KYBERNETIKA — VOLUME 9 (1973), NUMBER 2

Necessary and Sufficient Optimality Conditions for Average Reward of Controlled Markov Chains

KAREL SLADKÝ

On the base of the recurrence relation for expected reward necessary and sufficient optimality conditions for average reward and average overtaking reward of controlled finite Markov chains are inferred. Some results are also extended to semi-Markov decision processes in the Appendix.

0. INTRODUCTION AND NOTATIONS

Necessary and sufficient optimality conditions for discounted Markov chains were derived in [3]. A more general form of these conditions can be found [5]. Using an interesting recurrence relation for expected reward also in [5] necessary and sufficient optimality conditions for average reward of ergodic Markov chains were studied.

This paper presents a general form of the recurrence relation for expected reward of controlled (non-ergodic) Markov chains. On the base of this recurrence relation necessary and sufficient optimality conditions of controlled (non-ergodic) Markov chains are inferred for average reward and also for more selective Veinott's average overtaking reward optimality criterion (compare [2], [6]). For the average overtaking case the obtained results are compared with that inferred in [2]. In the Appendix the average reward optimality conditions are also extended to the case of ergodic semi-Markov decision processes and it is indicated how the obtained results can be employed for investigation of continuous time Markovian decision models. Notations used in [4], [5] are followed in this paper.

We shall investigate a controlled Markov chain with state space $I = \{1, 2, \ldots, r\}$ and the set of control parameter values $z \in J = \{1, 2, \ldots, s\}$ in any of the states. Choosing control parameter value $z \in J$ in state $j \in I$ state $k \in I$ will be reached in the next transition with given probability $p(j, k; z)$. $c(j, k; z)$ denotes the reward associated with such a transition. The values $p(j, k; z)$, $c(j, k; z)$ are supposed to be known for any pair j, $k \in I$ and any $z \in J$. For the sake of brevity we shall introduce the expected one stage reward $\bar{c}(j; z)$ in state j if control parameter takes value z. Obviously,

$$\bar{c}(j; z) = \sum_{k \in I} p(j, k; z)\, c(j, k; z)\,.$$

A control ω of the chain is given by a sequence of control parameter values z chosen with respect to the complete history of the chain. So we write $\omega = \{z_n(j_0, j_1, \ldots, j_n), n = 0, 1, \ldots\}$ where $z_n(j_0, j_1, \ldots, j_n)$ is the control parameter value chosen at the n-th transition following the occurrence of states $j_0, j_1, \ldots, j_n$. ω is called a Markovian (memoryless) control if $z_n(j_0, j_1, \ldots, j_n) = z_n(j_n)$, for $n = 0, 1, \ldots$. A Markovian control is called homogeneous if $z_n(j_n) = z(j_n)$. For homogeneous Markovian control we write $\omega \sim z(j)$. The chosen control ω together with the transition probabilities and the initial state $j \in I$ define the probability distribution P_j^ω of a sequence $\{X_n, n = 0, 1, \ldots\}$ of random variables describing the development of the chain under control ω (of course, $X_0 = j$). By E_j^ω we denote the mathematical expectation with respect to this probability distribution. For shortening we shall often delete the arguments $X_0, \ldots, X_k$ and $j_0, \ldots, j_k$ in $z_k(\ldots)$ e.g. we shall write z_2 instead of $z_2(X_0, X_1, X_2)$ or instead of $z_2(j_0, j_1, j_2)$.

The (random) reward up to the next n following transitions is given by

$$C_n = \sum_{m=1}^{n} c(X_{m-1}, X_m; z_{m-1})$$

(we set $C_0 = 0$). The quality of control ω of the considered controlled Markov chain is considered according to the limit behaviour (if $n \to \infty$) of $(1/n)\, E_j^\omega C_n$. Obviously,

$$E_j^\omega C_n = \sum_{m=1}^{n} E_j^\omega\, \bar{c}(X_{m-1}; z_{m-1})\,,$$

where $X_0 = j$. A control $\hat{\omega}$ will be called average optimal if for an arbitrary control ω and any $j \in I$

$$\lim_{L \to \infty} \frac{1}{L}\, E_j^{\hat{\omega}} C_L \geqq \limsup_{L \to \infty} \frac{1}{L}\, E_j^\omega C_L\,.$$

A more selective than average reward optimality criterion is the following average overtaking optimality criterion introduced by Veinott in [6]. A control ω^* will be called average overtaking optimal if for an arbitrary control ω and any $j \in I$

$$\liminf_{L \to \infty} \frac{1}{L} \sum_{M=0}^{L-1} (E_j^{\omega^*} C_M - E_j^\omega C_M) \geqq 0\,.$$

1. NECESSARY AND SUFFICIENT OPTIMALITY CONDITIONS FOR AVERAGE REWARD

Lemma 1.1. *There exists a homogeneous Markovian control* $\hat{\omega} \sim \hat{z}(j)$ *and numbers* $\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_r$; $\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_r$ *such that*

$$(1.1) \qquad \hat{g}_j = \sum_{k \in I} p(j, k; \hat{z}(j)) \,.\, \hat{g}_k \geqq \sum_{k \in I} p(j, k; z) \,.\, \hat{g}_k$$

for all $z \in J$ *and any* $j \in I$;

$$(1.1') \qquad \hat{w}_j + \hat{g}_j = \bar{c}(j; \hat{z}(j)) + \sum_{k \in I} p(j, k; \hat{z}(j)) \,.\, \hat{w}_k \geqq$$
$$\geqq \bar{c}(j; z) + \sum_{k \in I} p(j, k; z) \,.\, \hat{w}_k$$

for any $j \in I$ *and any* $z \in J$ *for which* (1.1) *holds with equality.*

Proof. The rigorous proof can be found in [4] or [6]. Homogeneous Markovian control $\hat{\omega} \sim \hat{z}(j)$ and the numbers $\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_r$; $\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_r$ can be found e.g. by Howard's policy iteration procedure described in [1] or [4]. □

We shall denote $J_j \subset J$ the set of all control parameter values in state j for which (1.1) holds with equality.

Remark 1.2. Let us denote

$$(1.2) \qquad \psi(j; z) = \sum_{k \in I} p(j, k; z) \,.\, \hat{g}_k - \hat{g}_j \,,$$

$$(1.2') \qquad \varphi(j; z) = \bar{c}(j; z) + \sum_{k \in I} p(j, k; z) \,.\, \hat{w}_k - \hat{w}_j - \hat{g}_j \,.$$

For homogeneous Markovian control $\hat{\omega} \sim \hat{z}(j)$ and any $j \in I$, $\psi(j; \hat{z}(j)) = 0$, $\varphi(j; \hat{z}(j)) = 0$. (Note that $\psi(j; z) \leqq 0$ for any $j \in I$, $z \in J$ and $\varphi(j; z) \leqq 0$ for any $j \in I$ and $z \in J_j$).

Theorem 1.3. *For an arbitrary control* $\omega \sim z_n$ *and all* $M = 0, 1, \ldots$

$$(1.3) \quad E_j^\omega C_M - \hat{g}_j \,.\, M = \sum_{n=0}^{M-1} [(M - 1 - n) \,.\, E_j^\omega \, \psi(X_n; z_n) + E_j^\omega \, \varphi(X_n; z_n)] +$$
$$+ \hat{w}_j - E_j^\omega \hat{w}_{X_M} \,.$$

Proof. By induction with respect to M. For $M = 0$ equation (1.3) is trivially fulfilled, for $M = 1$ equation (1.3) reads

$$\bar{c}(j; z) - \hat{g}_j = \varphi(j; z) + \hat{w}_j - \sum_{k \in I} p(j, k; z) \,.\, \hat{w}_k$$

that coincides with (1.2').

Let (1.3) hold for M. As $E_j^\omega C_{M+1} = E_j^\omega C_M + E_j^\omega \bar{c}(X_M; z_M)$ we have

$$(1.4)\ E_j^\omega C_{M+1} - \hat{g}_j . (M+1) = \sum_{n=0}^{M-1} [(M-1-n) . E_j^\omega \psi(X_n; z_n) + E_j^\omega \varphi(X_n; z_n)] +$$

$$+ \hat{w}_j - E_j^\omega \hat{w}_{X_M} + E_j^\omega \bar{c}(X_M; z_M) - \hat{g}_j + E_j^\omega \hat{w}_{X_{M+1}} - E_j^\omega \hat{w}_{X_{M+1}} +$$

$$+ E_j^\omega \hat{g}_{X_M} - E_j^\omega \hat{g}_{X_M} = \sum_{n=0}^{M} E_j^\omega \varphi(X_n; z_n) + \hat{w}_j - E_j^\omega \hat{w}_{X_{M+1}} +$$

$$+ \sum_{n=0}^{M-1} (M-1-n) . E_j^\omega \psi(X_n; z_n) + \sum_{n=1}^{M} (E_j^\omega \hat{g}_{X_n} - E_j^\omega \hat{g}_{X_{n-1}}) .$$

As $E_j^\omega \hat{g}_{X_{n+1}} - E_j^\omega \hat{g}_{X_n} = E_j^\omega \psi(X_n; z_n)$ setting into (1.4) we obtain

$$E_j^\omega C_{M+1} - \hat{g}_j . (M+1) = \sum_{n=0}^{M} E_j^\omega \varphi(X_n; z_n) + \hat{w}_j - E_j^\omega \hat{w}_{X_{M+1}} +$$

$$+ \sum_{n=0}^{M-1} (M-n) . E_j^\omega \psi(X_n; z_n) . \ \square$$

Remark 1.4. For the ergodic case $\hat{g}_j = \hat{g}$ for any $j \in I$ and the set of equations (1.1), (1.1′) takes the form

$$(1.1^*) \qquad w_j + g = \max_{z \in J} [\bar{c}(j; z) + \sum_{k \in I} p(j, k; z) . w_k] .$$

Under these assumptions $\psi(j; z) = 0$ for any $j \in I$, $z \in J$ (so $g_j = g$ for any $j \in I$) and the recurrence relation (1.3) reads

$$(1.3^*) \qquad E_j^\omega C_M - \hat{g}_j . M = \sum_{n=0}^{M-1} E_j^\omega \varphi(X_n; z_n) + \hat{w}_j - E_j^\omega \hat{w}_{X_M} .$$

Relation (1.3*) was inferred in [5] and it can be also extended to the case of semi-Markov decision processes as it is shown in the Appendix.

Theorem 1.5. *A control ω is average optimal if and only if the following conditions are satisfied for any $j \in I$ and all $n = 0, 1, \ldots$*

$$(1.5) \qquad E_j^\omega \psi(X_n; z_n) = 0 ,$$

$$(1.6) \qquad \lim_{M \to \infty} \frac{1}{M} \sum_{m=0}^{M-1} E_j^\omega \varphi(X_m; z_m) = 0 .$$

Proof. Obviously,

$$(1.7) \qquad \lim_{M \to \infty} \frac{1}{M} . E_j^\omega C_M = \hat{g}_j \Leftrightarrow \lim_{M \to \infty} \frac{1}{M} (E_j^\omega C_M - \hat{g}_j . M) = 0 .$$

(Notice that equation (1.7) is fulfilled for homogeneous Markovian control $\hat{\omega} \sim \hat{z}(j)$).

Let $\chi(j; z) = 1$ if $\psi(j; z) < 0$ (resp. $\chi(j, z) = 0$ if $\psi(j; z) = 0$) and let $\mathcal{D}$ be the set of all pairs $(j; z)$ for which $\chi(j; z) = 1$. As the control parameter z can take only a finite number of values there exists

$$\max_{\substack{j \in I \\ z \in J}} \varphi(j; z) = K_1 \geqq 0 \quad \text{and} \quad \max_{(j;z) \in \mathcal{D}} \psi(j; z) = K_2 < 0 \, .$$

First we shall show that condition (1.5) must be fulfilled for any average optimal control. Let us therefore suppose $E_j^\omega \, \psi(X_n; z_n)$ to be negative for certain n, say let $E_j^\omega \, \psi(X_{n_0}; z_{n_0}) < 0$. As $\psi(j; z) = 0 \Rightarrow \varphi(j; z) \leqq 0$ from (1.3) we obtain for any $M > n_0 + m_0$ (where integer $m_0 \geqq -K_1/K_2$, $m_0 < -K_1/K_2 + 1$)

$$\text{(1.8)} \qquad \frac{1}{M} (E_j^\omega C_M - \hat{g}_j \, . \, M - \hat{w}_j - E_j^\omega \hat{w}_{X_M}) =$$

$$= \frac{1}{M} \sum_{n=0}^{M-1} [(M - 1 - n) \, . \, E_j^\omega \, \psi(X_n; z_n) + E_j^\omega \, \varphi(X_n; z_n)] \leqq$$

$$\leqq \frac{1}{M} \, . \, [(M - 1 - n_0) \, . \, E_j^\omega \, \psi(X_{n_0}; z_{n_0}) + E_j^\omega \, \varphi(X_{n_0}; z_{n_0})] +$$

$$+ \frac{1}{M} \sum_{n=M-m_0}^{M-1} [(M - 1 - n) \, . \, K_2 + K_1] \, . \, E_j^\omega \, \chi(X_n; z_n) \, .$$

But for $M \to \infty$

$$\frac{1}{M} \sum_{n=M-m_0}^{M-1} [(M - 1 - n) \, . \, K_2 + K_1] \to 0 \, .$$

So from (1.8) we have

$$\text{(1.9)} \qquad \limsup_{M \to \infty} \frac{1}{M} E_j^\omega C_M - \hat{g}_j \leqq E_j^\omega \, \psi(X_{n_0}; z_{n_0}) < 0$$

and the control ω cannot be average optimal (compare (1.7)). So condition (1.5) must hold if ω is average optimal.

Under condition (1.5) equation (1.3) can be written as (1.3*). From (1.3*), (1.7) fulfilling (1.6) (in case that (1.5) holds) must be also the necessary and sufficient average reward optimality condition. □

From (1.8) we also simply obtain the following Corollary.

Corollary 1.6. *If $E_j^\omega \, \psi(X_{n_0}; z_{n_0}) < 0$ then there exists certain $\bar{K} < 0$ and M_0 such that for any $M \geqq M_0$*

$$\text{(1.10)} \qquad \frac{1}{M} \sum_{n=0}^{M-1} [(M - 1 - n) \, . \, E_j^\omega \, \psi(X_n; z_n) + E_j^\omega \, \varphi(X_n; z_n)] \leqq \bar{K} < 0 \, .$$

The properties of controls determined by the solution of the set of equations (1.1), (1.1′) are summarized in the following Corollary.

Corollary 1.7. *Homogeneous Markovian control $\hat{\omega} \sim \hat{z}(j)$ is average optimal. Moreover,*

$$E_j^{\hat{\omega}} C_M - \hat{g}_j \,.\, M = \hat{w}_j - E_j^{\hat{\omega}} \hat{w}_{X_M} \,. \tag{1.3'}$$

In case that state j belongs to an aperiodic class of states (with respect to transition probability matrix determined by control $\hat{\omega}$) then even

$$\lim_{M \to \infty} \left(E_j^{\hat{\omega}} C_M - \hat{g}_j \,.\, M \right) = \hat{w}_j - \sum_{k \in I} \pi(j, k; \hat{\omega}) \,.\, \hat{w}_k \tag{1.3''}$$

where $\pi(j, k; \hat{\omega})$ is the limit probability of transition probability matrix

$$\mathbf{P}^{\hat{\omega}} = \| p(j, k; \hat{z}(j)) \|_{j,k=1}^{r} \quad (\hat{\omega} \sim \hat{z}(j))$$

defined as

$$\| \pi(j, k; \hat{z}(j)) \|_{j,k=1}^{r} = \lim_{m \to \infty} \frac{1}{m+1} \sum_{n=0}^{m} (\mathbf{P}^{\hat{\omega}})^n \,.$$

2. NECESSARY AND SUFFICIENT OPTIMALITY CONDITIONS FOR AVERAGE OVERTAKING OPTIMAL CONTROLS

In this paragraph the recurrence relation inferred in paragraph 1 (compare Theorem 1.3) will be employed for investigating optimality conditions for average overtaking optimal controls. First we shall formulate some lemmas.

Comparing with Lemma 1.1 we shall need a deeper insight into the set of numbers fulfilling conditions (1.1), (1.1′) with equality. We can formulate

Lemma 2.1. *There exists a homogeneous Markovian control $\omega^* \sim z^*(j)$ and the (unique) numbers $g_1^*, g_2^*, \ldots, g_r^*$; $w_1^*, w_2^*, \ldots, w_r^*$; $u_1^*, u_2^*, \ldots, u_r^*$ corresponding to ω^* such that*

$$g_j^* = \sum_{k \in I} p(j, k; z^*(j)) \,.\, g_k^* \geqq \sum_{k \in I} p(j, k; z) \,.\, g_k^* \tag{2.1}$$

for all $z \in J$ and any $j \in I$;

$$\begin{aligned} w_j^* + g_j^* &= \bar{c}(j; z^*(j)) + \sum_{k \in I} p(j, k; z^*(j)) \,.\, w_k^* \geqq \\ &\geqq \bar{c}(j; z) + \sum_{k \in I} p(j, k; z) \,.\, w_k^* \end{aligned} \tag{2.1'}$$

for any $j \in I$ and any $z \in J$ for which (2.1) *holds with equality*;

(2.1″) $$u_j^* = -w_j^* + \sum_{k \in I} p(j, k; z^*(j)) \, . \, u_k^* \geqq$$
$$\geqq -w_j^* + \sum_{k \in I} p(j, k; z) \, . \, u_k^*$$

for any $j \in I$ and any $z \in J$ for which (2.1), (2.1′) *hold with equality if for any $j \in I$*

(2.2) $$\sum_{k \in I} \pi(j, k; \omega^*) \, . \, w_k^* = 0 \, ,$$

(2.2′) $$\sum_{k \in I} \pi(j, k; \omega^*) \, . \, u_k^* = 0 \, .$$

($\pi(j, k; \omega)$ *again denotes the limit probability of transition probability matrix* $\boldsymbol{P}^\omega = \|p(j, k; z(j))\|_{j,k=1}^r$ (*if* $\omega \sim z(j)$).

Proof. The proof can be found in [6]. Theorem 6 in [6] also provides an algorithm for finding homogeneous Markovian control $\omega^* \sim z^*(j)$ and the numbers $g_1^*, g_2^*, \ldots, g_r^*; w_1^*, w_2^*, \ldots, w_r^*; u_1^*, u_2^*, \ldots, u_r^*$. □

We shall again denote $J_j \subset J$ (resp. $J'_j \subset J_j$) the set of all control parameter values in state j for which (2.1) (resp. (2.1), (2.1′)) hold with equality.

Comparing with the values $\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_r; \hat{w}_1, \hat{w}_2, \ldots, \hat{w}_r$ determined in Lemma 1.1 we can easily see that $\hat{g}_j = g_j^*$ for any $j \in I$.

Remark 2.2. Let us denote

(2.3) $$\bar{\psi}(j; z) = \sum_{k \in I} p(j, k; z) \, . \, g_k^* - g_j^* \quad (\text{so } \bar{\psi}(j; z) = \psi(j; z)) \, ,$$

(2.3′) $$\bar{\varphi}(j; z) = \bar{c}(j; z) + \sum_{k \in I} p(j, k; z) \, . \, w_k^* - w_j^* - g_j^* \, ,$$

(2.3″) $$\bar{\gamma}(j; z) = -w_j^* + \sum_{k \in I} p(j, k; z) \, . \, u_k^* - u_j^* \, .$$

For homogeneous Markovian control $\omega^* \sim z^*(j)$ and any $j \in I$, $\bar{\psi}(j; z^*(j)) = 0$, $\bar{\varphi}(j; z^*(j)) = 0$, $\bar{\gamma}(j; z^*(j)) = 0$. (Note that $\bar{\psi}(j; z) \leqq 0$ for any $j \in I$, $z \in J$; $\bar{\varphi}(j; z) \leqq \leqq 0$ for any $j \in I$, $z \in J_j$; $\bar{\gamma}(j; z) \leqq 0$ for any $j \in I$, $z \in J'_j$).

Setting for any $j \in I$, $\hat{g}_j = g_j^*$, $\hat{w}_j = w_j^*$ equation (1.3) reads

(2.4) $$E_j^\omega C_M - g_j^* \, . \, M = \sum_{n=0}^{M-1} [(M - 1 - n) \, . \, E_j^\omega \, \bar{\psi}(X_n; z_n) + E_j^\omega \, \bar{\varphi}(X_n; z_n)] +$$
$$+ w_j^* - E_j^\omega w_{X_M}^* \, .$$

will be based on the following identity (2.5) obtained by summing (2.4) for $M = 0, 1, 2, \ldots, L - 1$:

$$(2.5) \quad \frac{1}{L} \sum_{M=0}^{L-1} \left[E_j^{\omega} C_M - \left(g_j^* \cdot \frac{L-1}{2} + w_j^* - E_j^{\omega^*} w_{X_M}^* \right) \right] =$$

$$= \frac{1}{L} \sum_{M=0}^{L-1} \left\{ \sum_{n=0}^{M-1} \left[(M - 1 - n) \, . \, E_j^{\omega} \, \bar{\psi}(X_n; z_n) + E_j^{\omega} \, \bar{\varphi}(X_n; z_n) \right] - \right.$$

$$\left. - \left(E_j^{\omega} w_{X_M}^* - E_j^{\omega^*} w_{X_M}^* \right) \right\} .$$

Remark 2.3. From (2.2)

$$\lim_{L \to \infty} \frac{1}{L} \sum_{M=0}^{L-1} E_j^{\omega^*} w_{X_M}^* = 0 .$$

So from (2.5) using homogeneous Markovian control $\omega^* \sim z^*(j)$ we have for any $j \in I$

$$(2.5') \qquad \lim_{L \to \infty} \frac{1}{L} \sum_{M=0}^{L-1} \left[E_j^{\omega^*} C_M - \left(g_j^* \cdot \frac{L-1}{2} + w_j^* \right) \right] = 0 .$$

Now we shall formulate the main result of this paragraph.

Theorem 2.4. *A control ω is average overtaking optimal if and only if the following conditions are fulfilled for all $n = 0, 1, \ldots$ and any $j \in I$*

$$(2.6) \qquad E_j^{\omega} \, \bar{\psi}(X_n; z_n) = 0 ,$$

$$(2.7) \qquad E_j^{\omega} \, \bar{\varphi}(X_n; z_n) = 0$$

and

$$(2.8) \qquad \lim_{M \to \infty} \frac{1}{M} \sum_{m=0}^{M-1} E_j^{\omega} \, \bar{\gamma}(X_m; z_m) = 0 .$$

Proof. As $E_j^{\omega} \, \bar{\psi}(X_n; z_n) = 0 \Rightarrow E_j^{\omega} \, \bar{\varphi}(X_n; z_n) \leqq 0$ using the results of Corollary 1.6 if for certain $n = n_0$, $E_j^{\omega} \, \bar{\psi}(X_{n_0}; z_{n_0}) < 0$ then there exists certain $\bar{K} < 0$ and integer M_0 such that for any $M \geqq M_0$

$$\frac{1}{M} \sum_{n=0}^{M-1} \left[(M - 1 - n) \, . \, E_j^{\omega} \, \bar{\psi}(X_n; z_n) + E_j^{\omega} \, \bar{\varphi}(X_n; z_n) \right] \leqq \bar{K} < 0 .$$

As

$$\frac{1}{L} \sum_{M=0}^{N-1} \sum_{n=0}^{M-1} (M - 1 - n) \, . \, E_j^{\omega} \, \bar{\psi}(X_n; z_n) + E_j^{\omega} \, \bar{\varphi}(X_n; z_n)] , \quad \frac{1}{L} \sum_{M=0}^{L-1} E_j^{\omega} w_{X_M}^*$$

are uniformly bounded from above for any $N \leqq M_0$ and any L the righthand side of (2.5) tends to $-\infty$ for $L \to \infty$ if for certain $n = n_0$, $E_j^\omega \bar{\psi}(X_{n_0}; z_{n_0}) < 0$. So (if we compare (2.5′) and (2.5)) condition (2.6) must be fulfilled for any control that is average overtaking optimal. As

$$\sum_{M=0}^{L-1} \sum_{n=0}^{M-1} E_j^\omega \bar{\varphi}(X_n; z_n) = \sum_{n=0}^{L-1} (L - 1 - n) \,.\, E_j^\omega \bar{\varphi}(X_n; z_n)$$

under condition (2.6) equation (2.5) reads

$$\text{(2.9)} \qquad \frac{1}{L} \sum_{M=0}^{L-1} \left[E_j^\omega C_M - \left(g_j^* \,.\, \frac{L-1}{2} + w_j^* - E_j^{\omega^*} w_{X_M}^* \right) \right] =$$

$$= \frac{1}{L} \sum_{n=0}^{L-1} [(L - 1 - n) \,.\, E_j^\omega \bar{\varphi}(X_n; z_n) - (E_j^\omega w_{X_n}^* - E_j^{\omega^*} w_{X_n}^*)] \,.$$

From (2.3″) we have

$$\text{(2.10)} \qquad E_j^\omega \bar{\gamma}(X_n; z_n) = -E_j^\omega w_{X_n}^* + E_j^\omega u_{X_{n+1}}^* - E_j^\omega u_{X_n}^* \,.$$

Summing (2.10) for $n = 0, 1, 2, \ldots, L - 1$ we obtain

$$\text{(2.11)} \qquad -\sum_{n=0}^{L-1} E_j^\omega w_{X_n}^* = \sum_{n=0}^{L-1} E_j^\omega \bar{\gamma}(x_n; z_n) + u_j^* - E_j^\omega u_{X_L}^*$$

and (as $\bar{\gamma}(j; z^*(j)) = 0$ for $\omega^* \sim z^*(j)$)

$$\text{(2.12)} \qquad -\sum_{n=0}^{L-1} (E_j^\omega w_{X_n}^* - E_j^{\omega^*} w_{X_n}^*) = \sum_{n=0}^{L-1} E_j^\omega \bar{\gamma}(X_n; z_n) - (E_j^\omega u_{X_L}^* - E_j^{\omega^*} u_{X_L}^*) \,.$$

Setting from (2.12) into (2.9) we obtain

$$\text{(2.13)} \qquad \frac{1}{L} \sum_{M=0}^{L-1} \left[E_j^\omega C_M - \left(g_j^* \,.\, \frac{L-1}{2} + w_j^* - E_j^{\omega^*} w_{X_M}^* \right) \right] =$$

$$= \frac{1}{L} \{ \sum_{n=0}^{L-1} [(L - 1 - n) \,.\, E_j^\omega \bar{\varphi}(X_n; z_n) + E_j^\omega \bar{\gamma}(X_n; z_n)] - (E_j^\omega u_{X_L}^* - E_j^{\omega^*} u_{X_L}^*) \} \,.$$

As

$$E_j^\omega \bar{\psi}(X_n; z_n) = 0 \,, \quad E_j^\omega \bar{\varphi}(X_n; z_n) = 0 \Rightarrow E_j^\omega \bar{\gamma}(X_n; z_n) \leqq 0$$

using again the results of paragraph 1 (compare (1.10), (2.13) and employ the results of Corollary 1.6) if for certain $n = n_0$, $E_j^\omega \bar{\varphi}(X_{n_0}; z_{n_0}) < 0$ then

$$\text{(2.14)} \qquad \limsup_{L \to \infty} \frac{1}{L} \sum_{M=0}^{L-1} \left[E_j^\omega C_M - \left(g_j^* \,.\, \frac{L-1}{2} + w_j^* \right) \right] < 0$$

and (compare (2.14) and (2.5′)) the control ω cannot be average overtaking optimal.

Thus condition (2.7) must be also fulfilled for any average overtaking optimal control and any average overtaking optimal control must be taken from the set $\bar{J}' = J'_1 \times J'_2 \times \ldots \times J'_r$. Under the conditions (2.6), (2.7) equation (2.13) reads

$$(2.15) \qquad \frac{1}{L} \sum_{M=0}^{L-1} \left[E_j^\omega C_M - \left(g_j^* \cdot \frac{L-1}{2} + w_j^* - E_j^{\omega^*} w_{X_M}^* \right) \right] =$$

$$= \frac{1}{L} \cdot \left[\sum_{n=0}^{L-1} E_j^\omega \, \bar{\gamma}(X_n; z_n) - (E_j^\omega u_{X_L}^* - E_j^{\omega^*} u_{X_L}^*) \right] .$$

As $\bar{\gamma}(j; z) \leqq 0$ for any $z \in J'_j$ and

$$\lim_{L \to \infty} \frac{1}{L} \sum_{M=0}^{L-1} E_j^{\omega^*} w_{X_M}^* = 0$$

from (2.15)

$$(2.16) \qquad \lim_{L \to \infty} \frac{1}{L} \sum_{M=0}^{L-1} \left[E_j^\omega C_M - \left(g_j^* \cdot \frac{L-1}{2} + w_j^* \right) \right] = 0$$

if and only if under conditions (2.6), (2.7) also condition (2.8) is fulfilled and for any other control ω

$$\liminf_{L \to \infty} \frac{1}{L} \sum_{M=0}^{L-1} \left(g_j^* \cdot \frac{L-1}{2} + w_j^* - E_j^\omega C_M \right) =$$

$$= \liminf_{L \to \infty} \frac{1}{L} \sum_{M=0}^{L-1} (E_j^{\omega^*} C_M - E_j^\omega C_M) > 0. \ \square$$

Corollary 2.5. *Homogeneous Markovian control $\omega^* \sim z^*(j)$ is not only average optimal but even average overtaking optimal.*

Remark 2.6. In case that $\|p(j, k; z)\|_{j,k=1}^r$ is ergodic for any control $\omega \sim z(j)$ condition (2.6) is fulfilled for any control ω.

Corollary 2.7. *Using the recurrence relation (2.4) it can be easily seen that conditions (2.6), (2.7) are fulfilled (for any $j \in I$ and all $n = 0, 1, \ldots$) if and only if it holds for all $M = 1, 2, \ldots$*

$$(2.17) \qquad E_j^\omega C_M = g_j^* \, . \, M + w_j^* - E_j^\omega w_{X_M}^* \, .$$

If the conditions $E_j^\omega \, \bar{\psi}(X_n; z_n) = 0$, $E_j^\omega \, \bar{\varphi}(X_n; z_n) = 0$ are satisfied for all $n = 0, 1, \ldots$ and any $j \in I$ employing (2.11) condition (2.8) will be fulfilled if and only if for any $j \in I$

$$(2.18) \qquad \lim_{L \to \infty} \frac{1}{L} \sum_{n=0}^{L-1} E_j^\omega \, w_{X_n}^* = 0 \, .$$

So (2.17) together with (2.18) forms also the necessary and sufficient optimality conditions for average overtaking optimal controls. Conditions (2.17), (2.18) were inferred by another approach in [2].

APPENDIX

AVERAGE REWARD OPTIMALITY CONDITIONS FOR ERGODIC SEMI-MARKOV DECISION PROCESSES AND CONTINUOUS TIME DECISION MODELS

Let us suppose that the transition of the considered Markov chain (that is ergodic for any possible control) from state j into state k is associated with the values $c(j, k; z)$; $d(j, k; z) \geqq 0$ that are supposed to be known for any j, $k \in I$, $z \in J$. $d(j, k; z)$ can be interpreted as the time spent in state j before transition into state k occurs and

$$\bar{d}(j; z) = \sum_{k \in I} p(j, k; z) \,.\, d(j, k; z)$$

as the expected time spent in state j under control parameter value z. Such an object is a particular case of a semi-Markov decision process. Denoting D_n the (random) time up to the n following transitions, obviously,

$$E_j^\omega D_n = \sum_{m=1}^{n} E_j^\omega \, \bar{d}(X_{m-1}; z_{m-1}) \,,$$

where $X_0 = j$.

By an analogy with Markov chains it can be shown the existence of the solution (denoted $\hat{g}$; $\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_r$) of the set of equations

$$\text{(A.1)} \quad w_j = \max_{z \in J} \left[\bar{c}(j; z) + \sum_{k \in I} p(j, k; z) \,.\, w_k - \bar{d}(j; z) \,.\, g\right] \quad (j = 1, 2, \ldots, r)$$

if $\bar{d}(j; z) > 0$ at least for one state that is recurrent for any admissible control. Denoting

$$\text{(A.1')} \qquad \varphi(j; z) = \bar{c}(j; z) + \sum_{k \in I} p(j, k; z) \,.\, \hat{w}_k - \hat{w}_j - \bar{d}(j; z) \,.\, \hat{g}$$

then for any control $\omega \sim z_n$ $(n = 0, 1, 2, \ldots)$

$$\text{(A.1'')} \quad E_j^\omega \, \bar{c}(X_n; z_n) - \hat{g} \,.\, E_j^\omega \, \bar{d}(X_n; z_n) = E_j^\omega \, \varphi(X_n; z_n) + E_j^\omega \, \hat{w}_{X_n} - E_j^\omega \hat{w}_{X_{n+1}} \,.$$

If we sum (A.1'') for $n = 0, 1, 2, \ldots, M - 1$ we immediately obtain

$$\text{(A.2)} \qquad E_j^\omega C_M - \hat{g} \,.\, E_j^\omega D_M = \sum_{n=0}^{M-1} E_j^\omega \, \varphi(X_n; z_n) + \hat{w}_j - E_j^\omega \hat{w}_{X_M} \,.$$

As $(1/M)\,E_j^\omega D_M$ is uniformly bounded and positive for $M \to \infty$

$$(\text{A.3}) \qquad \lim_{M\to\infty} \frac{E_j^\omega C_M}{E_j^\omega D_M} = \hat{g} \Leftrightarrow \lim_{M\to\infty} \frac{1}{M}\left(E_j^\omega C_M - \hat{g}\,.\,E_j^\omega D_M\right) = 0\,.$$

Let us denote by $C(t)$ the (random) reward up to time t and let $E_j^\omega\, C(t)$ be the expected value of $C(t)$ under control ω (if $X_0 = j$). If the random variable $M(t)$ denotes the total number of transitions up to time t then, obviously,

$$E_j^\omega\big(C(t) - C_{M(t)}\big) \leqq E_j^\omega\big(C_{M(t)+1} - C_{M(t)}\big) \leqq \max_{j,k;z} c(j, k; z)\,,$$

resp.

$$E_j^\omega\big(t - D_{M(t)}\big) \leqq E_j^\omega\big(D_{M(t)+1} - D_{M(t)}\big) \leqq \max_{j,k;z} d(j, k; z)\,,$$

and

$$(\text{A.4}) \qquad \lim_{t\to\infty} \frac{1}{t}\, E_j^\omega\big(C(t) - C_{M(t)}\big) = 0\,,$$

$$(\text{A.5}) \qquad \lim_{t\to\infty} \frac{1}{t}\, E_j^\omega D_{M(t)} = 1$$

(as $(1/t)\, E_j^\omega D_{M(t)} = 1 - \big(t - E_j^\omega D_{M(t)}\big)/t$).

As $t \to \infty \Rightarrow M(t) \to \infty$ from (A.4), (A.5) and the relation

$$\frac{1}{t}\, E_j^\omega\, C(t) = \frac{E_j^\omega C_{M(t)}}{E_j^\omega D_{M(t)}}\,.\,\frac{E_j^\omega D_{M(t)}}{t} + \frac{1}{t}\, E_j^\omega\big(C(t) - C_{M(t)}\big)$$

we can infer that it holds for an arbitrary control $\omega \sim z_n$ and all $j \in I$

$$(\text{A.6}) \qquad \limsup_{t\to\infty} \frac{1}{t}\, E_j^\omega\, C(t) = \limsup_{M\to\infty} \frac{E_j^\omega C_M}{E_j^\omega D_M}\,,$$

resp.

$$(\text{A.6}') \qquad \liminf_{t\to\infty} \frac{1}{t}\, E_j^\omega\, C(t) = \liminf_{M\to\infty} \frac{E_j^\omega C_M}{E_j^\omega D_M}\,.$$

(Similar result can be also found in the book S. M. Ross: Applied Probability Models with Optimization Applications, Holden-Day, San Francisco 1970.)

As $\varphi(j; z) \leqq 0$ from (A.2), (A.3), (A.6) it can be easily seen that

$$\lim_{t\to\infty} \frac{1}{t}\, E_j^\omega C(t) = \lim_{M\to\infty} \frac{E_j^\omega C_M}{E_j^\omega D_M} = \hat{g}$$

if and only if

$$\text{(A.7)} \qquad \lim_{M \to \infty} \frac{1}{M} \sum_{n=0}^{M-1} E_j^{\omega}\, \varphi(X_n;\, z_n) = 0$$

and that there exists no control ω for which

$$\limsup_{M \to \infty} \frac{E_j^{\omega} C_M}{E_j^{\omega} D_M} > \hat{g}\,.$$

Obviously, homogeneous Markovian control $\hat{\omega} \sim \hat{z}(j)$ for which equation (A.1) is fulfilled is average optimal.

Under the ergodic properties of the set of transition probability matrices it can be shown that (A.7) is the necessary and sufficient average reward optimality condition even if the set of control parameter values

$$Z_j = \bigcup_{i=1}^{s_j} Z_{ji}$$

where s_j is a given number, all Z_{ji} are compact and $p(j, k; z)$, $c(j, k; z)$, $d(j, k; z)$ are continuous functions of z on any Z_{ji} (under these assumptions solution of equation (A.1) exists). These results can be also employed for investigation of continuous time Markovian decision process X considered as a generalization of semi-Markov decision processes if we allow the control parameter value to be changed between transitions (we only suppose that the selected control parameter value must be unchanged at least for a given time $e > 0$ if the trajectory of X remains in any of the states).

Let us therefore consider a semi-Markov decision process X with state space I, transition probabilities p_{jk}^{u} from state j into state k if the control parameter value $u \in U_j = \{1, 2, \ldots, s_j\}$ is selected in state j being in state j for a random sojourn time $\Theta_j^u \leqq K$ (where $K > 0$ is a given number) with known distribution function $F(j; (u, \tau)) = P\{\Theta_j^u < \tau\}$. $F(j; (u, \tau))$ is supposed to be a continuous function of τ with $F(j; (u, 0)) = 0$, $F(j; (u, K)) = 1$ and the resulting transition probability matrix is supposed to be ergodic for any admissible control. As to the reward structure of X we shall suppose that the reward r_{jk} is associated with transition from state j into state k and the reward rate v_j^k is being obtained in state j if the control parameter value $u \in U_j$ is selected.

Let us construct a Markov chain embedded into the considered process X at the time instants at which new control parameter is selected. Of course, control parameter value $u \in U_k$ must be chosen whenever a new state $k \neq j$ is reached from state j and if the trajectory of X remains in state j after elapsing of an arbitrary chosen time $\tau \in \langle e; K \rangle$ new control parameter value from U_j can be used (choosing $\tau = K$ means that no new control parameter is selected if the trajectory of X remains in state j).

The transition probabilities of the embedded Markov chain are given by

$$p(j, j; (u, \tau)) = 1 - F(j; (u, \tau)) \tag{A.8}$$

and for $j \neq k$

$$p(j, k; (u, \tau)) = p^u_{jk} \,.\, F(j; (u, \tau)) \,. \tag{A.8'}$$

The expected time spent (resp. the expected reward obtained) in state j up to the next transition of the embedded Markov chain can be calculated as

$$\bar{d}(j; (u, \tau)) = \int_0^\tau \xi \,.\, \mathrm{d}F(j; (u, \xi)) + \tau \,.\, [1 - F(j; (u, \tau))] \,, \tag{A.9}$$

resp.

$$\bar{c}(j; (u, \tau)) = \sum_{k \in I} p(j, k; (u, \tau)) \,.\, r_{jk} + v^u_j \,.\, \bar{d}(j; (u, \tau)) \,. \tag{A.9'}$$

Considering the pair (u, τ) as a control parameter value $z \in Z_j = U_j \times \langle e; K \rangle$ in state j from the optimality properties of the embedded Markov chain (compare (A.1) – (A.7)) optimality conditions for continuous time Markovian decision process X can be inferred. From the average reward optimality properties of the embedded Markov chain follows that the long range average reward of the considered continuous time decision process X can be found among stationary decision rules specifying for each state j certain control parameter value $\hat{u}_j \in U_j$ and a fixed time $\hat{\tau}_j \in \langle e; K \rangle$ during which control parameter value $\hat{u}_j$ is being used if the trajectory of X remains in state j.

(Similar result was also obtained in the paper Chitgopekar, S. S.: Continuous Time Markovian Sequential Control Processes, SIAM Journal Control *7* (1969), 3, 367—389.)

(Received August 1, 1972.)

REFERENCES

[1] Howard, R. A.: Dynamic Programming and Markov Processes. M.I.T. and Wiley Press, New York 1960.

[2] Denardo, E. V., Miller, B. L.: An Optimality Condition for Discrete Dynamic Programming with No Discounting. Annals Mathem. Statistics *39* (1968), 4, 1220—1227.

[3] Lippman, S. A.: On the Set of Optimal Policies in Discrete Dynamic Programming. Journal Mathem. Analysis Applic. *24* (1968), 2, 440—445.

[4] Mandl, P.: Controlled Markov Chains (in Czech). Kybernetika *6* (1969), Supplement, 1—74.

[5] Mandl, P.: On the Variance in Controlled Markov Chains. Kybernetika *7*, (1971), 1, 1—12.

[6] Veinott, A. F.: On Finding Optimal Policies in Discrete Dynamic Programming with No Discounting. Annals Mathem. Statistics *37* (1966), 5, 1284—1294.

Ing. Karel Sladký, Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Vyšehradská 49, 128 48 Praha 2. Czechoslovakia.