Journal of the Operations Research
Society of Japan
Vol. 28, No. 3, September 1985

AVERAGE-OPTIMAL ADAPTIVE POLICIES IN SEMI-MARKOV DECISION PROCESSES INCLUDING AN UNKNOWN PARAMETER

Masami Kurano
Chiba University

(Received February 27, 1984: Revised May 8, 1985)

Abstract We consider the problem of minimizing the long-run average (expected) cost per unit time in a semi-Markov decision process including an unknown parameter. In the case of general state and action spaces and compact parameter space we construct the adaptive policy which has good properties under some identifiability conditions weaker than those for the strong consistency of the estimator. As example, we treat the age replacement with an unknown failure distribution.

1. Introduction

We consider the problem of minimizing the long-run average (expected) cost per unit time in a semi-Markov decision process (semi-MDP) including an unknown parameter. One possible approach to this problem is to estimate the unknown parameter by a maximum likelihood estimate from the observed history of the system and then take an action optimally in believing that it really equals the actual value of the parameter. This kind of adaptive policy has been studied by Kurano [12] for Markov decision models with finite state and action space. Mandl [15] has adopted the approach via minimum contrast estimates. These results are extended by Kolonko [10] to simi-MDP with denumerable state space and unbounded rewards. The consistency of a maximum likelihood estimator or a minimum contrast estimator has been studied by Kumar [11] for the finite state case and Kolonko [9] and Borkar and Varaiya [3] for the denumerable state case. Doshi and Shreve [4] has given identifiability conditions of a modified maximum likelihood estimator by randomization. They has given the conditions for the strong consistency of estimator under any policy, including the adaptive policy and under this strong consistency constructed

the optimal adaptive policy.

In this paper we treat the case of general state, action and parameter spaces and do not require the strong consistency of estimator. By the minimum contrast method ([15]) we define a modified η-minimum contrast estimator, the strong consistency of which does not necessarily hold under any policy. Further, we construct the adaptive policy which has good properties utilizing the idea which may be thought of as a modification of forced choice cycles proposed by Fox and Rolph [8]. Then, in case where the parameter space is compact the consistency of a modified η-minimum contrast estimator is shown under this adaptive policy.

In Section 2 we specify semi-MDP's including the parameter. In Section 3 some results are given for any fixed value of the parameter. In Section 4 we construct an optimal adaptive policy under some identifiability conditions. Finally in Section 5 we treat the age replacement with an unknown failure distribution as example.

2. Semi-MDP with an Unknown Parameter

In this section we formulate a semi-MDP with an unknown parameter referring to Federgruen and Tijms [5] and Kolonko [10].

A *Borel* set is a Borel subset of a complete separable metric space and for a Borel set X F_X denotes the Borel subsets of X. If X is a non-empty Borel set, $B_a(X)$ and $B_m(X)$ denote the set of all bounded lower semianalytic functions on X and the set of all bounded Baire functions on X respectively, and $P(X)$ the set of all probability measures on X. The product space of the sets D_1, $D_2, \ldots$ will be denoted by $D_1 D_2 \cdots$. For any non-empty Borel sets X and Y, a transition [probability] measure on Y given X is a function $p(\cdot|\cdot)$ on $F_Y X$ such that for each $x \in X$ $p(\cdot|x)$ is a [probability] measure on F_Y and for each $B \in F_Y$ $p(B|\cdot)$ is a Baire function on X. The set of all transition [probability] measures on Y given X is denoted by $T_m(Y|X)$ [$T_p(Y|X)$]. Also, we denote the set of all [analytically] measurable functions from X to Y by $B_m(X \to Y)$ [$B_a(X \to Y)$].

Semi-MDP's including the parameter are specified by six objects $(S, \Theta, Z, \{A(x), x \in S\}, Q, c, \tau)$, where S, Θ and Z are any Borel sets and denote the state space, the parameter space and the space of additional observations respectively, for each $x \in S$ $A(x)$ is a non-empty Borel subset of a Borel set A such that $\{(x,a) | x \in S, a \in A(x)\}$ is an element of $F_S \times F_A$, the set of actions available at state x, $Q \in T_p(ZS|SA\Theta)$ is the law of motion, $c \in B_m(SAZS\Theta)$ is one step cost

function and $\tau \in \mathbf{B}_m(SAZS\Theta \to (0,\infty))$ is a weighting function for defining the average cost.

Let Π denote the set of all policies, i.e. for $\pi = (\pi_0,\pi_1,\cdots) \in \Pi$ let $\pi_t \in T_p(A|S(AZS)^t)$ such that $\pi_t(A(x_t)|x_0,a_0,z_1,x_1,\cdots,z_t,x_t) = 1$ for all $(x_0,a_0,z_1,x_1,\cdots,z_t,x_t) \in S(AZS)^t$. A policy $\pi = (\pi_0,\pi_1,\cdots)$ is called [analytically measurable] stationary policy if there is a $f \in \mathbf{B}_m(S\to A)[\mathbf{B}_a(S\to A)]$ with $f(x) \in A(x)$ for all $x \in S$ such that $\pi_t(\{f(x_t)\}|x_0,a_0,z_1,x_1,\cdots,z_t,x_t) = 1$ for all $(x_0,a_0,z_1,x_1,\cdots,z_t,x_t) \in S(AZS)^t$ and each t. Such policy will be denoted f^∞.

The sample space is the product space $\Omega = S(AZS)^\infty$. Let X_t, Z_t and Δ_t be random quantities defined by $X_t(\omega) = x_t$, $Z_t(\omega) = z_t$ and $\Delta_t(\omega) = a_t$ for $\omega = (x_0,a_0,z_1,x_1,a_1,\cdots) \in \Omega$. For each $\theta \in \Theta$ we assume that given a policy $\pi \in \Pi$ and $x \in S$,

$$\text{Prob}(\Delta_t \in D_1 | X_0 = x, \Delta_0, \cdots, Z_t, X_t) = \pi_t(D_1 | X_0 = x, \Delta_0, \cdots, Z_t, X_t)$$

and

$$\text{Prob}(Z_{t+1} \in D_2, X_{t+1} \in D_3 | X_0 = x, \Delta_0, \cdots, Z_t, X_t, \Delta_t) = Q(D_2 D_3 | X_t, \Delta_t, \theta)$$

for each $t \geq 0$, $D_1 \in \mathcal{F}_A$, $D_2 \in \mathcal{F}_Z$ and $D_3 \in \mathcal{F}_S$.

Then, for each $\theta \in \Theta$, each starting state $x \in S$ and each policy $\pi \in \Pi$ we can define the probability measure $P^x_{\theta,\pi}$ on Ω in an obvious way.
We introduce the following conditions:

Condition(*): There are a $v \in \mathbf{B}_a(S\Theta)$ and a $g^* \in \mathbf{B}_a(\Theta)$ satisfying that

$$v(x,\theta) = \inf_{a \in A(x)} \{c(x,a,\theta) - g^*(\theta)\tau(x,a,\theta) + \int v(x',\theta) Q^S(dx'|x,a,\theta)\}, \tag{2.1}$$

where

$$c(x,a,\theta) = \int c(x,a,z,x',\theta) Q(d(z,x')|x,a,\theta),$$

$\tau(x,a,\theta) = \int \tau(x,a,z,x',\theta) Q(d(z,x')|x,a,\theta)$ and $Q^S \in T_p(S|SA\Theta)$ is the marginal distribution of Q on S defined by $Q^S(D|x,a,\theta) = \int_{ZD} Q(d(z,x')|x,a,\theta)$ for each $D \in \mathcal{F}_S$.

Condition 1. There are positive numbers τ^* and τ^{**} such that $\tau^* < \tau(x,a,\theta) < \tau^{**}$ for all $x \in S$, $a \in A(x)$ and $\theta \in \Theta$.

For $v \in \mathbf{B}_a(S\Theta)$ and $g^* \in \mathbf{B}_a(\Theta)$ as in Condition(*), let

$$\phi(x,a,z,x',\theta) = v(x,\theta) + \tau(x,a,z,x',\theta)g^*(\theta) - \{c(x,a,z,x',\theta) + v(x',\theta)\},$$

and

$$\phi(x,a,\theta) = \int \phi(x,a,z,x',\theta) Q(d(z,x') \mid x,a,\theta).$$

Then, we have the following.

Theorem 2.1. Assume that Condition(*) and Condition 1 hold. Then, we have, for each $\theta \in \Theta$, $\pi \in \Pi$ and $x \in S$,

$$\liminf_{T\to\infty} \tilde{W}_T(\theta) \geqq g^*(\theta) \quad P^x_{\theta,\pi} - \text{a.s.} . \tag{2.2}$$

where

$$\tilde{W}_T(\theta) = \frac{\sum_{t=0}^{T-1} c(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta)}{\sum_{t=0}^{T-1} \tau(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta)} .$$

Proof: By the stability theorem of Loeve [14] it holds that for any $\pi \in \Pi$, $\theta \in \Theta$ and $x \in S$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \{\phi(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta) - \phi(X_t,\Delta_t,\theta)\} = 0 \quad P^x_{\theta,\pi} - \text{a.s.} . \tag{2.3}$$

Since $\frac{1}{T} \sum_{t=0}^{T-1} \phi(X_t,\Delta_t,\theta) \leqq 0$ from Condition(*), (2.3) implies

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \phi(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta) \leqq 0 \quad P^x_{\theta,\pi} - \text{a.s.} . \tag{2.4}$$

We observe that

$$\sum_{t=0}^{T-1} \phi(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta) = g^*(\theta)\tilde{\tau}(T) - \sum_{t=0}^{T-1} c(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta) + v(X_0,\theta) - v(X_T,\theta),$$

so that

$$\tilde{\tau}(T)^{-1} \sum_{t=0}^{T-1} \phi(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta) = g^*(\theta) - \tilde{W}_T(\theta) + \tilde{\tau}(T)^{-1}(v(X_0,\theta) - v(X_T,\theta)), \tag{2.5}$$

where

$$\tilde{\tau}(T) = \sum_{t=0}^{T-1} \tau(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta).$$

Using the stability theorem again we get, from Condition 1, $\tau^* < \liminf_{T\to\infty} \frac{1}{T} \tilde{\tau}(T)$ and $\limsup_{T\to\infty} \frac{1}{T} \tilde{\tau}(T) < \tau^{**}$ $P_{\theta,\pi}$ - a.s. . Therefore, from (2.4) and (2.5), (2.2) follows. Q.E.D.

We say π^* is average-optimal if $\lim_{T\to\infty} \tilde{W}_T(\theta) = g^*(\theta)$ P^x_{θ,π^*} - a.s. for all $\theta \in \Theta$ and $x \in S$. In Section 4 an average-optimal policy will be constructed.

3. Condition(*) and Related Results

In this section, using the idea of the successive approximations and the

contraction mapping we shall give the sufficient conditions for which Condition(*) holds and derive some results used in the next section.

Condition(**): There is an $\gamma \in T_m(S|\Theta)$ such that

(a) $Q^S(D|x,a,\theta) \geqq \tau(x,a,\theta)\gamma(D|\theta)$ for any $D \in F_S$ and $(x,a,\theta) \in SA\Theta$

and

(b) there exists $0 < \beta < 1$ satisfying

$$Q^S(S|x,a,\theta) - \tau(x,a,\theta)\gamma(S|\theta) < \beta \quad \text{for any } (x,a,\theta) \in SA\Theta.$$

Define the map U on $B_a(S\Theta)$ by $Uu(x,\theta) = \inf_{a \in A(x)} U(x,a,u,\theta)$ for each $(x,\theta) \in S\Theta$, where for any $(x,a,u,\theta) \in SAB_a(S\Theta)\Theta$, $U(x,a,u,\theta) = c(x,a,\theta) + \int u(x',\theta)Q^S(dx'|x,a,\theta) - \tau(x,a,\theta)\int u(x',\theta)\gamma(dx'|\theta)$.

Then, similarly as the proof of Theorem 3.1 of [13], we can prove the following.

Theorem 3.1. Let Condition(**) hold. Then

(a) there is a $v(x,\theta) \in B_a(S\Theta)$ such that

(3.1) $v = Uv$,

and by putting $g^*(\theta) = \int v(x',\theta)\gamma(dx'|\theta)$ in (3.1) Condition(*) holds.

(b) for each $\varepsilon > 0$, there is a $f_\varepsilon \in B_a(S\Theta \to A)$ with $f_\varepsilon(x,\theta) \in A(x)$ for all $x \in S$ satisfying that

(3.2) $v(x,\theta) + \varepsilon \geqq U(x,f_\varepsilon(x,\theta),v,\theta)$ for any $(x,\theta) \in S\Theta$.

For a non-empty Borel set X, $B_c(X)[B_s(X)]$ denotes the set of all bounded continuous [lower semicontinuous] functions on X. To obtain the further results we need the following assumptions:

A1. $c \in B_s(SAZS\Theta)$ and $\tau \in B_c(SAZS\Theta)$.

A2. For $x \in S$, $A(x)$ is compact and a point to set map $A(\cdot)$ is upper semicontinuous, that is, if $x_n \in S \to x \in S$ and $a_n \in A(x_n) \to a$ as $n \to \infty$, $a \in A(x)$.

A3. $Q(\cdot|x,a,\theta)$ is weakly continuous in $(x,a,\theta) \in SA\Theta$, that is, whenever $x_n \to x$, $a_n \to a$ and $\theta_n \to \theta$, $Q(\cdot|x_n,a_n,\theta_n)$ converges weakly to $Q(\cdot|x,a,\theta)$.

A4. Condition(**) holds and $\gamma(\cdot|\theta)$ is weakly continuous in $\theta \in \Theta$.

Lemma 3.1. Suppose that A1–A4 hold, then

(a) $u \in B_s(S\Theta)$ implies $Uu \in B_s(S\Theta)$,

(b) the operator U on $B_s(S\Theta)$ is a contraction mapping.

Proof: For (a), let $\bar{Q}^S(dx'|x,a,\theta) = Q^S(dx'|x,a,\theta) - \tau(x,a,\theta)\gamma(dx'|\theta)$. Then, by A1 and A4 we have $\tau \in B_c(SA\Theta)$, so that $\bar{Q}^S(\cdot|x,a,\theta) \in T_m(S|SA\Theta)$ is weakly

continuous in $(x,a,\theta) \in SA\Theta$. Therefore, for any $u \in B_s(S\Theta)$, $U(x,a,u,\theta) \in B_s(SA\Theta)$ and $Uu \in B_s(S\Theta)$ (for example, [18]).

For any u, $u' \in B_s(S\Theta)$ $Uu(x,\theta) - Uu'(x,\theta) \leqq \int\{u(x,\theta) - u'(x,\theta)\}\bar{Q}^S(dx'|x,a,\theta)$ $\leqq ||u - u'||\beta$ for some $a \in A(x)$, where $||\cdot||$ is the supremum norm. Q.E.D.

Theorem 3.2. Suppose that A1–A4 hold, then

(a) there are a $v \in B_s(S\Theta)$ and a $g^* \in B_s(\Theta)$ satisfying the equation (2.1),

(b) there exists a $f^* \in B_m(S\Theta \to A)$ with $f^*(x,\theta) \in A(x)$ for all $x \in S$ such that (3.2) holds for $\varepsilon = 0$.

Proof: By Lemma 4.2 of [16] the metric space $(B_s(S\Theta), ||\cdot||)$ is complete, so that from Lemma 3.1 U has a unique fixed point in $B_s(S\Theta)$. Let v be the unique fixed point of U, i.e. $v = Uv$. If we put $g^*(\theta) = \int v(x',\theta)\gamma(dx'|\theta)$, $g^* \in B_s(\Theta)$ and (v,g^*) satisfies (2.1). Also, from the selection theorem (for example, [16] and [19]), (b) follows. Q.E.D.

Remark 3.1. For any $\pi \in \Pi$, $\theta \in \Theta$ and $x \in S$, define

$$\psi[\pi](x,\theta) = \limsup_{T\to\infty} \frac{E^x_{\theta,\pi}[\sum_{t=0}^{T-1} c(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta)]}{E^x_{\theta,\pi}[\sum_{t=0}^{T-1} \tau(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta)]} \tag{3.3}$$

if this expression exists, where $E^x_{\theta,\pi}$ is the expectation operator w.r.t. $P^x_{\theta,\pi}$. Then, for $f_\varepsilon \in B_a(S\Theta \to A)$ as in Theorem 3.1, the stationary policy f^∞_ε is ε-optimal for any fixed $\theta \in \Theta$ w.r.t. the expected criterion (3.3), i.e.

$\psi[f_\varepsilon(\cdot,\theta)^\infty](x,\theta) \leqq \psi[\pi](x,\theta) + \varepsilon$ for all $\pi \in \Pi$, $\theta \in \Theta$ and $x \in S$.

And for $f^* \in B_m(S\Theta \to A)$ as in Theorem 3.2, $f^*(\cdot,\theta)^\infty$ is 0-optimal for any fixed $\theta \in \Theta$. Since the proof is similar to that of Theorem 7.6 of [17], we omit it here.

Let X and Y be any Borel sets. Then $p(\cdot|y) \in T_p(X|Y)$ is said to be (*)-continuous if $y \to p(D|y)$ is uniformly continuous for $D \in \{$ colsed sets of $X \}$. The following is easily proved.

Lemma 3.2. Let X, Y and W be any Borel sets and $u \in B_s(XW)$. If $p(\cdot|y) \in T_p(X|Y)$ is (*)-continuous, $y \to \int u(x,w)p(dx|y)$ is uniformly continuous in $y \in Y$ for $w \in W$.

Theorem 3.3. Suppose that A1–A4 hold. Further, assume that (i) both $c(x,a,z,x',\theta)$ and $\tau(x,a,z,x',\theta)$ are uniformly continuous in $\theta \in \Theta$ and (ii) $Q(\cdot|x,a,\theta)$ and $\gamma(\cdot|\theta)$ are (*)-continuous in $\theta \in \Theta$ uniformly for $(x,a) \in SA$. Then, $v(x,\theta) \in B_s(S\Theta)$ and $g^* \in B_s(\Theta)$ as in Theorem 3.2 become to be uniformly continuous in $\theta \in \Theta$.

Proof: Under given assumptions, from Lemma 3.2, $c(x,a,\theta)$ and $\tau(x,a,\theta)$ are uniformly continuous in $\theta \in \Theta$, so that by the same proof as that of Thorem 3.2 we obtain Theorem 3.3. Q.E.D.

4. Estimation and Adaptive Policy

In this section we construct the average-optimal policy under some identifiability conditions. Hereafter we assume that Condition (**) holds. Extending the notion of a minimum contrast function ([15]) we give the following definitions.

Definition. Let $m \in B_m(SAZS\Theta)$ and $\delta \in T_p(A|S)$ with $\delta(A(x)|x) = 1$ for all $x \in S$. Then, (m,δ) is called informative w.r.t. semi-MDP's including the unknown parameter if the following conditions B1-B3 are satisfied.

B1. $m(x,a,z,x',\theta)$ is uniformly continuous in $\theta \in \Theta$ for $(x,a,z,x') \in SAZS$.

B2. Let $\bar{m}(x,a,\theta,\theta') = \int \bar{m}(x,a,z,x',\theta,\theta')Q(d(z,x')|x,a,\theta)$ where $\bar{m}(x,a,z,x',\theta,\theta') = m(x,a,z,x',\theta) - m(x,a,z,x',\theta')$. Then, $\bar{m}(x,a,\theta,\theta') \leqq 0$ for any $x \in S$, $a \in A(x)$ and any $\theta,\theta' \in \Theta$.

B3. Further let $\bar{m}(x,\theta,\theta') = \int \bar{m}(x,a,\theta,\theta')\delta(da|x)$. Then, $\int \bar{m}(x,\theta,\theta')\gamma(dx|\theta) < 0$ for any $\theta \neq \theta' \in \Theta$.

By using the minimum contrast method ([15]) and the idea which may be thought of as a modification of forced choice cycles introduced by Fox and Rolph [8], we shall construct the policy $\pi^* = (\pi_0^*,\pi_1^*,\cdots)$ which will be proved average-optimal. For a sequence $\{\sigma(t)\}_0^\infty$ of non-negative numbers with $\sigma(t) \leqq 1$ for each t we define the sequence $\{Y_t\}_0^\infty$ of independent random variables by $\mathrm{Prob}(Y_t = 1) = 1 - \mathrm{Prob}(Y_t = 0) = \sigma(t)$ for all $t \geqq 0$.

Now, assume that there exists $m \in B_m(SAZS\Theta)$ and $\delta \in T_p(A|S)$ with $\delta(A(x)|x) = 1$ for all $x \in S$ such that (m,δ) is informative.

Set for $T \geqq 1$ and $\theta \in \Theta$,

$$L_T(\theta) = \sum_{t=0}^{T-1} Y_t \cdot m(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta). \tag{4.1}$$

Then, the selection theorem (for example, see [19]) implies that for any $\eta > 0$ there is a $\hat{\theta}_{T,\eta} \in B_a(\{0,1\}^T S(AZS)^T \to \Theta)$ such that

$$\inf_{\theta \in \Theta} L_T(\theta) \geqq L_T(\hat{\theta}_{T,\eta}(\bar{H}_T)) - \eta \quad \text{for all } \bar{H}_T, \tag{4.2}$$

where $\bar{H}_T = (Y_0,\cdots,Y_{T-1},H_T)$ and $H_T = (X_0,\Delta_0,Z_1,X_1,\cdots,Z_T,X_T)$.

Let us call $\hat{\theta}_{T,\eta}$ the modified η-minimum contrast estimator at time T. Here, for any positive number η, a sequence $\{\varepsilon(t)\}_1^\infty$ of positive numbers and $\{Y_t\}$ defined above, we define the policy $\pi^* = (\pi_0^*, \pi_1^*, \cdots)$ as follows: For any $a_0 \in A(X_0)$ let π_0^* be $\pi_0^*(\{a_0\}|X_0) = 1$. And for any $t \geq 1$ define π_t^* by $\pi_t^*(\{f_{\varepsilon(t)}(X_t, \hat{\theta}_t)\}|\bar{H}_t) = 1$ if $Y_t = 0$ and $\pi_t^*(\cdot|\bar{H}_t) = \delta(\cdot|Y_t)$ if $Y_t = 1$, where $f_{\varepsilon(t)} \in B_a(S\Theta \to A)$ is given in Theorem 3.1 and $\hat{\theta}_t = \hat{\theta}_{t,\eta}$ is the modified η-minimum contrast estimator at time t.

Note that π^* is dependent on both $\{\varepsilon(t)\}_1^\infty$ and $\{\sigma(t)\}_0^\infty$. The policy π^* means that if $Y_t = 1$ we select the action according to δ by force and if $Y_t = 0$ we estimate the value of the parameter by a modified η-minimum contrast estimator using the observed history through the first t times and take the action which would be $\varepsilon(t)$-optimal if the estimated value was the actual value of the parameter. Let us call π^* the adaptive policy constructed by $(m, \delta, \{\varepsilon(t)\}, \{\sigma(t)\}, \eta)$.

Let $\bar{P}^x_{\theta,\pi^*}$ be a product measure $P^* \times P^x_{\theta,\pi^*}$ on the product space $\{0,1\}^\infty \Omega$, where P^* is a probability measure on $\{0,1\}^\infty$ such that the projection on the t-th factor $\{0,1\}$ is Y_t and $P^*(Y_t=1) = 1 - P^*(Y_t=0) = \sigma(t)$.

The next theorem shows the consistency of $\hat{\theta}_{T,\eta}$ under the policy π^*.

Theorem 4.1. Assume that Condition(**) and Condition 1 in Section 2 hold and Θ is compact. Let (m, δ) be informative and π^* the adaptive policy constructed by $(m, \delta, \{\varepsilon(t)\}, \{\sigma(t)\}, \eta)$. Then if $\sum_{t=0}^\infty \sigma(t) = \infty$ and $\sum_{T=1}^\infty \{\sum_{t=0}^{T-1} \sigma(t)\}^{-2} < \infty$, it holds that for any $\theta \in \Theta$ and $x \in S$ $\hat{\theta}_{T,\eta} \to \theta$ $\bar{P}^x_{\theta,\pi^*}$ - a.s. .

Proof: For any fixed $\theta_0 \in \Theta$, put $\bar{m}(x,a,z,x',\theta) = \bar{m}(x,a,z,x',\theta_0,\theta)$ and $\bar{m}(x,\theta) = \bar{m}(x,\theta_0,\theta)$ for simplicity. Now, let $C_{\theta,n} = \{x \in S | \bar{m}(x,\theta) < -\frac{1}{n}\}$ for each $\theta \in \Theta$ and n. Then, since $\int \bar{m}(x,\theta)\gamma(dx|\theta_0) < 0$ $(\theta_0 \neq \theta)$ by B3, we can find an integer j and a positive number α such that $\gamma(C_{\theta,j}|\theta_0) \geq \alpha$. By Condition (**) and Condition 1, we have

$$(4.3) \qquad P^x_{\theta_0,\pi^*}(X_t \in C_{\theta,j}|H_{t-1}) \geq \inf_{x \in S,\ a \in A(x)} Q^S(C_{\theta,j}|x,a,\theta_0)$$

$$\geq \tau^* \gamma(C_{\theta,j}|\theta_0) \geq \tau^* \alpha.$$

Let $\bar{\sigma}(T) = \sum_{t=0}^{T-1} \sigma(t)$. Then, the stability theorem of Loeve [14] says that

$$(4.4) \qquad \lim_{T\to\infty} \{\bar{\sigma}(T)\}^{-1} \sum_{t=0}^{T-1} \{Y_t \bar{m}(X_t, \Delta_t, Z_{t+1}, X_{t+1}, \theta)$$

$$- \bar{E}^x_{\theta_0,\pi^*}[Y_t \bar{m}(X_t, \Delta_t, Z_{t+1}, X_{t+1}, \theta)|\bar{H}_t]\} = 0 \quad \bar{P}^x_{\theta_0,\pi^*}\text{-a.s.},$$

where $\bar{E}^{X}_{\theta_0,\pi^*}$ is the expection operator w.r.t. $\bar{P}^{X}_{\theta_0,\pi^*}$.

Since $\limsup_{T\to\infty}\{\bar{\sigma}(T)\}^{-1}\sum_{t=0}^{T-1}\bar{E}^{X}_{\theta_0,\pi^*}[Y_t\bar{m}(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta)|\bar{H}_t]$

$= \limsup_{T\to\infty}\{\bar{\sigma}(T)\}^{-1}\sum_{t=0}^{T-1}\sigma(t)\bar{m}(X_t,\theta)$, from the definition of π^*

$\leq \limsup_{T\to\infty}\{\bar{\sigma}(T)\}^{-1}\sum_{t=0}^{T-1}\sigma(t)\bar{m}(X_t,\theta)I_{C_{\theta,j}}(X_t)$, from B2

$\leq -\frac{1}{j}\limsup_{T\to\infty}\{\bar{\sigma}(T)\}^{-1}\sum_{t=0}^{T-1}\sigma(t)I_{C_{\theta,j}}(X_t)$

$\leq -\frac{1}{j}\alpha\tau^*$ $\bar{P}^{X}_{\theta_0,\pi^*}$-a.s., from (4.3) and the stability theorem,

it holds from (4.4) that

(4.5) $\limsup_{T\to\infty}\{\bar{\sigma}(T)\}^{-1}\sum_{t=0}^{T-1}Y_t\bar{m}(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta) \leq -\frac{1}{j}\alpha\tau^*$ $\bar{P}^{X}_{\theta_0,\pi^*}$-a.s.,

where I_C is the indicator function of the set C.

The above discussion show that for any $\theta\neq\theta_0$ and $(n,p)\in K_\theta = \{(n,p)\,|\,\gamma(C_{\theta,n}|\theta_0) > \frac{1}{p}\}$, there is a null set $N_{\theta,n,p}$, that is, $\bar{P}^{X}_{\theta_0,\pi^*}(N_{\theta,n,p}) = 0$ such that for any $\omega^*\notin N_{\theta,n,p}$ it holds that $\limsup_{T\to\infty}\{\bar{\sigma}(T)\}^{-1}\{L_T(\theta_0) - L_T(\theta)\} \leq -\frac{1}{n}\frac{1}{p}\tau^*$ and $\limsup_{T\to\infty}\{\bar{\sigma}(T)\}^{-1}\sum_{t=0}^{T-1}Y_t = 1$.

Since Θ is separable, there exists a subset of Θ, $\{\bar{\theta}_i\,|\,i=1,2,\ldots\}$, which is dense in Θ. Here, we define $N = \bigcup_{i=1}^{\infty} N_{\bar{\theta}_i}$, where $N_{\bar{\theta}_i} = \bigcup_{(n,p)\in K_{\bar{\theta}_i}} N_{\bar{\theta}_i,n,p}$.

Note that $\bar{P}^{X}_{\theta_0,\pi^*}(N) = 0$. For simplicity, let $\hat{\theta}_T = \hat{\theta}_{T,\eta}$. Now, we assume that there exists $\omega^*\notin N$ and $\theta^*\in\Theta$ $(\theta^*\neq\theta_0)$ which satisfy that $\hat{\theta}_{T_j}(\omega^*)\to\theta^*$ for some subsequence $\{\hat{\theta}_{T_j}(\omega^*)\}_1^\infty$ of $\{\hat{\theta}_T(\omega^*)\}_1^\infty$. The denseness shows that there is a subsequence $\{\bar{\theta}_{i_j}\}_1^\infty$ of $\{\bar{\theta}_i\}_1^\infty$ such that $\bar{\theta}_{i_j}\to\theta^*$ as $j\to\infty$. By the uniform continuity of $\bar{m}$ with respect to θ, for any integer d, there exist k and n $(n>d)$ such that

$$\bigcap_{j\geq k} C_{\bar{\theta}_{i_j},n} \supset C_{\theta^*,d}\,.$$

Let d and p be such that $\gamma(C_{\theta^*,d}|\theta_0) > \frac{1}{p}$. Then, by the above result, there exist k and n such that

$$\gamma\Big(\bigcap_{j\geq k} C_{\bar{\theta}_{i_j},n}|\theta_0\Big) > \frac{1}{p},$$

which implies

$$\gamma(C_{\bar{\theta}_{i_j},n}|\theta_0) > \frac{1}{p} \quad \text{for all } j \geq k.$$

Therefore, we have

$$\limsup_{T\to\infty}\{\bar{\sigma}(T)\}^{-1}\{L_T(\theta_0) - L_T(\bar{\theta}_{i_j})\} \leq -\frac{1}{n}\frac{1}{p}\tau^* \quad \text{for all } j \geq k.$$

By the uniform continuity of m w.r.t. θ, for any $\varepsilon > 0$ there is an j^* such that $|\, m(x,a,z,x',\theta^*) - m(x,a,z,x',\bar{\theta}_{i_j})\,| < \varepsilon$ for all $j \geq j^*$.
In this case,

$$|\,\{\bar{\sigma}(T)\}^{-1}\{L_T(\theta_0) - L_T(\bar{\theta}_{i_j})\} - \{\bar{\sigma}(T)\}^{-1}\{L_T(\theta_0) - L_T(\theta^*)\}\,|$$

$$\leq \varepsilon\{\bar{\sigma}(T)\}^{-1}\textstyle\sum_{t=0}^{T-1} Y_t,$$

so that

$$\limsup_{T\to\infty}\{\bar{\sigma}(T)\}^{-1}\{L_T(\theta_0) - L_T(\theta^*)$$

$$\leq \limsup_{T\to\infty}\bar{\sigma}(T)^{-1}\{L_T(\theta_0) - L_T(\bar{\theta}_{i_j})\} + \varepsilon \limsup_{T\to\infty}\bar{\sigma}(T)^{-1}\textstyle\sum_{t=0}^{T-1} Y_t$$

$$\leq -\frac{1}{n}\frac{1}{p}\tau^* + \varepsilon \quad \text{as } j \to \infty.$$

Hence, we have

$$\limsup_{T\to\infty}\bar{\sigma}(T)^{-1}\{L_T(\theta_0) - L_T(\theta^*)\} < 0. \tag{4.6}$$

On the other hand, it holds from (4.2) that $\bar{\sigma}(T_j)^{-1}\{L_{T_j}(\theta_0) - L_{T_j}(\hat{\theta}_{T_j}) + \eta\} \geq 0$ for all $j \geq 1$, so that by repeating the above discussions we obtain $\liminf_{j\to\infty} \sigma(T_j)^{-1}\{L_{T_j}(\theta_0) - L_{T_j}(\theta^*)\} \geq 0$, which contradicts (4.6). Q.E.D.

Theorem 4.2. Assume that the assumptions of Theorem 4.1 hold. If $\frac{1}{T}\sum_{t=0}^{T-1} \varepsilon(t) \to 0$ and $\frac{1}{T}\sum_{t=0}^{T-1} \sigma(t) \to 0$ as $T \to \infty$ and $\phi(x,a,\theta)$ is uniformly continuous in $\theta \in \Theta$ for $(x,a) \in SA$, then π^* is average-optimal.

Proof: For any fixed $\theta_0 \in \Theta$ and $x \in S$, put $P^x = P^x_{\theta_0,\pi^*}$ and $E = E^x_{\theta_0,\pi^*}$ for simplicity. By the stability theorem [14], we have

$$\lim_{T\to\infty} \frac{1}{T}\textstyle\sum_{t=0}^{T-1} \{\phi(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta_0) - \phi(X_t,\Delta_t,\theta_0)\} = 0 \quad P\text{-a.s.}, \tag{4.7}$$

where $\phi(x,a,z,x',\theta)$ and $\phi(x,a,\theta)$ are defined in Section 2.
Since $\phi(X_t, f_{\varepsilon(t)}(X_t,\hat{\theta}_t),\hat{\theta}_t) \geqq -\varepsilon(t)$ by the definition of $f_{\varepsilon(t)}$, it follows that

$$(4.8)\qquad \frac{1}{T}\sum_{t=0}^{T-1}\phi(X_t,\Delta_t,\theta_0) = \frac{1}{T}\sum_{t=0}^{T-1}(1-Y_t)\phi(X_t,\Delta_t,\theta_0) + \frac{1}{T}\sum_{t=0}^{T-1}Y_t\phi(X_t,\Delta_t,\theta_0)$$

$$\geqq \frac{1}{T}\sum_{t=0}^{T-1}(1-Y_t)\{\phi(X_t,f_{\varepsilon(t)}(X_t,\hat{\theta}_t),\theta_0) - \phi(X_t,f_{\varepsilon(t)}(X_t,\hat{\theta}_t),\hat{\theta}_t) - \varepsilon(t)\} + \frac{1}{T}\sum_{t=0}^{T-1}Y_t\phi(X_t,\Delta_t,\theta_0).$$

We observe that $\limsup_{T\to\infty}\left|\frac{1}{T}\sum_{t=0}^{T-1}Y_t\phi(X_t,\Delta_t,\theta_0)\right| \leqq C \times \limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}Y_t = 0$ P^*-a.s. for some constant $C > 0$.
Since, by Theorem 4.1 and the assumptions of Theorem 4.2, $\phi(x,a,\theta_0) - \phi(x,a,\hat{\theta}_t) \to 0$ uniformly for $(x,a) \in SA$ as $t \to \infty$, it follows from (4.7) and (4.8) that

$$(4.9)\qquad \liminf_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\phi(X_t,\Delta_t,Z_{t+1},X_{t+1},\theta_0) \geqq 0 \quad P\text{-a.s.}\ .$$

By repeating the proof of Theorem 2.1, we have $\limsup_{T\to\infty}\tilde{W}_T(\theta_0) \leqq g^*(\theta_0)$ P-a.s., so that it holds from Theorem 2.1 that $g^*(\theta_0) = \lim_{T\to\infty}\tilde{W}_T(\theta_0)$ P-a.s. which shows that π^* is average-optimal. Q.E.D.

Remark 4.1. By Theorem 3.3, if the conditions of Theorem 3.3 are satisfied, then $\phi(x,a,\theta)$ is uniformly continuous in $\theta \in \Theta$.
We can prove that under the assumptions of Theorem 4.2 π^* is average-optimal in the expected criterion, i.e. $\psi[\pi^*](x,\theta) \leqq \psi[\pi](x,\theta)$ for all $\pi \in \Pi$, $\theta \in \Theta$ and $x \in S$.

Remark 4.2. For any integer N and $0 < s < \frac{1}{2}$, let $\sigma(t) = 1$ if $t < N$ and $= t^{-s}$ if $t \geqq N$. Then, the sequence $\{\sigma(t)\}$ satisfies the conditions of Theorem 4.2.

Remark 4.3. Let $\{m_1,m_2,\cdots\}$ and $\{n_1,n_2,\cdots\}$ be increasing sequences of positive integers such that (i) $\liminf_{k\to\infty}k^{-(1+s)/2}\max\{j \mid m_j \leqq k\} > 1$ for some $s > 0$ and (ii) $\limsup_{k\to\infty}k^{-1}\max\{j \mid m_j \leqq k\} = 0$ and (iii) $\limsup_{k\to\infty}k^{-1}\max\{j \mid n_j \leqq k\} = 0$. For example, if $m_j = j^{\alpha}$ $(2>\alpha>1)$ for all $j \geqq 1$, $\{m_j\}$ satisfies (i) and (ii). Now, define the sequence $\{\sigma(t)\}$ by $\sigma(t) = 1$ if $t \in \{m_1,m_2,\cdots\}$, $\sigma(t) = 0$ elsewhere and the sequence $\{\varepsilon(t)\}$ by $\varepsilon(t) = 1$ if $t \in \{n_1,n_2,\cdots\}$, $\varepsilon(t) = 0$ elsewhere. Then, $\{\sigma(t)\}$ and $\{\varepsilon(t)\}$ satisfy the conditions of Theorem 4.2.

5. Age Replacement

We consider the problem of minimizing the long-run average cost per unit time in the age replacement (for example, [2], [6]) with an unknown failure distribution. Fox [7] has treated the discounted case assuming that the true failure distribution belongs to the parametric family of a Weibul with a Gamma prior.

Also, Arukumar [1] has given the nonparametric estimate of optimum age under a sample of size n and shows its consistency.

Under an age replacement policy a unit is always replaced at failure or at the end of a specified time interval a, whichever occurs first, with respectively cost c_1 and c_2 $(c_1>c_2>0)$.

Age replacement corresponds to a semi-MDP with one state ([6]). Define $S = \{x_0\}$, $A = [b,\infty)$ $\{\infty\}$ and $Z = R^+$, where b is any positive number and $R^+ = (0,\infty)$. A stage is the period starting just after a replacement and ending just after the next replacement. The length of each stage is represented as the element of Z and $a \in [b,\infty)$ and ∞ correspond to the action of the planned replacement time a and non-planned replacement respectively. Let $A(x_0) = A$.

For any positive numbers M_1 and M_2 $(M_1< M_2)$, let

$$\bar{\Theta} = \{ F \mid F \text{ is a continuous failure distribution on } R^+, \text{ and } M_1 \leqq \int_0^b x dF \text{ and } \int_0^\infty x dF \leqq M_2 \}.$$

Note that $\bar{\Theta}$ is complete and separable w.r.t. the suppremum norm. Let Θ be a compact subset of $\bar{\Theta}$. Further define $Q(D\{x_0\}|x_0,a,F) = \mu(F^a)(D)$ for all $D \in \mathcal{F}_{R^+}$, where $\mu(F)$ is the probability measure induced by the distribution function $F \in \Theta$ and $F^a(x) = F(x)$ if $x < a$, $= 1$ if $x \geqq a$.

Finally define $c(x_0,a,z,x_0,F) = c_1(c_2)$ if $z \leqq (>)\ a$, and $\tau(x_0,a,z,x_0,F) = \min\{a,z\}$. Since $0 < M_1 \leqq \tau(x_0,a,F) \leqq M_2$ for all $a \in R^+$ and $F \in \Theta$, Condition(**) is satisfied with $\gamma(\{x_0\} \mid F) = (2M_2)^{-1}$. In this case, the optimal equation (3.1) is as follows:

$$(5.1) \qquad v(F) = \inf_{a \in A} \{ c_1F(a) + c_2\bar{F}(a) + v(F) - (2M_2)^{-1}v(F)\tau(x_0,a,F) \}$$

for all $F \in \Theta$,

where $\bar{F}(a) = 1 - F(a)$.

Putting $g^*(F) = (2M_2)^{-1}v(F)$, (5.1) becomes

$$(5.2) \qquad g^*(F) = \inf_{a \in A} \{ c_1F(a) + c_2\bar{F}(a) \} / \int x dF^a(x),$$

which agrees with the well-known results ([2] and [6]).

Assume that for any $\varepsilon > 0$, there exists $d > 0$ for which $\int_d^\infty x dF < \varepsilon$ for all $F \in \Theta$. Then from the fundamental calculus, we observe that $g^*(F)$ is continuous in $F \in \Theta$, so that $\phi(x_0,a,F) = (2M_2)^{-1}v(F)\tau(x_0,a,F) - c_1F(a) - c_2\bar{F}(a)$ is uniformly continuous in $F \in \Theta$ for $a \in A$.

Let $\{r_j \mid j=1,2,\cdots\}$ be the set of all rational numbers belonging to R^+. Now, define $m(x_0,a,z,x_0,F) = \sum_{j=0}^{\infty} \alpha^j\{ I_{(-\infty,r_j]}(z) - F^a(r_j) \}^2$ for some $0 < \alpha < 1$ and $\delta(\{\infty\}|x_0) = 1$. Then, we shall show that (m,δ) is informative w.r.t. this age replacement. In fact, $\bar{m}(x_0,a,F,F') = -\sum_{j=1}^{\infty} \alpha^j\{ F^a(r_j) - F'^a(r_j) \}^2 \leqq 0$ and $\bar{m}(x_0,F,F') = -\sum_{j=1}^{\infty} \alpha^j\{ F(r_j) - F'(r_j) \}^2 < 0$ for all $F \neq F' \in \Theta$.

For L_T defined by (4.1),

$$L_T(F) = \sum_{j=1}^{\infty} \alpha^j \sum_{t=0}^{T-1} Y_t\{ I_{(-\infty,r_j]}(Z_{t+1}) - F(r_j) \}^2$$

$$= \sum_{j=1}^{\infty} \alpha^j \sum_{t=0}^{T-1} Y_t\{I_{(-\infty,r_j]}(Z_{t+1}) - \tilde{F}_T(r_j)\}^2$$

$$+ (\sum_{t=0}^{T-1} Y_t)\sum_{j=1}^{\infty}\alpha^j\{\tilde{F}_T(r_j) - F(r_j)\}^2,$$

where $\tilde{F}_T(z) = \{\sum_{t=0}^{T-1} Y_t\}^{-1}\sum_{t=0}^{T-1} Y_t I_{(-\infty,z]}(Z_{t+1})$ is the empirical distribution function, so that the modified η-minimum contrast estimator $\hat{\theta}_{T,\eta}$ is dependent only on $\tilde{F}_T$. We note that by applying Theorem 4.2 an average-optimal adaptive replacement policy is constructed.

Acknowledgement

The author wishes to express his thanks to referees for their very helpful comments which led to an improvement of the presentation of the material.

References

[1] Arunkumar, S.: Nonparametric age replacement policy. *Sankhya* A, 34 (1972), 251-256.

[2] Barlow, R. E. and Proschan, F.: *Mathematical Theory of Reliability*, Wiley, New York, 1965.

[3] Borkar, V. and Varaiya, P.: Identification and adaptive control of Markov chains. *SIAM J. Contr. Optimiz.* 20 (1982), 470-489.

[4] Doshi, B. and Shreve, S. E.: Strongly consistency of a modified maximum likelihood estimator for controlled Markov chains. *J. Appl. Prob.* 17 (1980), 726-734.

[5] Federgruen, A. and Tijms, H. C.: The optimalty equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms. *J. Appl. Prob.*, 15 (1978), 356-373.

[6] Fox, B. L.: Age replacement with discounting. *Operations Research*, 14 (1966), 533-537.

[7] Fox, B. L.: Adaptive age replacement. *J. Math. Anal. Appl.*, 18 (1968), 365-376.

[8] Fox, B. L. and Rolph, T. E.: Adaptive policies for markov renewal programs. *Annals of Statist.* 1 (1973), 334-341.

[9] Kolonko, M.: Strongly consistent estimation in a controlled Markov renewal madel. *J. Appl. Prob.* 19 (1982), 532-545.

[10] Kolonko, M.: The average-optimal adaptive control of a Markov renewal model in presence of an unknown parameter. *Optimization*, 13 (1982), 567-591.

[11] Kumar, P. R.: Adaptive control with a compact parameter set. *STAM J. Contr. optimiz.* 20 (1982), 9-13.

[12] Kurano, M.: Discrete-time Markovian decision processes with an unknown parameter- average return criterion. *J. Op. Res. Soc. Japan* 15 (1972), 67-76.

[13] Kurano, M.: Semi-Markov decision processes and their applications in replacement models. *J. Op. Res. Soc. Japan* 28 (1985), 18-29.

[14] Loeve, M.: *Probability theory,* Second Edition. D. Van Nostrand Co. Inc., New Jersy, (1960).

[15] Mandl, P.: Estimation and Control in Markov chains. *Adv. Appl. Prob.* 6 (1974), 40-60.

[16] Maitra, A.: Discounted dynamic programming on compact metric spaces. *Sankhya*, Series A, 30 (1968), 211-216.

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

[18] Schäl, M.: Conditions for Optimality in Dynamic Programming and for Limit of n-stage Optimal Policies to be Optimal. *Zritschrift Wahrscheinlichkeitst-theorie verw.* 32 (1975), 179-196.

[19] Shreve, S. E. and Bertsekas, D. P.: Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control. *SIAM J. Contr. Optimiz.* 16 (1978), 953-978.

Masami KURANO: Department of Mathematics
Faculty of Education Chiba University
Yayoi-cho, Chiba, 260, Japan