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

Kazuyoshi Wakuta
Nagaoka Technical College

(Received April 20, 1981; Final July 2, 1982)

Abstract In this paper, we study semi-Markov decision processes with the incomplete state observation under the discounted cost criterion. We show that this model can be transformed to ordinary semi-Markov decision processes. Furthermore, we show that if the action space is countable, the optimal value function is Borel measurable and satisfies the optimality equation and that if the action space is finite, there exists an optimal stationary l-policy.

1. Introduction

Markov decision processes with the incomplete state observation have been investigated by many authors, for example, [5], [7], [8] and [9]. 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 a semi-Markov decision process with the incomplete state observation (SMDP-II). A semi-Markov decision process with the complete state observation (SMDP-I), i.e., the ordinary semi-Markov decision process was introduced by Jewell [4] and has been studied by several authors, for example, Ross [6]. SMDP-II was first formulated as a partially observed semi-Markov optimization problem by White [12], where the finite planning horizon, finite state, discrete time case was analyzed. Wakuta [11] has studied the infinite planning horizon, countable state, continuous time case under the average cost criterion. We shall consider here the discounted cost case and show that our model can be transformed to the ordinary semi-Markov decision process, where the states are probabilities on the set of the states of the original model.

First we give the notations and definitions. Let X and Y be nonempty
Borel sets (Borel set means a Borel subset of a complete separable metric space). A probability on $X$ is a probability measure defined on Borel subsets of $X$, and the set of all probabilities on $X$ is denoted by $P(X)$. A conditional probability on $Y$ given $X$ is a function $q(\cdot|x)$ such that for each $x \in X$, $q(\cdot|x)$ is a probability on $Y$ and for each Borel subset of $Y$, $q(B|\cdot)$ is a Borel measurable function on $X$. $Q(Y|X)$ is the set of all conditional probabilities on $Y$ given $X$. We denote the Cartesian product of $X$ and $Y$ by $XY$. For any $p \in P(X)$ and $q \in Q(Y|X)$, $p \otimes q$ is a unique probability on $XY$ such that for all Borel subsets $A$ and $B$,

$$p \otimes q(AB) = \int_A q(B|x) dp(x).$$

Specially, if $q(\cdot|x)$ is degenerate for each $x$, i.e., if there is a Borel measurable function $f$ from $X$ to $Y$ such that $q(\cdot|x) = I_f(x)(\cdot)$, then $p \otimes q$ is also denoted by $p \otimes f$. Every probability $m \in P(XY)$ has a factorization $m = p \otimes q$ where $p \in P(X)$ and $q \in Q(Y|X)$. This notation extends to a finite or infinite sequence of Borel sets $X_1, X_2, \ldots$ (cf. Hinderer [3], Appendix 3). $M(X)$ is the set of all bounded Borel measurable functions on $X$.

SMDP-II is specified by $(S, M, A, Ps', Pt', q, s_0, c, a)$. $S$ 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, the set of actions. $Ps$ is an element of $Q(S|SA)$. $Pt$ is an element of $Q(R_+|SAS)$ where $R_+ = [0, +\infty)$. $Pt$ has a density $f(t|s,a,s')$ with respect to some $\sigma$-finite measure $\lambda$ (for example, Lebesgue measure or counting measure if the process is discrete time like [12]), where $f$ is a Borel measurable function of $(t,s,a,s')$. $Ps$ and $Pt$ are the laws of the motion of the system. $q$ is an element of $Q(M|S)$, the characteristic of the measuring system. $s_0$ is an element of $\Phi(M|S)$, the initial distribution of the system. $c$ is an element of $M(R_+SA)$, the cost function. $a$ is a discount factor ($a > 0$).

Now, we shall give a brief description of SMDP-II. The initial distribution $s_0$ of $S_0$ is given at time 0. An action $a_0$ must be chosen. If the system is in state $s_n$ at time $T_n = t_1 + \ldots + t_n$, where each $t_n$ is the time interval of the $n$-th transition, and action $a_n$ is chosen, then

(i) the next state of the system is chosen according to $Ps(\cdot|s_n,a_n)$;

(ii) conditional on the event that the next state is $s_{n+1}$, the time until the transition from $s_n$ to $s_{n+1}$ occurs is a random variable.
Semi-Markov Decision Processes

with a conditional probability $p_t(\cdot|s_n, a_n, s_{n+1})$;

(iii) state $s_n$ cannot be directly observed, but the observation signal $m_n$ generated according to $q(\cdot|s_n)$ is given;

(iv) the cost incurred during $[\tau_n, \tau_{n+1}]$ is given by
\[ e^{-at} \int_0^{T_{n+1} - \tau_n} c(t, s_n, a_n) e^{-at} dt. \]

After the transition occurs, an action is again chosen and (i), (ii), (iii) and (iv) are repeated.

We note that we can observe the time $t$ when the transition occurs and that the times of the process transition, observation, and control reset occur simultaneously.

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

Condition 1 ([6]). There exist $\delta > 0$ and $\varepsilon > 0$ such that for all $s$ and $a$,
\[ \sum_{s', t^*} p_t([0, \delta]|s, a, s')p_s(s'|s, a) \leq 1 - \varepsilon. \]

To select actions, a policy is needed. A policy $\omega$ is a sequence $(\omega_0, \omega_1, \ldots)$, where $\omega_n$ is an element of $Q(A|H_n)$, where $H_n = \phi(AR^+_M)^n$. $H_n$ is the set of all observable histories up to the $n$-th stage. $\phi$ is metrizable by introducing the distance
\[ d(\phi, \phi') = \sum_{s \in S} |\phi(s) - \phi'(s)|, \phi, \phi' \in \phi. \]

The topology introduced by this metric is equivalent to the weak topology (cf. Billingsley [1], Appendix II, Scheffe's Theorem). Then, $\phi$ is a complete separable metric space by introducing the discrete topology on $S$, and the Borel $\sigma$-algebra of $\phi$ is identical with the $\ast$-$\sigma$-algebra of $\phi$. Then, we define an element $q^P \in Q(S|\phi)$ by
\[ q^P(s|\phi) = \phi(s), s \in S. \]

Hence, any policy $\omega$, together with $q^P, p^P$, $p^T$ and $q$ defines an element
\[ p_\omega \in Q(S(AR^+_M)^N|\phi) \] where $S(AR^+_M)^N$ is the set of all futures of the system ($N$ denotes the set of all natural numbers), i.e., it defines
\[ p_\omega(\cdot|\phi) = q^P \otimes \otimes_{n=0}^{\infty} (\omega_n \otimes p^P \otimes p^T \otimes q). \]
When a policy $\omega$ is applied, the expected total discounted cost function on $\phi$ is defined by

$$J_\omega^\alpha(\phi_0) = E_\omega \left[ \sum_{n=0}^{\infty} e^{-\alpha T} \int_0^{T_n+1} c(t, s_n, a_n) e^{-\alpha t} dt | \phi_0 \right],$$

where $E_\omega[\cdot | \phi_0]$ denotes the conditional expectation by $p_\omega(\cdot | \phi_0)$. $p_\omega(\cdot | \phi_0)$ is defined by putting the initial distribution $\phi_0$ in (1.1). Then our optimization problem is to minimize $J_\omega^\alpha(\phi_0)$ among all policies. We say that $\omega^*$ is optimal if $J_\omega^\alpha(\phi_0) \leq \inf_{\omega} J_\omega^\alpha(\phi_0)$ for all $\phi_0$.

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.

We denote the conditional probability of $s_n$ given the observable history $h_n$ by $q_n(\cdot | h_n)$. Using the Bayesian formula, we obtain the following relation of $q_n$:

$$q_{n+1}(s_{n+1} | h_{n+1}) = q_{n+1}(s_{n+1} | h_n, a_n, t_{n+1}, m_{n+1})$$

$$= \frac{\sum_{s_n} \nu(s_n, a_n, s_{n+1}, t_{n+1}, m_{n+1}) q_n(s_n | h_n)}{\sum_{s_n} \nu(s_n, a_n, s_{n+1}, t_{n+1}, m_{n+1}) q_n(s_n | h_n)},$$

(2.1)

where $\nu = p_s(s_{n+1} | s_n, a_n)f(t_{n+1} | s_n, a_n, s_{n+1})q(m_{n+1} | s_{n+1})$.

Letting $q_n$ correspond to an element $\phi_n \in \phi$ by

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

we see that the right-hand 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 $u: \Phi_{AR,A'} \rightarrow \phi$ defined by

$$\phi_{n+1}(s_{n+1}) = \frac{\sum_{s_n} \nu(s_n, a_n, s_{n+1}, t_{n+1}, m_{n+1}) \phi_n(s_n)}{\sum_{s_n} \nu(s_n, a_n, s_{n+1}, t_{n+1}, m_{n+1}) \phi_n(s_n)}$$

(2.2)

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

(cf, Hinderer [3], Remarks, p.85).
By repeated use of $u$, corresponding to any observable history $h_n$, $b_n = (\phi_0, a_0, t_1, \phi_1, \ldots, a_{n-1}, t_n, \phi_n) \in B_n$ is determined Borel measurable, where $B_n = \phi(\mathbb{R}_+ \phi)^N$. $B_n$ is the set of the possible information concerning the histories of the system. Then, we define a new policy $\pi$ which depends only on the possible information. We call this policy as information policy (I-policy) according to [8]. An I-policy $\pi$ is a sequence $\{\pi_0, \pi_1, \ldots\}$, where each $\pi_n$ is an element of $Q(A|B_n)$. An I-policy $\pi$ is said to be stationary if there exists a Borel measurable map $f: \phi \to A$ such that $\pi_n(f(\phi)|\phi_0, a_0, t_1, \phi_1, \ldots, a_{n-1}, t_n, \phi_n) = 1$ for all $\phi_n$. For any I-policy $\pi$, we define a policy $\omega^\pi = (\omega_0^\pi, \omega_1^\pi, \ldots)$ by $\omega_n^\pi(\cdot|b_n) = \pi_n(\cdot|b_n^h)$, where $b_n^h$ is an element of $B_n$ which corresponds to $h_n \in H_n$. Then, $\pi$ and $\omega^\pi$ 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 $\pi$, together $q^P$, $p_s$, $p_t$, $q$ and $u$, defines an element $\pi_0^\pi$ of $Q(S(\mathbb{R}_+ \mathbb{R})^N|\phi)$, i.e., it defines

$$(2.3) \quad \pi_0^\pi(\cdot|\phi) = q^P \otimes \otimes_{n=0}^{\infty} (\otimes_n) \otimes p_s \otimes p_t \otimes q \otimes u).$$

For any I-policy $\pi$, the expected total discounted cost function on $\phi$ is defined by

$$J^{\pi}_0(\phi) = \bar{E}_n \left[ \sum_{n=0}^{\infty} e^{-\alpha t_n} \int_0^{T_n+1} e^{-\alpha t} c(t, s_n, a_n) e^{-\alpha t} dt | \phi_0 \right],$$

where $\bar{E}_n[\cdot|\phi_0]$ denotes the conditional expectation by $\pi_0^\pi(\cdot|\phi_0)$. $\bar{E}_n[\cdot|\phi_0]$ is defined by putting the initial distribution $\phi_0$ in (2.3). We also define $\bar{E}_\omega$ and $\bar{E}_\omega$ in the same way for policy $\omega$. Then, $J^{\pi}_0(\phi_0)$ can be rewritten by $\bar{E}_\omega$ in place of $E_\omega$.

Let for all $s$ and $a$,

$$(2.4) \quad c_{\alpha}(t, s, a) = \int_0^{t} c(t, s, a) e^{-\alpha t} dt,$$

and

$$(2.5) \quad \bar{c}_{\alpha}(\phi, a) = \sum_{s} \int_0^{\infty} c_{\alpha}(t, s, a) d\rho_{ts}(t|s, a) \rho_{ts}(s'|s, a) \phi(s).$$

Proposition 1.

$$J^{\pi}_0(\phi) = \bar{E}_\omega \left[ \sum_{n=0}^{\infty} e^{-\alpha t_n} \bar{c}_{\alpha}(\phi_n, a_n) | \phi_0 \right].$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.
Proof.

\[ J^\alpha_0(\phi_0) = \int_\omega \left[ \sum_{n=0}^{\infty} e^{-\alpha T_n} \int_0^{T_{n+1}} c(t, s_n, a_n) e^{-\alpha t} dt \right] d\mu_{\phi_0} \]

\[ = \sum_{n=0}^{\infty} \int_\omega e^{-\alpha T_n} \left[ \int_0^{T_{n+1}} c(t, s_n, a_n) e^{-\alpha t} dt \right] d\mu_{\phi_0} \]

where \( h'_n = (\phi_0, a_0, t_1, m_1, \phi_1, \ldots, \phi_n, a_n) \).

Then

\[ \int_\omega \left[ c(\pi_{n+1}, s_n, a_n) | h'_n \right] d\mu_{\phi_0} \]

\[ = \sum_{n=0}^{\infty} \int_\omega \left[ \int_0^{T_{n+1}} c(\pi_{n+1}, s_n, a_n) e^{-\alpha t} dt \right] d\mu_{\phi_0} \]

\[ = \sum_{n=0}^{\infty} \int_\omega \left[ \int_0^{T_{n+1}} c(\pi_{n+1}, s_n, a_n) e^{-\alpha t} dt \right] d\mu_{\phi_0} \]

\[ = \int_\omega \left[ \int_0^{T_{n+1}} c(\pi_{n+1}, s_n, a_n) e^{-\alpha t} dt \right] d\mu_{\phi_0} \]

Hence, we have

\[ J^\alpha_0(\phi_0) = \int_\omega \left[ \sum_{n=0}^{\infty} e^{-\alpha T_n} c(\pi_{n+1}, a_n) | \phi_0 \right] d\mu_{\phi_0} \]

Remark 1. Proposition 1 is also true for I-policy \( \pi \).

The following theorem is an extension of one that was pointed out in [12], and is basic in the later sections.

**Theorem 2.1** ([11]). For any fixed sequence of actions \( \{a_0, a_1, \ldots, a_n\} \),

where \( a_n \) is the action to be chosen during the \( n \)-th transition interval, (i) the stochastic process \( (\pi, t) = (\pi_n, t_n; n \in N) \) is a Markov renewal process (see Cinlar [2] for definition), (ii) given that the process has just entered \( \pi_n \), the probability that the next state \( \pi_{n+1} \) will be into \( \Gamma \) depends only on \( \pi_n \) and \( a_n \), and is given by

\[ \overline{q}(\Gamma | \pi_n, m_{n+1}) = \sum_{s_n, s_{n+1}} \sum_{m=0}^{\infty} \int_{\Gamma_m} v(s_n, a_n, s_{n+1}, t_{n+1}, m_{n+1}) \]

\[ \times d\lambda(t_{n+1}) \psi_n(s_n), \]

where for any Borel subset \( \Gamma \) of \( \psi \),

\[ \overline{\Gamma} = \overline{\Gamma}(\pi_n, a_n; \Gamma) = \{t_{n+1}, m_{n+1} \in \Gamma \}, \]

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.
\[ \bar{T}_m = \bar{T}_m(\phi_n, a_n, m_{n+1}; \Gamma) = \{ t_{n+1}; (t_{n+1}, m_{n+1}) \in \bar{T} \}. \]

(iii) conditional on the event that the next state is \( \phi_{n+1} \), the time until the transition from \( \phi_n \) to \( \phi_{n+1} \) occurs depends only on \( \phi_n, a_n \) and \( \phi_{n+1} \). Its conditional probability \( \bar{p}(\cdot|\phi_n, a_n, \phi_{n+1}) \) satisfies for any Borel subset \( B \) of \( \mathbb{R}_+ \),

\[
\int_{\phi_n} \bar{p}(B|\phi_n, a_n, \phi_{n+1}) \, d\bar{q}(\phi_{n+1}|\phi_n, a_n) = \sum_{s_n} \sum_{s_{n+1}} p(t|s_n, a_n, s_{n+1}) p_{s_n} (s_{n+1} | s_n, a_n) \phi_n(s_n).
\]

Theorem 2.2. The set of all I-policies is enough, i.e. for any policy \( \omega \), there exists an I-policy \( \pi \) which satisfies

\[ J^\alpha_\pi(\phi_0) = J^\alpha_\omega(\phi_0), \phi_0 \in \phi. \]

Proof. Using Proposition and Theorem 2.1, this theorem can be proved in a similar way to Lemma 2 of [11].

3. The Transformation of SMDP-II to SMDP-I and the Existence of an Optimal Stationary I-Policy

In this section we shall show that SMDP-II \( (S, M, A, p_s, p_t, q, \phi_0, c, \alpha) \) can be transformed to SMDP-I specified by \( (\phi, A, \bar{q}, \bar{p}, \bar{c}_\alpha, \alpha) \). \( \phi \) is the set of states of a new model. \( \bar{q} \) and \( \bar{p} \) are defined in Theorem 2.1, and \( \bar{c}_\alpha \) is defined in (2.5). \( \bar{q} \) and \( \bar{p} \) are the laws of motion, and \( \bar{c}_\alpha \) 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 \( \pi \), the expected total discounted cost function on \( \phi \) is defined by

\[ I^\alpha_\pi(\phi_0) = E^\pi \left[ \sum_{n=0}^\infty e^{-\alpha T_n} \bar{c}_\alpha(\phi_n, a_n) \mid \phi_0 \right], \]

where \( E^\pi[\cdot|\phi_0] \) denotes the conditional expectation by \( p^\pi(\cdot|\phi_0) \). \( p^\pi(\cdot|\phi_0) \) is defined by putting the initial distribution \( \phi_0 \) in the following

\[
p^\pi(\cdot|\phi_0) = \bigotimes_{n=0}^\infty (\pi_n \otimes \bar{q} \otimes \bar{p}).
\]

Theorem 3.1. SMDP-II \( (S, M, A, p_s, p_t, q, \phi_0, c, \alpha) \) and SMDP-I \( (\phi, A, \bar{q}, \bar{p}, \bar{c}_\alpha, \alpha) \) are equivalent in the sense that for any I-policy \( \pi \),

\[ J^\alpha_\pi(\phi_0) = I^\alpha_\pi(\phi_0), \phi_0 \in \phi. \]
Proof. Using Remark 1 after Proposition 1 and Theorem 2.1, this theorem can be proved in a similar way to Theorem 3.1 of [11].

As [11] showed in Proposition 1, Condition 1 of SMMD-II ($S, M, A, P_s, P_t, q, \phi_0, c, a$) is inherited to SMMD-I ($\phi, A, \tilde{q}, \tilde{P}, \tilde{c}, \tilde{a}$). Hence, we have the following proposition.

Proposition 2. There exist $\delta > 0$ and $\epsilon > 0$ such that all $\phi$ and $\tilde{a}$,

$$\int_{\phi} p([0, \delta] | \phi, a, \tilde{a}') d\tilde{q}(\tilde{a}' | \phi, a) \leq 1 - \epsilon.$$

Next, we shall state the main results.

Let

$$I^\alpha(\phi) = \inf_{\pi} I^\alpha_{\pi}(\phi), \phi \in V.$$

Theorem 3.2. Suppose that $A$ is countable. Then, $I^\alpha(\phi)$ is a Borel measurable function on $\phi$, and is the unique solution to

$$I^\alpha(\phi) = \inf_{\alpha \in A} \{ \tilde{c}^\alpha(\phi, a) + \int_{\phi} \int_0^\infty e^{-at} I^\alpha_{\alpha}(\phi') \times dp(t | \phi, a, \tilde{a}') d\tilde{q}(\tilde{a}' | \phi, a), \phi \in V,$$

which is equivalent to

$$I^\alpha(\phi) = \inf_{\alpha \in A} \{ \tilde{c}^\alpha(\phi, a) + \sum_{S} \sum_{S'} \int_0^\infty e^{-at} I^\alpha_{\alpha}(u(\phi, a, t, m')) \times dp(t | s, a, s') dp_{\tilde{P}}(s' | s, a) q(m' | s') \phi(s), \phi \in V.$$

If the infimum in (3.2) (or (3.3)) is achieved, any stationary I-policy $f^\alpha_A$ which in each state selects the action which minimizes the right-hand side of (3.2) (or (3.3)) is optimal.

Proof. First we shall translate Strauch's results [10] for Markov decision processes into the case of semi-Markov decision processes. Let $X = (A \times R^+) \times \mu$ and factorize a probability measure $\nu \in P(x)$ like Lemma 7.2 of [10].

Then, Lemma 7.2 also holds in the case of SMMD-I. Hence, applying Lemma 7.1 and Theorem 7.1 of [10] to the case of SMMD-I, we can easily prove that $I^\alpha(\phi)$ is universally measurable. Also, by the first half of Theorem 8.1 of [10], there exists a $(p, \epsilon)$-optimal I-policy for SMMD-I. Using these results we shall show that $I^\alpha(\phi)$ is a solution to (3.2). Let

$$T^\nu_{\alpha}(\phi) = \tilde{c}^\alpha(\phi, a) + \int_{\phi} \int_0^\infty e^{-at} \nu(\phi') \times dp(t | \phi, a, \tilde{a}') d\tilde{q}(\tilde{a}' | \phi, a).$$

We note that because of Condition 1,

$$\int_0^\infty e^{-at} dp(t | \phi, a, \tilde{a}') d\tilde{q}(\tilde{a}' | \phi, a) \leq 1 - \epsilon + \epsilon e^{-\alpha \delta} < 1.$$
At first, we have

$$\pi^a(\phi_0) = E \left[ \sum_{n=0}^{\infty} e^{-\alpha T_n} \zeta_n(a_n, \phi_n) | \phi_0 \right],$$


$$\geq E \left[ \sum_{n=0}^{\infty} \eta_{a_0} T_{a_0} \pi^a(\phi_0), \right),$$

where $\eta_{a_0} = \eta(0, \phi_0), a_0 \in \Lambda$. Hence,

$$\pi^a(\phi_0) \geq \inf_{a_0 \in \Lambda} T_{a_0} \pi^a(\phi_0).$$

To get the other way, let $\pi$ be the policy that chooses $a_0$ at time 0 and follows $\pi'$ after the first transition occurs, where $\pi'$ is $(p, \epsilon)$-optimal I-policy with respect to $p = \tilde{q}(\cdot | \phi_0, a_0)$. Then,

$$\pi^a(\phi_0) = T_{a_0} \pi^a(\phi_0) \quad \leq \quad T_{a_0} (\pi^a + \epsilon)(\phi_0) \quad \leq \quad T_{a_0} \pi^a(\phi_0) + \epsilon (1 - \epsilon + \epsilon e^{-\alpha \delta}).$$

Hence,

$$\pi^a(\phi_0) \leq \inf_{a_0 \in \Lambda} T_{a_0} \pi^a(\phi_0) + \epsilon (1 - \epsilon + \epsilon e^{-\alpha \delta}).$$

Since $\epsilon$ is arbitrary,

$$\pi^a(\phi_0) \leq \inf_{a_0 \in \Lambda} T_{a_0} \pi^a(\phi_0).$$

Thus, $\pi^a(\phi)$ is a solution to (3.2).

Next, we shall prove that $\pi^a(\phi)$ is Borel measurable. By the proof for the discounted case of Theorem 8.2 of [10], there does not exist another bounded solution to (3.2). Hence, $\pi^a(\phi)$ is the unique solution among the bounded functions. On the other hand, we note that

$$\| \inf_{a_0} T_{a_0} u - \inf_{a_0} T_{a_0} v \| \leq (1 - \epsilon + \epsilon e^{-\alpha \delta}) \| u - v \|,$$

where $\| \cdot \|$ denotes the supremum norm. Then, $\sup_{a_0} T_{a_0}$ has the unique fixed point, i.e., there exists the unique solution to (3.2) among the bounded Borel measurable functions. Hence, $\pi^a(\phi)$ must be Borel measurable.

Next, we shall prove the second half of the theorem. Let for stationary policy $f$

$$T_{f^\phi} (\phi) = T_{f(\phi)}^\phi (\phi), \phi \in \phi.$$
By the definition of $f_{\alpha}$,

$$I_{\alpha}^{\phi}(\psi) = T_{f_{\alpha}} I_{\alpha}^{\phi}(\psi).$$

On the other hand $I_{f_{\alpha}}^{\phi}(\psi)$ is the unique solution to

$$v(\psi) = T_{f_{\alpha}} v(\psi).$$

Hence,

$$I_{f_{\alpha}}^{\phi}(\psi) = I^{\phi}(\psi), \; \phi \in \Phi.$$

Therefore, $f_{\alpha}$ is optimal.

Finally we shall show that (3.2) and (3.3) are equivalent. Let $g(\phi, t)$ be Borel measurable in $(\phi, t)$. Then by the proof of Theorem 2.1 (i) of [11],

$$\int \int_{0}^{\infty} g(\phi', t) d\tilde{p}(t|\phi, a, \phi') d\tilde{q}(\psi'|\phi, a)$$

$$= \int \int_{0}^{\infty} g(\phi', t) d\tilde{p} \otimes \tilde{q}(\psi', t|\phi, a)$$

$$= \sum \sum \int_{0}^{\infty} g(u(\phi, a, t, m'), t) dp_{t}(t|s, a, s')$$

$$\times p_{s}(s'|s, a) q(m'|s') \psi(s),$$

which shows that (3.2) is equivalent to (3.3).

Acknowledgement

The author wishes to express his hearty thanks to Professor N. Furukawa of Kyushu University for many valuable discussions. The author is grateful to Professor K. Tanaka of Niigata University for helpful advices. The author is also thankful to referees for their helpful comments.

References


non-discounted Markov decision process with incomplete state information. 


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