Journal of the Operations Research
Society of Japan
Vol. 24, No. 2, June 1981

SEMI-MARKOV DECISION PROCESSES WITH INCOMPLETE STATE OBSERVATION —AVERAGE COST CRITERION—

Kazuyoshi Wakuta
Nagaoka Technical College

(Received September 28, 1978; Final September 26, 1980)

Abstract In this paper we study the infinite planning horizon, countable state, semi-Markov decision processes (SMDP's) with the incomplete state observation under the average cost criterion. We show that this model can be transformed to ordinary SMDP's, i.e., SMDP's with the complete state observation. Furthermore, when the action space is assumed to be finite, we present some sufficient conditions under which there exists an optimal stationary I-policy, and the method of successive approximations is applicable for obtaining a solution of the optimality equation.

1. Introduction

Markov decision processes with the incomplete state observation have been investigated by many authors, for example, [4], [7], [8] and [10]. The applicability of this model, however, is restricted because the time spent in a state is always required to be a unit time. Dropping this requirement, we have semi-Markov decision processes (SMDP's) with the incomplete state observation. Ordinary SMDP's, i.e., SMDP's with the complete state observation were introduced by Jewell [3] and have been studied by several authors, for example, Ross [5]. SMDP's with the incomplete state observation were formulated as a partially observed semi-Markov optimization problem by White [11], where the finite planning horizon, finite state, discrete time case was analyzed. We consider here the infinite planning horizon, countable state, continuous time case under the average cost criterion. Approaching to this problem by the way analogous to Sawaragi and Yoshikawa [8], we show that our model defined below can be transformed to SMDP's with the complete state observation, where the states are probabilities on the set of the states

of the original model. Furthermore, when the action space is assumed to be finite, we present some sufficient conditions imposed on the original model under which there exists an optimal stationary I-policy, and the method of successive approximations is applicable for obtaining a solution of the optimality equation.

Throughout this paper, we follow [2] and [8] for the probabilistic notations, terms and properties.

We denote a class of SMDP's with the incomplete state observation defined below by SMDP-II and a class of SMDP's with the complete state observation by SMDP-I. SMDP-II is defined by $(S, M, A, p_S, p_F, q, \phi_0, c)$. S is a countable set, the set of states of a system. M is a countable set, the set of observation signals. A is a Borel set of a complete separable metric space, the set of actions. p_S is a conditional probability $p_S(s'|s,a)$ of the next state s', given the current state s and the action a to be chosen. p_F is a conditional probability $p_F(\cdot|s,a,s')$ of the time until the transition from s to s' occurs, given that the next state is s'. p_F has a density $f(t|s,a,s')$ with respect to Lebesgue measure λ where f is a Borel measurable function of (t,s,a,s'). p_S and p_F determine the law of motion of the system. q is a conditional probability $q(m|s)$ of the observation signal m, given that the state of the system is s, the characteristic of the observation system. ϕ_0 is an element of $\Phi = P(S)$, where $P(S)$ is the set of all probabilities on S, the initial distribution of the system. c is a bounded Borel measurable function $c(t|s,a)$ of (t,s,a), the cost function.

We cannot directly observe the state of the system. But we can only obtain the observation signal generated according to q. We note that we can observe the time when the transition occurs.

In order to ensure that an infinite number of transitions does not occur in a finite interval of time, the following condition is imposed throughout.

Condition 1 (Ross [6]). There exists $\delta > 0$, $\varepsilon > 0$ such that for all s and a,

$$\sum_{s'} p_F([0,\delta]|s,a,s')p_S(s'|s,a) \leq 1-\varepsilon.$$

To select actions, a policy is needed. A policy ω is a sequence $\{\omega_0, \omega_1, \ldots\}$, where each ω_n is a conditional probability $\omega_n(\cdot|h_n)$ on A, given the observable history $h_n = (\phi_0,a_0,t_1,m_1,\ldots,a_{n-1},t_n,m_n) \in H_n = \Phi\times(A\times R_+\times M)^n$, where $R_+ = [0,+\infty)$. Φ is metrizable by introducing the distance

$$d(\phi,\phi') = \sum_{i\in S} |\phi(i) - \phi'(i)|, \quad \phi, \phi' \in \Phi.$$

The topology introduced by this metric is equivalent to the weak topology.

We introduce the discrete topology on S. Then, Φ is a complete separable metric space, and $*$-σ-algebra in Φ is identical to the σ-algebra generated by the metric. Then, we define a conditional probability $q^P(i|\phi)$ on S, given Φ by $q^P(i|\phi)=\phi(i)$, $\phi\in\Phi$. Hence, any policy ω, together with q^P, p_S, p_F and q, defines a conditional distribution p_ω on the infinite product set $S\times(A\times S\times R_+\times M)^N$ of futures of the system, given the initial distribution ϕ_0 (N denotes the set of natural numbers), i.e., it defines

$$p_\omega\{\cdot|\phi_0\} = q^P \otimes \bigotimes_{n=0}^{\infty} (\omega_n \otimes p_S \otimes p_F \otimes q),$$

where $\otimes$ means the product of conditional probabilities (cf. Hinderer [2], Appendix 3). When a policy ω is applied, the expected average cost function on Φ is defined by

$$J_\omega(\phi_0) = \overline{\lim_{n\to\infty}} \frac{E_\omega\left[\sum_{i=0}^{n-1} c(t_{i+1}|s_i,a_i)|\phi_0\right]}{E_\omega\left[\sum_{i=1}^{n} t_i \mid \phi_0\right]},$$

where E_ω denotes the expectation by the conditional distribution p_ω. Then, our optimization problem is to minimize $J_\omega(\phi_0)$ among all policies. We say that ω^* is optimal if $J_\omega{}^*(\phi_0)\le\inf_\omega J_\omega(\phi_0)$ for all ϕ_0.

2. The Construction of a New Model

In this section we shall construct a new model with the complete state observation equivalent to one defined in the preceding section.

Let the conditional probability of s_n be denoted by $q_n=q_n(\cdot|h_n)$, given the observable history h_n. Using the Bayesian formula, we obtain the following relation of q_n: for any $s_{n+1}\in S$,

$$\begin{aligned} & q_{n+1}(s_{n+1}|h_{n+1}) \\ &= q_{n+1}(s_{n+1}|h_n,a_n,t_{n+1},m_{n+1}) \\ &= \frac{\sum_{s_n} v(s_n,a_n,s_{n+1},t_{n+1},m_{n+1}) q_n(s_n|h_n)}{\sum_{s_n}\sum_{s_{n+1}} v(s_n,a_n,s_{n+1},t_{n+1},m_{n+1}) q_n(s_n|h_n)}, \end{aligned} \tag{2.1}$$

where $v=f(t_{n+1}|s_n,a_n,s_{n+1})q(m_{n+1}|s_{n+1})p_S(s_{n+1}|s_n,a_n)$. Letting q_n correspond to an element $\phi \in \Phi$ by

$$q_n(s_n|h_n) = \phi_n(s_n) = q^P(s_n|\phi_n),\ s_n \in S,$$

we see that the right side of (2.1) is Borel measurable in $(\phi_n,a_n,t_{n+1},m_{n+1})$. Hence, there exists a Borel measurable map u: $\Phi \times A \times R_+ \times M \to \Phi$ defined by

$$\phi_{n+1}(s_{n+1}) = \frac{\sum_{s_n} v(s_n,a_n,s_{n+1},t_{n+1},m_{n+1})q^P(s_n|\phi_n)}{\sum_{s_n}\sum_{s_{n+1}} v(s_n,a_n,s_{n+1},t_{n+1},m_{n+1})q^P(s_n|\phi_n)} \tag{2.2}$$

$$= u(\phi_n,a_n,t_{n+1},m_{n+1})(s_{n+1}),\ s_{n+1} \in S,$$

(cf. Hinderer [2], Remarks, p.85).

By repeated use of u, corresponding to any observable history h_n, $b_n = (\phi_0,a_0,t_1,\phi_1,\cdots,a_{n-1},t_n,\phi_n) \in B_n = \Phi \times (A \times R_+ \times \Phi)^n$ is determined Borel measurably, where B_n is the set of the possible information concerning the histories of the system. Then, we define a new policy π which depends only on the possible information. We call this policy as information policy (I-policy) according to Sawaragi and Yoshikawa [8].

An I-policy π is a sequence $\{\pi_0, \pi_1,..\}$, where each π_n is a conditional probability $\pi_n(\cdot|b_n)$ on A, given the possible information b_n. An I-policy π is said to be stationary if there exists a Borel measurable map $f:\Phi \to A$ such that

$$\pi_n(f(\phi_n)|\phi_0,a_0,t_1,\phi_1,\cdots,a_{n-1},t_n,\phi_n) = 1 \text{ for all } \phi_n.$$

For any I-policy π, we define a policy $\omega^\pi=\{\omega_0^\pi, \omega_1^\pi,..\}$ by

$$\omega_n^\pi(\cdot|h_n) = \pi_n(\cdot|b_n^h),$$

where b_n^h is an element of B_n which corresponds to $h_n \in H_n$. Then π and ω^π assign the same conditional probability to A. Hence, the set of all I-policies is regarded as a subset of the set of all policies. Any I-policy π, together with q^P, p_S, p_F, q and u, defines a conditional distribution $p_{\pi\phi}$ on the infinite product set $S \times (A \times S \times R_+ \times M \times \Phi)^N$, i.e., it defines

$$p_{\pi\phi}\{\cdot|\phi_0\} = q^P \otimes \bigotimes_{n=0}^{\infty} (\pi_n \otimes p_S \otimes p_F \otimes q \otimes u).$$

For any I-policy π, the expected average cost function on Φ is defined by

$$J_\pi(\phi_0) = \overline{\lim_{n\to\infty}} \frac{E_{\pi\phi}\left[\sum_{i=0}^{n-1} c(t_{i+1}|s_i,a_i) \mid \phi_0\right]}{E_{\pi\phi}\left[\sum_{i=1}^{n} t_i \mid \phi_0\right]}$$

where $E_{\pi\phi}$ denotes the expectation by the conditional distribution $p_{\pi\phi}$. We also define $p_{\omega\phi}$ and $E_{\omega\phi}$ in the same way for any policy ω. Then, $J_\omega(\phi_0)$ can be rewritten by $E_{\omega\phi}$ in place of E_ω.

Let for all s and a,

$$\bar{c}(s,a) = \sum_{s'} \int_0^\infty c(t|s,a)p_F(dt|s,a,s')p_s(s'|s,a)$$

and

$$\bar{\tau}(s,a) = \sum_{s'} \int_0^\infty t p_F(dt|s,a,s')p_s(s'|s,a).$$

We note that the expected cost during the transition interval and the expected length of the transition interval depend only on the parameters of the process through $\bar{c}$, $\bar{\tau}$ and p_s. Then, for any ω and π,

$$J_\cdot(\phi_0) = \overline{\lim_{n\to\infty}} \frac{E_{\cdot\phi}\left[\sum_{i=0}^{n-1} \bar{c}(s_i,a_i) \mid \phi_0\right]}{E_{\cdot\phi}\left[\sum_{i=0}^{n-1} \bar{\tau}(s_i,a_i) \mid \phi_0\right]}.$$

Let for any ϕ and a,

$$c^\phi(\phi,a) = \sum_s \bar{c}(s,a)q^p(s|\phi)$$

and

$$\tau^\phi(\phi,a) = \sum_s \bar{\tau}(s,a)q^p(s|\phi).$$

Lemma 1. For any ω and π,

(a) $E_{\cdot\phi}[\bar{c}(s_n,a_n)|\phi_0] = E_{\cdot\phi}[c^\phi(\phi_n,a_n)|\phi_0]$,

(b) $E_{\cdot\phi}[\bar{\tau}(s_n,a_n)|\phi_0] = E_{\cdot\phi}[\tau^\phi(\phi_n,a_n)|\phi_0]$, $\phi_0 \in \Phi$.

Proof: We prove only (a) for ω.

$$E_{\omega\phi}[\bar{c}(s_n,a_n)|\phi_0]$$

$$= \int_{S\times\Phi\times A} p_{\omega\phi}\{d(s_n,\phi_n,a_n)|\phi_0\}\ \bar{c}(s_n,a_n)$$

$$= \int_{\Phi\times A} p_{\omega\phi}\{d(\phi_n,a_n)|\phi_0\} \sum_{s_n} p_{\omega\phi}\{s_n|\phi_0,\phi_n,a_n\}\ \bar{c}(s_n,a_n)$$

$$= \int_{\Phi\times A} p_{\omega\phi}\{d(\phi_n,a_n)|\phi_0\} \sum_{s_n} q^{p}(s_n|\phi_n)\bar{c}(s_n,a_n)$$

$$= \int_{\Phi\times A} p_{\omega\phi}\{d(\phi_n,a_n)|\phi_0\}\ c^{\phi}(\phi_n,a_n)$$

$$= E_{\omega\phi}[c^{\phi}(\phi_n,a_n)|\phi_0].$$

Theorem 2.1. For any fixed sequence of actions $\{a_0,a_1,\ldots\}$, where a_n is the action to be chosen during the n-th transition interval, (i) the stochastic process $(\phi,t) = \{\phi_n,t_n;\ n\in\mathbf{N}\}$ is a Markov renewal process, (ii) given that the process has just entered ϕ_n, the probability that the next transition will be into ϕ_{n+1} depends only on ϕ_n and a_n, and is given by

$$q^{\phi}(\Gamma|\phi_n,a_n) = \sum_{s_n}\sum_{s_{n+1}}\sum_{m_{n+1}} \int_{\bar{\Gamma}_m} v(s_n,a_n,s_{n+1},t_{n+1},m_{n+1})\lambda(dt_{n+1})q^{p}(s_n|\phi_n),$$

where for any Borel set Γ of Φ,

$$\bar{\Gamma} = \bar{\Gamma}\ (\phi_n,a_n,\Gamma) = \{(t_{n+1},m_{n+1});\ u(\phi_n,a_n,t_{n+1},m_{n+1})\in\Gamma\}$$

and

$$\bar{\Gamma}_m = \bar{\Gamma}_m(m_{n+1}) = \{t_{n+1};\ (t_{n+1},m_{n+1})\in\bar{\Gamma}\},$$

(iii) conditional on the event that the next state is ϕ_{n+1}, the time until the transition from ϕ_n to ϕ_{n+1} occurs is a random variable with the probability $p^{\phi}(\cdot|\phi_n,a_n,\phi_{n+1})$ which satisfies for any Borel set B of R_+,

$$p^{\phi}(B|\phi_n,a_n) = \int_{\Phi} p^{\phi}(B|\phi_n,a_n,\phi_{n+1})q^{\phi}(d\phi_{n+1}|\phi_n,a_n)$$

$$= \sum_{s_n}\sum_{s_{n+1}} p_F(B|s_n,a_n,s_{n+1})p_s(s_{n+1}|s_n,a_n)q^{p}(s_n|\phi_n).$$

Proof:

(i) $$p\{\phi_{n+1}\in\Gamma,t_{n+1}\in B|\phi_0,a_0,t_1,\phi_1,\ldots,a_n\}$$

$$= p\{(t_{n+1},m_{n+1})\in\bar{\Gamma},t_{n+1}\in B|\phi_0,a_0,t_1,\phi_1,\ldots,a_n\}$$

$$= \sum_{s_n}\sum_{s_{n+1}} p\{(t_{n+1},m_{n+1})\in\bar{\Gamma},t_{n+1}\in B|s_n,s_{n+1},\phi_0,\ldots,a_n\}$$

$$\times\ p\{s_{n+1}|s_n,\phi_0,..,a_n\}q^p(s_n|\phi_n)$$

$$= \sum_{s_n}\sum_{s_{n+1}}\sum_{m_{n+1}} \int_{\bar{\Gamma}m\cap B} v(s_n,a_n,s_{n+1},t_{n+1},m_{n+1})\lambda(dt_{n+1})q^p(s_n|\phi_n).$$

Hence, we have

$$p\{\phi_{n+1}\in\Gamma, t_{n+1}\in B|\phi_0,a_0,t_1,\phi_1,..,a_n\}$$

$$= p\{\phi_{n+1}\in\Gamma, t_{n+1}\in B|\phi_n,a_n\}.$$

Therefore, $(\phi,t) = \{\phi_n,t_n;\ n\in\mathbf{N}\}$ is a Markov renewal process in the sense of Çinlar [1].

(ii) $p\{\phi_{n+1}\in\Gamma|\phi_0,a_0,t_1,\phi_1,..,a_n\}$

$$= p\{(t_{n+1},m_{n+1})\in\bar{\Gamma}|\phi_0,a_0,t_1,\phi_1,..,a_n\}$$

$$= \sum_{s_n}\sum_{s_{n+1}} p\{(t_{n+1},m_{n+1})\in\bar{\Gamma}|s_n,s_{n+1},\phi_0,..,a_n\}$$

$$\times\ p\{s_{n+1}|s_n,\phi_0,..,a_n\}q^p(s_n|\phi_n)$$

$$= \sum_{s_n}\sum_{s_{n+1}}\sum_{m_{n+1}} \int_{\bar{\Gamma}m} v(s_n,a_n,s_{n+1},t_{n+1},m_{n+1})\lambda(dt_{n+1})q^p(s_n|\phi_n)$$

Hence, we have

$$p\{\phi_{n+1}\in\Gamma|\phi_0,a_0,t_1,\phi_1,\ldots,a_n\}$$

$$= p\{\phi_{n+1}\in\Gamma|\phi_n,a_n\}$$

$$= q^{\phi}(\Gamma|\phi_n,a_n).$$

(iii) $p\{B|\phi_0,a_0,t_1,\phi_1,\ldots,a_n\}$

$$= \sum_{s_n}\sum_{s_{n+1}} p\{B|s_n,s_{n+1},\phi_0,\ldots,a_n\}$$

$$\times\ p\{s_{n+1}|s_n,\phi_0,..,a_n\}q^p(s_n|\phi_n)$$

$$= \sum_{s_n}\sum_{s_{n+1}} p_F(B|s_n,a_n,s_{n+1})p_s(s_{n+1}|s_n,a_n)q^p(s_n|\phi_n).$$

Hence, we have

$$p\{B|\phi_0,a_0,t_1,\phi_1,\ldots,a_n\}$$

$$= p\{B|\phi_n,a_n\}$$

$$= p^{\phi}(B|\phi_n,a_n).$$

Then, (iii) follows from (i) and (ii).

Lemma 2. For any policy ω, there exists an I-policy π which satisfies

$$\text{(a)}\quad E_{\pi\phi}[c^{\phi}(\phi_n,a_n)|\phi_0] = E_{\omega\phi}[c^{\phi}(\phi_n,a_n)|\phi_0],$$

(2.3)

$$\text{(b)}\quad E_{\pi\phi}[\tau^{\phi}(\phi_n,a_n)|\phi_0] = E_{\omega\phi}[\tau^{\phi}(\phi_n,a_n)|\phi_0],\ \phi_0\in\Phi.$$

Proof: For any given policy ω, we define an I-policy $\pi^{\omega} = \{\pi_0^{\omega},\pi_1^{\omega},\ldots\}$ by

$$\pi_0^{\omega}(\cdot|\phi_0) = \omega_0(\cdot|\phi_0)$$

and

$$\pi_n^{\omega}(\cdot|\phi_0,a_0,t_1,\phi_1,\ldots,\phi_n)$$

$$= \sum_{m_1,\ldots,m_n} \omega_n(\cdot|\phi_0,a_0,t_1,m_1,\ldots,m_n)$$

$$\times\ p_{\omega\phi}\{(m_1,\ldots,m_n)|\phi_0,a_0,t_1,\phi_1,\ldots,\phi_n\},$$

where each $b_n^h = (\phi_0,a_0,t_1,\phi_1,\ldots,\phi_n)$ corresponds to $h_n = (\phi_0,a_0,t_1,m_1,\ldots,m_n)$. From the construction of π^{ω}, both ω and π^{ω} assign the same conditional probability

$$p_{\pi^{\omega}\phi}\{\cdot|\phi_0,a_0,t_1,\phi_1,\ldots,\phi_n\} = p_{\omega\phi}\{\cdot|\phi_0,a_0,t_1,\phi_1,\ldots,\phi_n\}$$

to A. By Theorem 2.1,

$$p_{\omega\phi}\{\cdot|\phi_0,a_0,t_1,\phi_1,\ldots,\phi_n,a_n\} = p_{\pi^{\omega}\phi}\{\cdot|\phi_0,a_0,t_1,\phi_1,\ldots,\phi_n,a_n\}$$

$$= q^{\phi}(\cdot|\phi_n,a_n)$$

and

$$p_{\omega\phi}\{\cdot|\phi_0,a_0,t_1,\phi_1,\ldots,\phi_n,a_n,\phi_{n+1}\} = p_{\pi^{\omega}\phi}\{\cdot|\phi_0,a_0,t_1,\phi_1,\ldots,\phi_n,a_n,\phi_{n+1}\}$$

$$= p^{\phi}(\cdot|\phi_n,a_n,\phi_{n+1}).$$

Then, this policy π^{ω} satisfies (2.3).

Now, from Lemmas 1 and 2, we have the following theorem.

Theorem 2.2. The set of all I-policies is enough, i.e., for any policy ω, there exists an I-policy π which satisfies

$$J_\pi(\phi_0) = J_\omega(\phi_0), \quad \phi_0 \in \Phi.$$

3. The Transformation of SMDP-II to SMDP-I

In this section we shall show that SMDP-II $(S, M, A, p_S, p_F, q, \phi_0, c)$ can be transformed to SMDP-I defined by $(\Phi, A, q^\phi, p^\phi, c^\phi)$. Φ is the set of states of this model. q^ϕ and p^ϕ determine the law of motion. c^ϕ is the immediate cost. We note that we can completely observe the state of this model.

A policy for this model is the same one as I-policy, and is also denoted by $\pi = \{\pi_0, \pi_1, \ldots\}$. Hence, for any I-policy π, the expected average cost function on Φ is defined by

$$I_\pi(\phi_0) = \overline{\lim_{n\to\infty}} \frac{E_\pi\left[\sum_{i=0}^{n-1} c^\phi(\phi_i, a_i) \mid \phi_0\right]}{E_\pi\left[\sum_{i=1}^{n} t_i \mid \phi_0\right]},$$

(3.1)

$$= \overline{\lim_{n\to\infty}} \frac{E_\pi\left[\sum_{i=0}^{n-1} c^\phi(\phi_i, a_i) \mid \phi_0\right]}{E_\pi\left[\sum_{i=0}^{n-1} \tau^\phi(\phi_i, a_i) \mid \phi_0\right]},$$

where E_π denotes the expectation by the conditional distribution

$$p_\pi\{\cdot \mid \phi_0\} = \bigotimes_{n=0}^{\infty} (\pi_n \otimes q^\phi \otimes p^\phi)$$

on the infinite product set $(A \times \Phi \times R_+)^N$.

Remark 1. (3.1) follows from that

$$\tau^\phi(\phi, a) = \int_0^\infty t \, p^\phi(dt \mid \phi, a).$$

Theorem 3.1. SMDP-II $(S, M, A, p_S, p_F, q, \phi_0, c)$ and SMDP-I $(\Phi, A, q^\phi, p^\phi, c^\phi)$ are equivalent in the sense that for any I-policy π,

$$J_\pi(\phi_0) = I_\pi(\phi_0), \quad \phi_0 \in \Phi.$$

Proof: By Theorem 2.1, the conditional distribution $p_{\pi\phi}\{\cdot \mid \phi_0\}$ on $(A \times \Phi \times R_+)^n \times A$ is rewritten as

$$p_{\pi\phi}\{\cdot|\phi_0\} = p_{\pi\phi}\{\{a_0\}|\phi_0\} \otimes p_{\pi\phi}\{\{\phi_1\}|\phi_0,a_0\} \otimes p_{\pi\phi}\{\{t_1\}|\phi_0,a_0,\phi_1\}$$
$$\otimes \cdots \otimes p_{\pi\phi}\{\{a_n\}|\phi_0,a_0,\phi_1,\cdots,\phi_n\}$$
$$= \pi_0 \otimes q^{\phi} \otimes p^{\phi} \otimes \pi_1 \otimes \cdots \otimes \pi_n$$
$$= p_{\pi}\{\cdot|\phi_0\}.$$

Then, we have

$$E_{\pi\phi}[\bar{c}(s_n,a_n)|\phi_0] = E_{\pi\phi}[c^{\phi}(\phi_n,a_n)|\phi_0] = E_{\pi}[c^{\phi}(\phi_n,a_n)|\phi_0]$$

and

$$E_{\pi\phi}[\bar{\tau}(s_n,a_n)|\phi_0] = E_{\pi\phi}[\tau^{\phi}(\phi_n,a_n)|\phi_0] = E_{\pi}[\tau^{\phi}(\phi_n,a_n)|\phi_0].$$

Hence, we have

$$J_{\pi}(\phi_0) = I_{\pi}(\phi_0), \phi_0 \in \Phi.$$

4. The Existence of an Optimal Stationary I-Policy and the Method of Successive Approximations

In this section we shall assume that A is finite, and present some sufficient conditions under which there exists an optimal stationary I-policy, and the method of successive approximations is applicable for obtaining a solution of the optimality equation.

First, we shall show that Condition 1 of SMDP-II (S, M, A, p_s, p_F, q, ϕ_0, c) is inherited to SMDP-I (Φ, A, q^{ϕ}, p^{ϕ}, c^{ϕ}).

Proposition 1. There exists $\delta > 0$, $\varepsilon > 0$ such that for all ϕ and a,

$$p^{\phi}([0,\delta]|\phi,a) = \int_{\Phi} p^{\phi}([0,\delta]|\phi,a,\phi')q^{\phi}(d\phi'|\phi a)$$
$$\leq 1 - \varepsilon.$$

Proof: By Theorem 2.1(iii) and Condition 1,

$$p^{\phi}([0,\delta]|\phi,a) = \sum_{s}\sum_{s'} p_F([0,\delta]|s,a,s')p_s(s'|s,a)q^{p}(s|\phi)$$
$$\leq 1 - \varepsilon.$$

Then, we have the following theorem.

Theorem 4.1 (Theorem 2 of [6]). If there exists a bounded Borel measurable function $h(\phi)$ and a constant g such that

$$(4.1)\qquad h(\phi) = \min_{a}\{c^{\phi}(\phi,a) + \int_{\Phi} h(\phi')q^{\phi}(d\phi'|\phi,a) - g\tau^{\phi}(\phi,a)\},\ \phi\in\Phi,$$

then there exists an optimal stationary I-policy π^*, and π^* is any policy which, for each ϕ, prescribes an action which minimizes the right side of (4.1).

Hence, we shall discuss on the existence of $h(\phi)$ and g of (4.1). The following assumption is imposed.

Assumption 1. There exists a state s^*, an observation signal m^* and positive numbers α and γ ($0<\alpha$, $\gamma\leq 1$) such that

$$p_s(s^*|s,a)\geq\alpha \quad \text{for all } s \text{ and } a$$

and

$$q(m^*|s) = \begin{cases}\gamma & \text{if } s = s^*\\ 0 & \text{otherwise.}\end{cases}$$

Remark 2. The assumption imposed on q implies that m^* is possible only when the system is in s^*. Particularly, if $\gamma = 1$, s^* is completely observable.

Lemma 3. Under Assumption 1,

$$q^{\phi}(\{\delta_{s^*}(\cdot)\}|\phi,a) \geq \alpha\gamma \text{ for all } \phi \text{ and } a,$$

where

$$\delta_{s^*}(s) = \begin{cases}1 & \text{if } s = s^*\\ 0 & \text{otherwise.}\end{cases}$$

Proof: According to (2.1), we have

$$q_{n+1}(\cdot|h_n,a_n,t_{n+1},m^*) = \delta_{s^*}(\cdot) \text{ for any } t_{n+1}\geq 0.$$

By (2.2), when $\Gamma = \{\delta_{s^*}(\cdot)\}$ in Theorem 2.1,

$$\bar{\Gamma} \supset R_+ \times \{m^*\}, \text{ and then, } \bar{\Gamma}_m(m^*) = R_+.$$

Hence, we have

$$\begin{aligned} &q^{\phi}(\{\delta_{s^*}(\cdot)\}|\phi_n,a_n)\\ &= \sum_{s_n}\sum_{s_{n+1}}\sum_{m_{n+1}}\int_{\bar{\Gamma}_m} v(s_n,a_n,s_{n+1},t_{n+1},m_{n+1})\lambda(dt_{n+1})q^{p}(s_n|\phi_n)\\ &\geq \sum_{s_n}\sum_{s_{n+1}} q(m^*|s_{n+1})p_s(s_{n+1}|s_n,a_n)q^{p}(s_n|\phi_n)\\ &= \sum_{s_n} q(m^*|s^*)p_s(s^*|s_n,a_n)q^{p}(s_n|\phi_n) \geq \alpha\gamma. \end{aligned}$$

Furthermore, the following assumption is imposed.

Assumption 2. There exists $\tau_2 > 0$ such that

$$\bar{\tau}(s,a) \leq \tau_2 \quad \text{for all } s \text{ and } a.$$

On the other hand, by Condition 1,

$$\bar{\tau}(s,a) \geq \varepsilon\delta = \tau_1 > 0 \text{ for all } s \text{ and } a.$$

Hence, we have

$$\tau_1 \leq \tau^{\phi}(\phi,a) \leq \tau_2.$$

Then, in a similar way to Schweitzer [9] and White [12], we have the following theorem.

Theorem 4.2. Under Assumptions 1 and 2, a bounded Borel measurable function $h(\phi)$ and a constant g satisfying (4.1) exist, and are obtained by the method of successive approximations.

Proof: Define for any Borel set B of Φ,

$$\tilde{q}(B|\phi,a) = \delta_{\phi}(B) + (q^{\phi}(B|\phi,a) - \delta_{\phi}(B))\frac{\tau_1}{\tau^{\phi}(\phi,a)}$$

where

$$\delta_{\phi}(B) = \begin{cases} 1 & \text{if } \phi \in B \\ 0 & \text{otherwise.} \end{cases}$$

Then, $\tilde{q}$ is a conditional probability. By lemma 3, we can easily see that

$$\tilde{q}(\{\delta_{s^*}(\cdot)\}|\phi,a) \geq \alpha\gamma\left(\frac{\tau_1}{\tau_2}\right).$$

Then, White's iterative scheme:

$$V_n(\phi) = \min_a \left[\frac{c^{\phi}(\phi,a)}{\tau^{\phi}(\phi,a)}\tau_1 + \int_{\Phi} v_n(\phi')\tilde{q}(d\phi'|\phi,a)\right],$$

$$d_n = V_n(\phi^*), \tag{4.2}$$

$$v_{n+1}(\phi) = V_n(\phi) - d_n, \qquad n \geq 1,$$

where $\phi^* = \delta_{s^*}(\cdot)$, converges. We denote the limits of $v_n(\phi)$ and $\tau_1^{-1}d_n$ by $h(\phi)$ and g, respectively. Then, we have

$$h(\phi) + g\tau_1 = \min_a \left[\frac{c^{\phi}(\phi,a)}{\tau^{\phi}(\phi,a)}\tau_1 + \int_{\Phi} h(\phi')\tilde{q}(d\phi'|\phi,a)\right].$$

We replace $\tilde{q}$ with q^{ϕ}. Then, we have

$$g = \min_a \left[\frac{c^\phi(\phi,a) + \int_\Phi h(\phi')q^\phi(d\phi'|\phi,a) - h(\phi)}{\tau^\phi(\phi,a)} \right]$$

or equivalently

$$h(\phi) = \min_a \left[c^\phi(\phi,a) + \int_\Phi h(\phi')q^\phi(d\phi'|\phi,a) - g\tau^\phi(\phi,a) \right].$$

Then, the proof is completed.

By this theorem and Theorem 4.1, under Assumptions 1 and 2, there exists an optimal stationary I-policy, and a solution to (4.1) can be constructed by the method of successive approximations.

Remark 3. In this paper we have treated the case when S and M are countable and A is finite. We can easily extend the results of this paper to the following cases: (i) S and M are Borel sets of a complete separable metric space; A is finite; c and τ are Borel measurable; p_S, p_F and q have densities, (ii) S and M are Borel sets of a complete separable metric space; A is a compact metric space; c^ϕ and τ^ϕ are continuous; p_S, p_F and q have densities; q^ϕ is weakly continuous, i.e., if $(\phi_n,a_n) \to (\phi,a)$, then $q^\phi(\cdot|\phi_n,a_n)$ converges weakly to $q^\phi(\cdot|\phi,a)$.

Acknowledgement

The author is grateful to Professor N. Furukawa of Kyushu University for valuable discussions. The author wishes to thank Professor K. Tanaka of Niigata University for helpful advices.

References

[1] Çinlar, E.: *Introduction to stochastic processes.* Prentice-Hall. 1975.

[2] Hinderer, K.: *Foundations of non-stationary dynamic programming with discrete time parameter.* Springer-Verlag. 1970.

[3] Jewell, W.: Markov renewal programming I and II. *Operations Res.* Vol.2, No.6(1963), 938-971.

[4] Kurano, M.: On the existence of an optimal stationary I-policy in non-discounted Markov decision processes with incomplete state information. *Bull. Math. Statist.* Vol.17, No.3-4(1977), 75-81.

[5] Ross, S. M.: *Applied probability models with optimization applications.* Holden-Day. 1970.

[6] Ross, S. M.: Average cost semi-Markov decision processes, *J. Appl. Prob.* Vol.7(1970), 645-656.

[7] Sawaki, K and Ichikawa, A.: Optimal control for partially observable Markov decision processes over an infinite horizon. *J. of the Operations Res. Soc. of Japan.* Vol.21, No.1(1978), 1-16.

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

[9] Schweitzer, P. J.: Iterative solution of the functional equations of undiscounted Markov renewal programming. *J. Math. Anal. Appl.* Vol.34 (1971), 495-501.

[10] Sondik, E.: The optimal control of partially observable Markov processes over the infinite horizon; Discounted costs. *Operations Res.* Vol.26, No.2(1978), 282-304.

[11] White, C. C.: Procedure for the solutions of a finite-horizon, partially observed, semi-Markov optimization problem. *Operations Res.* Vol.24(1976), 348-358.

[12] White, D. J.: Dynamic programming, Markov chains, and the method of successive approximations. *J. Math. Anal. Appl.* Vol.6(1963), 373-376.

Kazuyoshi WAKUTA: Nagaoka Technical College, Nagaoka-shi, Niigata-ken 940, Japan.