Journal of the Operations Research
Society of Japan
Vol. 25, No. 4, December 1982

© 1982 The Operations Research Society of Japan

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 I-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 $q(\cdot|\cdot)$ 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 $q(\cdot|x) = 1_{f(x)}(\cdot)$, 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 $(S, M, A, p_s, p_t, q, \phi_0, c, \alpha)$. 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. p_s is an element of $Q(S|SA)$. p_t is an element of $Q(R_+|SAS)$ where $R_+ = [0, +\infty)$. p_t has a density $f(t|s,a,s')$ with respect to some σ - finite measure λ (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'). p_s and p_t are the laws of the motion of the system. q is an element of $Q(M|S)$, the characteristic of the measuring system. ϕ_0 is an element of Φ, where $\Phi = P(S)$, the initial distribution of the system. c is an element of $M(R_+SA)$, the cost function. α is a discount factor ($\alpha > 0$).

Now, we shall give a brief description of SMDP-II. The initial distribution ϕ_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 $p_s(\cdot|s_n,a_n)$;

(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

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

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 $[T_n, T_{n+1}]$ is given by

$$e^{-\alpha T_n}\int_0^{T_{n+1}-T_n} c(t,s_n,a_n)e^{-\alpha t}dt.$$

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,

$$\sum_{s'} p_t([0,\delta]|s,a,s')p_s(s'|s,a) \leqq 1 - \varepsilon.$$

To select actions, a policy is needed. A policy ω is a sequence $\{\omega_0, \omega_1, \ldots\}$, where ω_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. Φ is metrizable by introducing the distance

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

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

$$q^p(s|\phi) = \phi(s), \quad s \in S.$$

Hence, any policy ω, together with q^p, p_s, p_t and q defines an element $p_\omega \in Q(S(ASR_+M)^N \mid \Phi)$ where $S(ASR_+M)^N$ is the set of all futures of the system (N denotes the set of all natural numbers), i.e., it defines

$$p_\omega\{\cdot|\phi\} = q^p \otimes \bigotimes_{n=0}^{\infty} (\omega_n \otimes p_s \otimes p_t \otimes q). \tag{1.1}$$

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

When a policy ω is applied, the expected total discounted cost function on Φ is defined by

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

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 ϕ_0 in (1.1). Then our optimization problem is to minimize $J_\omega^\alpha(\phi_0)$ among all policies. We say that ω^* is optimal if $J_{\omega^*}^\alpha(\phi_0) \leqq \inf_\omega J_\omega^\alpha(\phi_0)$ 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.

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 : for $s_{n+1} \in S$,

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

$$= \frac{\sum_{s_n} v(s_n, a_n, s_{n+1}, t_{n+1}, m_{n+1})q_n(s_n|h_n)}{\sum_{s_n}\sum_{s'_{n+1}} v(s_n, a_n, s'_{n+1}, t_{n+1}, m_{n+1})q_n(s_n|h_n)}, \tag{2.1}$$

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

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 $u: \Phi AR_+M \to \Phi$ defined by

$$\phi_{n+1}(s_{n+1}) = \frac{\sum_{s_n} v(s_n, a_n, s_{n+1}, t_{n+1}, m_{n+1})\phi_n(s_n)}{\sum_{s_n}\sum_{s'_{n+1}} v(s_n, a_n, s'_{n+1}, t_{n+1}, m_{n+1})\phi_n(s_n)} \tag{2.2}$$

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

(cf, Hinderer [3] , Remarks, p.85).

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

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 measurably, where $B_n = \Phi(AR_+\Phi)^n$. 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 [8]. An I-policy π is a sequence $\{\pi_0, \pi_1, \ldots\}$, where each π_n is an element of $Q(A|B_n)$. An I-policy π is said to be stationary if there exists a Borel measurable map $f: \Phi \to A$ such that $\pi_n(f(\phi_n)|\phi_0, a_0, t_1, \phi_1, \ldots, a_{n-1}, t_n, \phi_n) = 1$ for all ϕ_n. For any I-policy π, we define a policy $\omega^\pi = \{\omega_0^\pi, \omega_1^\pi, \ldots\}$ 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 q^p, p_s, p_t, q and u, defines an element $\bar{p}_\pi \in Q(S(ASR_+M\Phi)^N|\Phi)$, i.e., it defines

$$\bar{p}_\pi\{\cdot|\phi\} = q^p \otimes \bigotimes_{n=0}^{\infty} (\pi_n \otimes p_s \otimes p_t \otimes q \otimes u). \tag{2.3}$$

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

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

where $\bar{E}_\pi[\cdot|\phi_0]$ denotes the conditional expectation by $\bar{p}_\pi\{\cdot|\phi_0\}$. $\bar{p}_\pi\{\cdot|\phi_0\}$ is defined by putting the initial distribution ϕ_0 in (2.3). We also define $\bar{p}_\omega$ and $\bar{E}_\omega$ in the same way for policy ω. Then, $J_\omega^\alpha(\phi_0)$ can be rewritten by $\bar{E}_\omega$ in place of E_ω.

Let for all s and a,

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

and

$$\bar{c}_\alpha(\phi, a) = \sum_s \sum_{s'} \int_0^\infty c_\alpha(t, s, a) dp_t(t|s, a, s') p_s(s'|s, a) \phi(s) \tag{2.5}$$

Proposition 1.

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

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

Proof.

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

$$= \sum_{n=0}^{\infty} \bar{E}_\omega\left[\bar{E}_\omega\left[e^{-\alpha T_n}\int_0^{t_{n+1}} c(t, s_n, a_n)\, e^{-\alpha t}dt\,|h_n'\right]|\phi_0\right]$$

$$= \sum_{n=0}^{\infty} \bar{E}_\omega\left[e^{-\alpha T_n}\,\bar{E}_\omega\left[c_\alpha(t_{n+1}, s_n, a_n)|h_n'\right]|\phi_0\right],$$

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

Then

$$\bar{E}_\omega\left[c_\alpha(t_{n+1}, s_n, a_n)|h_n'\right]$$

$$= \sum_{s_n}\sum_{s_{n+1}} \int_0^\infty c_\alpha(t, s_n, a_n)dp_t(t|s_n, a_n, s_{n+1})p_s(s_{n+1}|s_n, a_n)q_n(s_n|h_n')$$

$$= \sum_{s_n}\sum_{s_{n+1}} \int_0^\infty c_\alpha(t, s_n, a_n)dp_t(t|s_n, a_n, s_{n+1})p_s(s_{n+1}|s_n, a_n)\phi_n(s_n)$$

$$= \bar{c}_\alpha(\phi_n, a_n).$$

Hence, we have

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

Remark 1. Proposition 1 is also true for I-policy π.

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 $\{a_0, a_1, \ldots\}$, where a_n is the action to be chosen during the n-th transition interval, (i) the stochastic process $(\phi, t) = \{\phi_n, t_n; n \in N\}$ is a Markov renewal process (see Çinlar [2] for definition), (ii) given that the process has just entered ϕ_n, the probability that the next state ϕ_{n+1} will be into Γ depends only on ϕ_n and a_n, and is given by

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

$$\times\, d\lambda(t_{n+1})\phi_n(s_n),$$

where for any Borel subset Γ of Φ,

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

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

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

(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 depends only on ϕ_n, a_n and ϕ_{n+1}. Its conditional probability $\bar{p}(\cdot|\phi_n, a_n, \phi_{n+1})$ satisfies for any Borel subset B of R_+,

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

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

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

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

Proof. Using Proposition 1 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 $(\Phi, A, \bar{q}, \bar{p}, \bar{c}_\alpha, \alpha)$. Φ 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 $\pi = \{\pi_0, \pi_1, ..\}$. Hence, for any I-policy π, the expected total discounted cost function on Φ is defined by

$$I_\pi^\alpha(\phi_0) = E_\pi\left[\sum_{n=0}^{\infty} e^{-\alpha T_n} \bar{c}_\alpha(\phi_n, a_n) | \phi_0\right],$$

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 ϕ_0 in the following

$$p_\pi\{\cdot|\phi\} = \bigotimes_{n=0}^{\infty} (\pi_n \otimes \bar{q} \otimes \bar{p}). \tag{3.1}$$

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 π,

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

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

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 SMDP-II (S, M, A, p_s, p_t, q, ϕ_0, c, α) is inherited to SMDP-I (Φ, A, $\bar{q}$, $\bar{p}$, $\bar{c}_\alpha$, α). Hence, we have the following proposition.

Proposition 2. There exist $\delta > 0$ and $\varepsilon > 0$ such that all ϕ and a,

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

Next, we shall state the main results.

Let

$$I^\alpha(\phi) = \inf_\pi I_\pi^\alpha(\phi), \quad \phi \in \Phi.$$

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

$$(3.2) \qquad I^\alpha(\phi) = \inf_{a\in A} \{\bar{c}_\alpha(\phi,a) + \int_\Phi\int_0^\infty e^{-\alpha t} I^\alpha(\phi') \times d\bar{p}(t|\phi,a,\phi')d\bar{q}(\phi'|\phi,a)\}, \quad \phi \in \Phi,$$

which is equivalent to

$$(3.3) \qquad I^\alpha(\phi) = \inf_{a\in A} \{\bar{c}_\alpha(\phi,a) + \sum_s \sum_{s'} \sum_{m'} \int_0^\infty e^{-\alpha t} I^\alpha(u(\phi,a,t,m')) \times dp_t(t|s,a,s')p_s(s'|s,a)q(m'|s')\phi(s)\}, \quad \phi \in \Phi.$$

If the infimum in (3.2) (or (3.3)) is achieved, any stationary I-policy f_α 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 $X = (A\Phi R_+)^N$ 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 SMDP-I. Hence, applying Lemma 7.1 and Theorem 7.1 of [10] to the case of SMDP-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,ε)-optimal I-policy for SMDP-I. Using these results we shall show that $I^\alpha(\phi)$ is a solution to (3.2). Let

$$T_a v(\phi) = \bar{c}_\alpha(\phi,a) + \int_\Phi\int_0^\infty e^{-\alpha t} v(\phi') \times d\bar{p}(t|\phi,a,\phi')d\bar{q}(\phi'|\phi,a).$$

We note that because of Condition 1,

$$\int_\Phi\int_0^\infty e^{-\alpha t} d\bar{p}(t|\phi,a,\phi')d\bar{q}(\phi'|\phi,a) \leq 1 - \varepsilon + \varepsilon e^{-\alpha\delta} < 1.$$

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

At first, we have

$$I_\pi^\alpha(\phi_0) = E_\pi\left[\sum_{n=0}^{\infty} e^{-\alpha T_n}\bar{c}_\alpha(\phi_n, a_n)\,|\,\phi_0\right],$$

$$\geqq \sum_{a_0\in A} p_{a_0} T_{a_0} I^\alpha(\phi_0),$$

where $p_{a_0} = \pi_0(a_0|\phi_0)$, $a_0 \in A$. Hence,

$$I^\alpha(\phi_0) \geqq \inf_{a_0\in A} T_{a_0} I^\alpha(\phi_0).$$

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

$$I_\pi^\alpha(\phi_0) = T_{a_0} I_\pi^\alpha,(\phi_0)$$

$$\leqq T_{a_0} (I^\alpha+\varepsilon)(\phi_0)$$

$$\leqq T_{a_0} I^\alpha(\phi_0) + \varepsilon(1 - \varepsilon + \varepsilon e^{-\alpha\delta}).$$

Hence,

$$I^\alpha(\phi_0) \leqq \inf_{a_0\in A} T_{a_0} I^\alpha(\phi_0) + \varepsilon(1 - \varepsilon + \varepsilon e^{-\alpha\delta}).$$

Since ε is arbitrary,

$$I^\alpha(\phi_0) \leqq \inf_{a_0\in A} T_{a_0} I^\alpha(\phi_0).$$

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

Next, we shall prove that $I^\alpha(\phi)$ 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, $I^\alpha(\phi)$ is the unique solution among the bounded functions. On the other hand, we note that

$$\left\| \inf_a T_a u - \inf_a T_a v \right\| \leqq (1 - \varepsilon + \varepsilon e^{-\alpha\delta}) \|u - v\|,$$

where $\|\cdot\|$ denotes the supremum norm. Then, $\sup_a T_a$ has the unique fixed point, i.e., there exists the unique solution to (3.2) among the bounded Borel measurable functions. Hence, $I^\alpha(\phi)$ must be Borel measurable.

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

$$T_f v\ (\phi) = T_{f(\phi)} v\ (\phi),\ \phi \in \Phi.$$

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

By the definition of f_α,

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

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

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

Hence,

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

Therefore, f_α is optimal.

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

$$\int_\Phi \int_0^\infty g(\phi',t)d\bar{p}(t|\phi,a,\phi')d\bar{q}(\phi'|\phi,a)$$

$$= \int\int_{\Phi\times R_+} g(\phi',t)d\bar{p} \otimes \bar{q}(\phi',t|\phi,a)$$

$$= \sum_s \sum_{s'} \sum_{m'} \int_0^\infty g(u(\phi,a,t,m'),t)dp_t(t|s,a,s')$$

$$\times p_s(s'|s,a)q(m'|s')\phi(s),$$

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

[1] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley.

[2] Çinlar, E. (1969). Markov renewal theory. *Adv. Appl. Prob.* 1, 123-187.

[3] Hinderer, K. (1970). *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*. Springer-Verlag.

[4] Jewell, W. (1963). Markov renewal programming I and II. *Oper. Res.* 2, 938-971.

[5] Kurano, M. (1977). On the existence of an optimal stationary I-policy in

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

non-discounted Markov decision process with incomplete state information. *Bull. Math. Statist.* 17, 75-81.

[6] Ross, S. M. (1970). *Applied Probability Models with Optimization Applications*. Holden-Day.

[7] Sawaki, K. and A. Ichikawa. (1978). Optimal control for partially observable Markov decision processes over an infinite horizon. *J. Oper. Res. Soc.* Japan. 21, 1-16.

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

[9] Sondik, E. (1978). The optimal control of partially observable Markov processes over the infinite horizon; discounted costs. *Oper. Res.* 26, 282-304.

[10] Strauch, R. E. (1966). Negative dynamic programming. *Ann. Math. Statist.* 37, 871-890.

[11] Wakuta, K. (1981). Semi-Markov decision processes with incomplete state observation - Average cost criterion-. *J. Oper. Res. Soc. Japan.* 24, 95-108.

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

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

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