Journal of the
Operations Research
Society of Japan

VOLUME 15 *June 1972* NUMBER 2

DISCRETE-TIME MARKOVIAN DECISION PROCESSES WITH AN UNKNOWN PARAMETER

– AVERAGE RETURN CRITERION –

MASAMI KURANO

Osaka University

(Received June 30, 1971 and Revised October 19, 1971)

1. Introduction

Discrete-time Markovian decision processes (MDP's) with an infinite planning horizon have been investigated by many authors in case that the transition probabilities are known to the decision maker. When the state and action spaces are finite, the non-randomized stationary optimal policy exists, and an iteration algorithm for searching the optimal policy is given in the average return criterion (for example [1] [2]). We treat MDP's in which nothing is known about transition probabilities to the decision maker. The Bayesian Analysis of MDP's with an unknown parameter was studied by J.J. Martin [3] and others. In this paper, we investigate the MDP's with an unknown parameter, a finite number of states, a finite number of actions and an infinite planning horizon, and it is shown that the policy $\pi_{a_0}{}^*$ which is constructed by using the maximum likelihood estimator for the unknown transition probabilities is optimal in the average return criterion.

2. Definitions and Notations on MDP's with an Unknown Parameter

In this section we shall give the definitions and notations on a class of MDP's with an unknown parameter. An MDP with an unknown parameter is a controlled dynamic stochastic system defined by S, A, $\boldsymbol{\theta}$, q and r. $S=\{1,2,\cdots,N\}$ is the set of states, $A=\{1,2,\cdots,K\}$ is the set of actions, $\Theta=\{1,2,\cdots,M\}$ is the set of possible values of a parameter, q is the set of stochastic matrices including an unknown parameter, and r is a return function defined on $S\times A$. At every $t=0, 1,\cdots$ one of a finite number of states $1,\cdots, N$ is observed. After each observation, the system is controlled by taking one of a finite number of actions $1,\cdots, K$. The action taken determines the probability distribution of the next state. Let $X_0, X_1, \cdots$ denote the sequence of observed states and Δ_0, $\Delta_1, \cdots$ the sequence of actions. The class C is the collection of all policies

$$R = \{D^1, D^2, \cdots\}$$

where

$$D^t = \{D_k{}^t;\ 1\leqq k\leqq K\}$$

$$D_k{}^t(X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t) = P(\Delta_t = k\,|\,X_0,\Delta_t, \cdots, \Delta_{t-1}, X_t)$$

for $t=0, 1, \cdots$ and for every $X_0, \Delta_0,\cdots, X_t$, $t=0, 1, \cdots$

$$D_k{}^t(X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t) \geqq 0$$

and

$$\sum_{k=1}^{K} D_k{}^t(X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t) = 1\,.$$

We shall assume throughout that

$$P(X_{t+1}=j\,|\,X_0, \Delta_0, \cdots, X_t=i, \theta)$$

$$=\sum_{k=1}^{K} q(j\,|\,i, k, \theta)\, D_k{}^t(X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t=i)$$

for $i, j \in S$, $\theta \in \Theta$; $t=0, 1, \cdots$ where $\{q(j/i, k, \theta)\}$ are the given stochastic matrices with an unknown parameter:

$$q(j\,|\,i, k, \theta) \geqq 0$$

and

$$\sum_{j=1}^{N} q(j\,|\,i, k, \theta) = 1$$

for $i \in S$, $k \in A$, $\theta \in \Theta$. Let C$'$ be the class of all policies R such that

$$D_k{}^t(X_0, \Delta_0, \cdots, X_t=i) = D_{ik}{}^t \qquad i \in S,\ k \in A$$

independent of $X_0, \cdots, X_{t-1}, \Delta_{t-1}$ and that $D_{ik}{}^t=0$ or $D_{ik}{}^t=1$. Let C$''$ be the class of all policies R (non-randomized stationary policies) such that

$$D_k{}^t(X_0, \Delta_0, \cdots, X_t=i) = D_{ik} \qquad i \in S,\ k \in A$$

independent of $X_0, \cdots, X_{t-1}, \Delta_{t-1}$, t and that $D_{ik}=0$ or $D_{ik}=1$. Denote by $F(X)$ the class of all functions from X to A. Thus there exists one to one correspondence between C$'$ and the class of all sequences $\{f_t, t=0, 1, \cdots\}$ of functions belonging to $F(S)$ by $D_{if_t(i)}{}^t=1$ and we can denote the policy contained in C$'$ by $\pi=\{f_0, f_1, \cdots\}$ and the policy contained in C$''$ by $f^{(\infty)}=\{f, f, \cdots\}$ in the same way. If $f^{(\infty)} \in$ C$''$ is used and the true value of a parameter is θ, the sequence X_t, $t=0, 1, \cdots$ is a Markov chain with stationary transition probabilities $\{p_{ij}(\theta)\}$ such that

$$p_{ij}(\theta) = q(j\,|\,i, f(i), \theta) \qquad i, j \in S\,.$$

Let $r^t(i, k)$ be the expected return ascribed to time t given that the system is observed in state i at time t and the action k is taken. We assume throughout that $r^t(i, k)=r(i, k)$ independent of t and that r is a given return function. Let $w_t(R, \theta)$, $t=0, 1, \cdots$ denote the expected return ascribed to time t when a policy R is used and the actual value of a parameter is θ. We suppose $X_0=i$ and define

$$g(i, R, \theta)=\liminf_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T} w_t(R, \theta) .$$

We shall make use of the following conditions.

(A) $|r(i, k)|\leq D<\infty, \quad i\in S,\ k\in A$.

(B) The Markov chain induced by any non-randomized stationary policy $f^{(\infty)}\in C''$ is completely ergodic for each $\theta\in\Theta$.

We have the following Lemma.

Lemma 1 ([1], [2]).

If conditions (A) and (B) hold, there exists a set of numbers $\{g(\theta), v_j(\theta)\}$ $j\in S,\ \theta\in\Theta$, satisfying

$$g(\theta)+v_i(\theta)=\underset{k\in A}{\mathrm{Max}}\left\{r(i, k)+\sum_{j=1}^{N} q(j|i, k, \theta)\, v_j(\theta)\right\},$$

$$\text{for} \quad i\in S \cdots (*)$$

and the following hold.

(i) $g(\theta)$ is uniquely determined by (*).

(ii) There exists a non-randomized stationary policy $f_*^{(\infty)}(\theta)$ such that

$$g(i, f_*^{(\infty)}(\theta), \theta)=\sup_{R\in C} g(i, R, \theta)$$

independently of $i\in S$, and $g(i, f_*^{(\infty)}(\theta), \theta)=g(\theta)$.

(iii) $f_*(\theta)=\{f_*(i, \theta), i\in S\}$ is a function which, for each i, prescribes the action that maximizes the right-hand side of (*).

3. Construction of Optimal Policies $\pi_{a_0}{}^*$ and Theorem

In this section we define an optimal policy of MDP's with an unknown parameter and construct the optimal policies $\pi_{a_0}{}^*$. We say that the policy $R^*\in C$ is optimal if $g(i, R^*, \theta)=g(\theta)$ for any $\theta\in\Theta$ under the conditions (A) and (B). The optimality of the policy R^* means that the average return per unit time induced by R^* is equal to the one induced when the decision maker behaves optimally as if the true value of the parameter was known to him. Denote by $H_t=S\times A\times\cdots\times S(2t+1$ factors) the set of possible histories of the process when some action is chosen in the t-th period. The policy R and the fixed θ determine the probabilities of H_t. Let us denote the probability by $p(h_t/\theta, R)$ where $h_t=(X_0, \Delta_0,\cdots, X_t)\in H_t$. We denote by $\theta_t(h_t)$ the parameter value which maximizes the likelihood function $p(h_t/\theta, R)$. That is to say, $\theta_t(h_t)$ is the maximum likelihood estimator of θ. We want to construct the policies $\pi_{a_0}{}^*$ which will be shown to be optimal. Let a_0 be any action and s_0 be any initial state. Policy $\pi_{a_0}{}^*=\{a_0, f_1^*, f_2^*,\cdots\}$ where $f_t^*\in F(H_t)$ is the following: "The decision-maker observes the initial state $X_0=s_0$ and makes the action $\Delta_0=a_0$. If $X_1=s_1$ is observed in the 1st period, compute $\theta_1(h_1)$ where $h_1=(s_0, a_0, s_1)$ and take the action $\Delta_1=f_1^*(h_1)=f_*(s_1, \theta_1(h_1))$ as the 2nd action where f_* is given by (iii) of Lemma 1. Similarly, if $h_t=(s_0, a_0,\cdots, s_t)\in H_t$, the history of the system until the t-th period is observed, compute the maximum likelihood estimator $\theta_t(h_t)$ and take $\Delta_t=f_t^*(h_t)=f_*(s_t, \theta_t(h_t))$ as the $(t+1)$-th action and so on." The policy $\pi_{a_0}{}^*$ means that we use the maximum likelihood estimator concerning the unknown parameter and we take an action optimally in believing that it really equals the actual value of the parameter. We need the following condition.

(C) For each $\theta_1, \theta_2\in\Theta$ such that $\theta_1\neq\theta_2$, $s\in S$ and $a\in A$

$$q(s'/s, a, \theta_1)\neq q(s'/s, a, \theta_2) \text{ for all } s'.$$

We can state the following theorem.

Theorem.

If the conditions (A), (B) and (C) hold, $\pi_{a_0}{}^*$ is optimal for any $a_0\in A$.

4. Proof of Theorem

Two lemmas will be given to prove the theorem. We associate with each $f\in F(S)$ (1) *an* $N\times 1$ column vector $r(f)$ whose s-th element is $r(s, f(s))$, and (2) *an* $N\times N$ stochastic matrix $Q_\theta(f)$ whose (s, s') element is

$q(s'/s, f(s), \theta)$. If we use a policy $\pi=(f_0, f_1, \cdots)\in C'$ and the system is initially in state s, the probability that the system will be in state s' at the t-th period is the (s, s') element of the matrix $Q_t(\pi, \theta)=Q_\theta(f_0)\cdots Q_\theta(f_t)$. Thus the average expected return from π up to the T-th time is

$$G^T(\pi, \theta) = \frac{1}{T+1}\left\{r(f_0)+\sum_{t=0}^{T-1} Q_t(\pi, \theta)\, r(f_{t+1})\right\}$$

and the average expected return per unit time is

$$G(\pi, \theta) = \liminf_{T\to\infty} Q^T(\pi, \theta)\,.$$

If $\pi=(f, f, \cdots)=f^{(\infty)}$, then

$$G^T(\pi, \theta) = \frac{1}{T+1}\left\{r(f)+\sum_{t=1}^{T} Q_\theta{}^t(f)\, r(f)\right\}$$

and

$$G(\pi, \theta)= Q_\theta{}^*(f)\, r(f)$$

where

$$Q_\theta{}^*(f) = \lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T} Q_\theta{}^t(f)$$

holds under the conditions (A) and (B), $Q_\theta{}^*(f)$ being the matrix of limiting state probabilities. It is well-known that all row vectors of $Q_\theta{}^*(f)$ are identical.

Lemma 2.

Let $\pi=\{f_0, f_1, \cdots\}$ and $f^{(\infty)}$ be any two policies where $f_i \in F(S)$ and $f\in F(S)$. For any integer m, if $\pi^m=(f_0', f_1', \cdots)$ such that $f_j'=f_j$, $j=0, 1, \cdots, m$, $f_j'=f$, $j=m+1, \cdots$, then $G(\pi^m, \theta)=G(f^{(\infty)}, \theta)$ under the conditions (A) and (B).

Proof.

Since

$$G^{m+T}(\pi^m, \theta) = \frac{1}{m+T+1}\left[r(f_0)+\sum_{k=0}^{m-1}\prod_{j=0}^{k} Q_\theta(f_j)\, r(f_{k+1})\right.$$

$$+\prod_{j=0}^{m} Q_\theta(f_j)\,\{\sum_{k=0}^{T} Q^k(f)\,r(f)\}]$$

$$=\frac{1}{m+T+1}\,\{r(f_0)+\sum_{k=0}^{m-1}\prod_{j=0}^{k} Q_\theta(f_j)\,r(f_{k+1})\}$$

$$+\frac{T+1}{m+T+1}\cdot\prod_{j=0}^{m} Q_\theta(f_j)\,G^T(f^{(\infty)},\theta)$$

we have

$$G(\pi^m,\theta)=\lim_{T\to\infty} G^T(\pi^m,\theta)=\lim_{T\to\infty} G^{m+T}(\pi^m,\theta)$$

$$=\prod_{j=0}^{m} Q_\theta(f_j)\,Q_\theta{}^*(f)\,r(f)=Q_\theta{}^*(f)\,r(f)=G(f^{(\infty)},\theta)\,.$$

This completes the proof.

Lemma 3 ([4]).

Let $\pi=(a_0, a_1,\cdots)$ be any sequence of actions. If condition (C) holds, for any integer n and any initial state s_0 there exists a number λ such that

(i) $0<\lambda<1$

(ii) $P\,\{\theta_n(h_n)\neq\theta/\theta,\pi\}\leq(N-1)\,\lambda^n$.

Proof.

$$P_l=P\,\{\theta_n(h_n)\neq l/\theta=l,\pi\}\leq\sum_{j\neq l}P\left\{\frac{p(h_n/j,\pi)}{p(h_n/l,\pi)}>1/l,\pi\right\}$$

$$=\sum_{j\neq l}P\left(\left\{\frac{p(h_n/j,\pi)}{p(h_n/l,\pi)}\right\}^\alpha>1/l,\pi\right)$$

for any α such that $0<\alpha<1$. Consequently,

$$P_l\leq\sum_{j\neq l}E_l\left(\frac{p(h_n/j,\pi)}{p(h_n/l,\pi)}\right)^\alpha$$

Then,

$$E_l\left(\frac{p(h_n/j,\pi)}{p(h_n/l,\pi)}\right)=\sum_{h_n\in H_n}p^{\alpha}(h_n/j,\pi)\,p^{1-\alpha}(h_n/l,\pi)$$

$$=\sum_{s_1\in S}q^{\alpha}(s_1/s_0,a_0,j)\,q^{1-\alpha}(s_1/s_0,a_0,l)\cdots$$

$$\times\sum_{s_n\in S}q^{\alpha}(s_n/s_{n-1},a_{n-1},j)\,q^{1-\alpha}(s_n/s_{n-1},a_{n-1},l)\,.$$

For any s, a, by Hölder's inequality,

$$\sum_{s'\in S}q^{\alpha}(s'/s,a,j)\,q^{1-\alpha}(s'/s,a,l)$$

$$<\{\sum_{s'\in S}q(s'/s,a,j)\}^{\alpha}\,\{\sum_{s'\in S}q(s'/s,a,l)\}^{1-\alpha}=1$$

under the condition (C).

Therefore, since

$$\operatorname*{Max}_{s,a,j\neq h}\{\sum_{s'}q^{\alpha}(s'/s,a,j)\,q^{1-\alpha}(s'/s,a,h)\}=\lambda<1\,,$$

we have

$$E_l\left(\frac{p(h_n/j,\pi)}{p(h_n/l,\pi)}\right)^{\alpha}\leqq\lambda^n\qquad\text{and}\qquad P_l\leq(N-1)\,\lambda^n\,.$$

This completes the proof.

For any two policies $\pi^1=\{f_0^1,f_1^1,f_2^1,\cdots\},\pi^2=\{f_0^2,f_1^2,\cdots\}$, where $f_j^i\in F(H_j)$, we write $\pi^1=\pi^2$ if $f_j^1=f_j^2$ for all j and for any two column vectors W_1, W_2 and we write $W_1\geqq W_2$ if every component of W_1 is larger than or equal to the corresponding component of W_2.

Proof of Theorem.

We shall fix any θ, and for any integer m, we define $\pi_{a_0}{}^m$ by

$$\pi_{a_0}{}^m=\{a_0,f_1^*,f_2^*,\cdots,f_m^*,f_*^{(\infty)}(\theta)\}\,.$$

Let $\pi_{a_0}{}^*=\{a_0,f_1^*,f_2^*,\cdots\}$. From Lemma 3,

$$P\left(\bigcup_{t=m+1}^{\infty}\{\theta_t(h_t)\neq\theta\}/\theta,\pi\right)\leqq\sum_{t=m+1}^{\infty}P(\theta_t(h_t)\neq\theta/\theta,\pi)$$

$$\leqq (N-1) \sum_{t=m+1}^{\infty} \lambda^t = (N-1) \frac{\lambda^{m+1}}{1-\lambda}.$$

Thus, since

$$f_t^*(h_t) = f_*(s_t, \theta_t(h_t)) \quad \text{where} \quad h_t = (s_0, a_0, \cdots, s_t),$$

$$P(\pi_{a_0}{}^m \neq \pi_{a_0}{}^* / \theta, \pi) = P\left(\bigcup_{t=m+1}^{\infty} \{f_*(s_t, \theta_t(h_t)) \neq f_*(s_t, \theta)\} / \theta, \pi\right)$$

$$\leqq P\left(\bigcup_{t=m+1}^{\infty} \{\theta_t(h_t) \neq \theta\} / \theta, \pi\right) \leq (N-1) \frac{\lambda^{m+1}}{1-\lambda}.$$

Therefore, if $W_t(R, \theta)$ is the $N \times 1$ column vector whose s-th component is the expected return ascribed to time t in using the policy R, the actual value of a parameter being θ and starting from the initial state $X_0 = s$,

$$G^{m+T}(\pi_{a_0}{}^m, \theta) - G^{m+T}(\pi_{a_0}{}^*, \theta)$$

$$= \frac{1}{m+T+1} \sum_{t=m+1}^{m+T} \{W_t(\pi_{a_0}{}^m, \theta) - W_t(\pi_{a_0}{}^*, \theta)\}$$

$$\leqq \frac{1}{m+T+1} 2T(N-1) D \frac{\lambda^{m+1}}{1-\lambda} I.$$

where I is the $N \times 1$ column vector whose components are all 1. When $T \to \infty$, we get

$$G(\pi_{a_0}{}^m, \theta) - G(\pi_{a_0}{}^*, \theta) \leqq 2(N-1) D \frac{\lambda^{m+1}}{1-\lambda} I.$$

However, by Lemma 1 and Lemma 2

$$G(\pi_{a_0}{}^m, \theta) = G(f_*{}^{(\infty)}(\theta), \theta) = g(\theta) I.$$

When $m \to +\infty$, we have $g(\theta)I - G(\pi_{a_0}{}^*, \theta) \leq 0$.

On the other hand, since $g(\theta)I$ is the optimal average return when the true value of the parameter is θ,

$$g(\theta) I - G(\pi_{a_0}{}^*, \theta) \geqq 0.$$

Therefore,

$$g(\theta) I = G(\pi_{a_0}{}^*, \theta).$$

This completes the proof.

Acknowledgement

The author would like to express his heartiest thanks to Professor M. Sakaguchi of Osaka University for his kind suggestions and encouragements while this paper was being prepared. The author also thanks the referee for his various suggestions for improving the original version of the paper.

References

[1] Howard, Ronald A., *Dynamic Programming and Markov Process*, The M.I.T. Press, 1960.

[2] Derman, Cyrus, "On sequential decisions and Markov chains," *Management Sci.*, **9** (1962), 16–24.

[3] Martin, J.J., *Bayesian Decision Problems and Markov Chains*, John Wiley & Sons, 1967.

[4] Korsh, James F., "On decision and information concerning an unknown parameter," *Information and Control*, **16** (1970), 123–127.