\gamma(y)=G(x,\pi^*)=\min_{\omega\in\Omega}G(x,\omega)$ *for all* $x\in\mathfrak{X}$ *and* π^* *is any policy which, for each* x, *prescribes an actin which minimizes the right side of* (4).

We shall obtain the sufficient conditions satisfying the assumption of Theorm 1.

LEMMA 2. (Satia and Lave [6], Proposition 4)

$g_n(\cdot,\beta)$ *is concave, that is,*

$$g_n\Big(\sum_1^p\lambda_i x^i,\beta\Big)\geqq\sum_1^p\lambda_i g_n(x^i,\beta)$$

$$\textit{for any } \lambda_i\geqq 0,\ \sum_1^p\lambda_i=1,\ \textit{any } x^i\in\mathfrak{X} \textit{ and } p=2,3,\cdots$$

Let $f_\beta(x,y)=g(x,\beta)-g(y,\beta)$. Then, we get, from (3),

$$f_\beta(x,y)+\gamma_\beta(y)=\min_a[c(x,a)+\beta\sum_{m\in M}f_\beta(T_{am}x,y)p(m/x,a)], \tag{5}$$

where $\gamma_\beta(y)=(1-\beta)g(y,\beta)$.

We shall set $A=\max_{s',s,a}q^s(s'/s,a)$, $B=\min_{s',s,a}q^s(s'/s,a)$, $C=\max_{m,s}q^m(m/s)$ and $D=\min_{m,s}q^m(m/s)$. Then, we can state the following Theorem.

THEOREM 2. *Let* $\lambda=\frac{BD}{AC}$. *If* $\sum_{s'}\max_{s,a,\bar{s},\bar{a}}|q^s(s'/s,a)-\lambda q^s(s'/\bar{s},\bar{a})|=\alpha<\lambda$, *then* $\{f_\beta(x,y),$ $0<\beta<1\}$ *is a uniformly bounded family of functions.*

PROOF. Choose E such that $\max_{s,a}|c(s,a)|\leqq E$ and let F be such that

$$2E\leqq 2E/(\lambda-\alpha)\leqq F. \tag{6}$$

Define $f_\beta^n(x,y)=g_n(x,\beta)-g_n(y,\beta)$. By (2), there are a^* and a^{**} such that

$$\begin{aligned} f_\beta^n(x,y)=c(x,a^{**})-c(y,a^*)+\beta\sum_m g_{n-1}(T_{a^{**}m}x,\beta)p(m/x,a^{**}) \\ -\beta\sum_m g_{n-1}(T_{a^*m}y,\beta)p(m/y,a^*)\,. \end{aligned} \tag{7}$$

Let $\lambda_{s'}(m)=[T_{a^*m}y]_{s'}-\lambda[T_{a^{*\prime}m}x]_{s'}$. For any $a\in A$, $x\in\mathfrak{X}$ and $m\in M$, we have $B\leqq p(s'/a,x)\leqq A$ for all s' and $D\leqq p(m/a,x)\leqq C$.

Therefore,

$$\begin{aligned} \lambda_{s'}(m)&=q^m(m/s')\Big(\frac{p(s'/a^*,y)}{p(m/a^*,y)}-\lambda\frac{p(s'/a^{**},x)}{p(m/a^{**},x)}\Big) \\ &\geqq q^m(m/s')\Big(\frac{\mathrm{B}}{\mathrm{C}}-\lambda\frac{\mathrm{A}}{\mathrm{D}}\Big), \end{aligned}$$

From $\lambda=\frac{\mathrm{BD}}{\mathrm{AC}}$, we get $\lambda_{s'}(m)\geqq 0$ for all $s'\in S$.

Thus, we can write

$$(8) \qquad [T_{a^*m}y]=\lambda[T_{a^{**}m}x]+\sum_s \lambda_s(m)\, I_s$$

where I_s is the unit probability distribution having one measure at s and $\sum_s \lambda_s(m)=1-\lambda$.

That is to say, $[T_{a^*m}y]$ is a mixture of $[T_{a^{**}m}x]$ and $\{I_s, s\in S\}$.

We shall show by induction for n that

$$(9) \qquad |f_\beta^n(x,y)|\leqq F\,.$$

By the definitin, $|f_\beta^1(x,y)|\leqq F$

Assume that (9) is true for $n=n-1$.

From (7), (8) and Lemma 2,

$$\begin{aligned} f^n(x,y) &\leqq c(x,a^{**})-c(y,a^*)+\beta\sum_m g_{n-1}(T_{a^{**}m}x,\beta)p(m/x,a^{**}) \\ &\quad -\beta\sum_m[g_{n-1}(T_{a^{**}m}x,\beta)\lambda p(m/y,a^*)+\sum_s g_{n-1}(I_s,\beta)\lambda_s(m)p(m/y,a^*)] \\ &\leqq 2E+\beta\sum_m g_{n-1}(T_{a^{**}m}x,\beta)\,(p(m/x,a^{**})-\lambda p(m/y,a^*)) \\ &\qquad +\beta\sum_{m,s} g_{n-1}(I_s,\beta)\,(-\lambda_s(m)p(m/y,a^*)) \\ &=2E+\beta\sum_m([g_{n-1}(T_{a^{**}m}x,\beta)-g_{n-1}(y,\beta)] \\ &\qquad \times[p(m/x,a^{**})-\lambda p(m/y,a^*)]) \\ &\qquad +\beta\sum_{m,s}(g_{n-1}(I_s,\beta)-g_{n-1}(y,\beta))\,(-\lambda_s(m)p(m/y,a^*))\,. \end{aligned}$$

Thus, by the assumption of induction,

$$\begin{aligned} f_\beta^n(x,y) &\leqq 2E+\beta F[\sum_m(p(m/x,a^{**})-\lambda p(m/y,a^*)) \\ &\qquad +\sum_{m,s}\lambda_s(m)p(m/y,a^*)] \\ &\leqq 2E+F[\sum_m|p(m/x,a^{**})-\lambda p(m/y,a^*)|+(1-\lambda)]\,. \end{aligned}$$

However,

$$\begin{aligned} &\sum_m|p(m/x,a^{**})-\lambda p(m/y,a^*)| \\ &\quad \leqq\sum_{m,s} q^m(m/s')|p(s'/a^{**},x)-\lambda p(s'/a^*,y)| \\ &\quad \leqq\sum_{m,s',s,\bar{s}} q^m(m/s')|q^s(s'/s,a^{**})-\lambda q^s(s'/\bar{s},a^*)|x_s y_{\bar{s}} \\ &\quad \leqq\sum_{s'}\max_{a,s,\bar{a},\bar{s}}|q^s(s'/s,a)-\lambda q^s(s'/\bar{a},\bar{s})| \\ &\quad =\alpha\,. \end{aligned}$$

Therefore, we obtain

$$f_\beta^n(x,y)\leqq 2E+F[\alpha+(1-\lambda)]\,.$$

From the definition of F, $f_\beta^n(x,y)\leqq F$.

By the similar way, we also get

$$f_\beta^n(x,y)\geqq -2E-F[\alpha+(1-\lambda)]\geqq -F\,.$$

That is to say, we obtain $|f^n_\beta(x,y)|\leqq F$.

Thus, the theorem is proved from (1). Q.E.D.

COROLLARY. *If* $\frac{C}{D}>\frac{NA^2}{NB^2+B}$, *then* $\{f_\beta(x,y),0<\beta<1\}$ *is a uniformly bounded family.*

PROOF.

$$\sum_{s'}\max_{s,a,\bar{s},\bar{a}}|q^s(s'/s,a)-\lambda q(s'/\bar{s},\bar{a})|\leqq N(A-\lambda B)\,.$$

$$N(A-\lambda B)>\lambda \text{ is equivalent to } \frac{D}{C}>\frac{NA^2}{NB^2+B}\,.$$

This completes the proof of the corollary.

REMARK. For example, $N=2$, $S=\{s_1,s_2\}$, $M=\{m_1,m_2\}$ and $A=\{a_1,a_2\}$. Suppose

$$\begin{bmatrix} q^m(m_1/s_1),\, q^m(m_2/s_1) \\ q^m(m_1/s_2),\, q^m(m_2/s_2) \end{bmatrix}=\begin{bmatrix} {}^4/_9,\, {}^5/_9 \\ {}^5/_9,\, {}^4/_9 \end{bmatrix}$$

and

$$\begin{bmatrix} (q^s(s_1/s_1,a_1),\, q^s(s_1/s_1,a_2)),\, (q^s(s_2/s_1,a_1),\, q^s(s_2/s_1,a_2)) \\ (q^s(s_1/s_2,a_1),\, q^s(s_1/s_2,a_2)),\, (q^s(s_2/s_2,a_1),\, q^s(s_2/s_2,a_2)) \end{bmatrix}=\begin{bmatrix} ({}^5/_4,\, {}^7/_{15}),\, ({}^4/_9,\, {}^8/_{15}) \\ ({}^4/_9,\, {}^8/_{15}),\, ({}^5/_9,\, {}^7/_{15}) \end{bmatrix}.$$

Then, $\frac{D}{C}=\frac{4}{5}$ and $\frac{NA^2}{NB^2+B}=\frac{50}{68}$.

Thus, $\frac{D}{C}>\frac{NA^2}{NB^2+B}$.

LEMMA 3. *Under the assumption of Theorem 2,* $\{f_\beta(x,y),0<\beta<1\}$ *is an equicontinuous family of functions.*

PROOF.

$$|f_\beta(x,y)-f_\beta(x',y')|\leqq|f_\beta(x,x')|+|f_\beta(y,y')|\,. \tag{10}$$

From (1) and Lemma 2, $g(\cdot,\beta)$ is concave.

Now, let, for any $\lambda', \lambda''_{s',a}(m)=[T_{am}x]_{s'}-\lambda'[T_{am}x']_{s'}$ $\lambda'(x,x')=\max\{\lambda'/\lambda''_{s'a}(m)\geqq 0$ for all $s',m,a\}$ and when $\lambda'=\lambda'(x,x')$ we set $\lambda''(x,x')=\max\limits_{s',m,a}\lambda_{s',a}(m)$. Then $\lambda'(x,x')\to 1$ and $\lambda''(x,x')\to 0$ as $x\to x'$. In a similar way to the proof of Theorm 2, we get

$$\begin{aligned}|f_\beta(x,x')|\leqq&\max_{a\in A}|c(x,a)-c(x',a)|+F\max_a\sum_m|p(m/x,a)-p(m/x',a)|\\&+F(1-\lambda'(x,x'))+F\times N\lambda''(x,x')\,.\end{aligned}$$

However, $\sum\limits_m|p(m/x,a)-p(m/x',a)|\leqq AB\sum\limits_i|x_i-x'_i|$.

Hence, by (10), $|f_\beta(x,y)-f_\beta(x',y')|\to 0$, independently of β, as $x\to x'$ and $y\to y'$.

We can state the following theorem. Q.E.D.

THEOREM 3. *Under the condition of Theorem 2 or Corollary,*

(a) *there exist* $\{f(x,y)\}$, $\{\gamma(y)\}$ *satisfying the equation* (4)

(b) *there exists an optimal stationary I-policy in the average cost criterion and*

(c) $\gamma_\beta(y)\to\gamma$ *for all* $y\in\mathscr{X}$ *as* $\beta\to 1$.

PROOF. From Theorem 2 and Lemma 3, $\{f_\beta(x,y)\}$ is a uniformly bounded equicontinuous family of functions.

Now $\{\gamma_\beta(y)\}$ is uniformly bounded.

Therefore, by the Ascoli-Arzela theorem, the bounded convergence theorem and (5), the theorem is easily proved. Q.E.D.

4. Policy improvement

We now discuss the problem of finding the optimal policy under the assumption of Theorem 2.

The following lemma can be proved in a similar way to the theorem. By any stationary I-policy π and any initial information $x\in\mathcal{X}$, the stationary Markov process on $\mathcal{X}$ is induced.

We denote by $Q^*(\cdot,x,\pi)$ the limiting state probability of the induced Markov process.

LEMMA 3. *Under the assumption of Theorem 2, for any stationary I-policy π, there exist an $f^\pi(x)\in F(\mathcal{X})$ and a constant γ^π such that*

$$f^\pi(x)+\gamma^\pi=c(x,\pi(x))+\sum_{m\in M}f^\pi(T_{\pi(x)m}x)p(m/x,\pi(x))$$

where $\pi(x)$ is the action chosen by π.

Define

$$\mathcal{E}_\pi(x,a)=f^\pi(x)+\gamma^\pi-[c(x,a)+\sum_{m\in M}f^\pi(T_{am}x)p(m/x,a)].$$

Let

$$\mathcal{E}_\pi(x)=\max_{a\in A}\mathcal{E}_\pi(x,a) \tag{11}$$

and

$$\mathcal{X}_\pi=\{x/\mathcal{E}_\pi(x)=0,x\in\mathcal{X}\}.$$

We define the new stationary I-policy π' such that if $x\in\mathcal{X}_\pi$, $\pi'(x)=\pi(x)$ and if $x\in\mathcal{X}_\pi$, $\pi(x)$ prescribes the action that maximizes the right hand side of (11).

We can state an extension of the improvement routines given by Howard [3].

THEOREM 4. *Under the assumption of Theorem 2,*

(a) $G(x,\pi')\leqq G(x,\pi)$ *for all* $x\in\mathcal{X}$

and

(b) *if* $Q^*(\mathcal{X}_\pi^c/x,\pi)>0$, then $G(x,\pi')<G(x,\pi)$.

References

[1] ÅSTRÖM, K. J,: *Optimal control of Markov processes with incomplete state information.* J. Math. Anal. Appl. **10** (1965), 174-205.

[2] BLACKWELL, D.: *Discounted dynamic programming.* Ann. Math. Statist. **36** (1965), 226-235.

[3] HOWARD, R. A.: Dynamic programming and Markov processes. John Wiley & Sons, Inc., New York, 1960.

[4] Rhenius, D.: *Incomplete information in Markovian decision models.* Ann. Statist. 2 (1974), 1327–1334.

[5] Ross, S.M.: *Arbitrary state Markovian decision processes.* Ann. Math. Statist. 39 (1968), 2118–2122,

[6] Satia, J.K. and Lave, R.E. J.: *Markovian decision processes with probabilistic observation of state.* Management Science, 20 (1973), 1–13.

[7] Sawaragi, Y. and Yoshikawa, T.: *Discrete-time Markovian decision processes with incomplete state observation.* Ann. Math. Statist. 41 (1970), 78–86.

[8] Taylor, H.: *Markovian sequential replacement processes.* Ann. Math. Statist. 36 (1965), 1677–1694.