J. Austral. Math. Soc. Ser. B 38(1997), 542–562

MARKOV DECISION PROGRAMMING – THE MOMENT OPTIMAL PROBLEM FOR THE FIRST-PASSAGE MODEL

LIU JIANYONG and LIU KE[1]

(Received 21 July 1994; revised 21 July 1995)

Abstract

In this paper, we discuss MDP-the moment optimal problem for the first-passage model. A policy improvement iteration algorithm is given for finding the k-moment optimal stationary policy.

1. Introduction

Allowing for the risk factor Jaquette [5, 6] posed a moment optimality model for a discounted Markov decision process. Sobel [15] presented a formula for the k-th moment of the total discounted return. A minimal variance problem (that is, a two-moment optimal problem) in optimal policies for the discounted MDP was discussed in [2, 12]. A moment optimality model in which the discount factor is dependent on history was discussed in [10]. For other works in the field see also Baykal- Gürsoy and Ross [1], Filar, Kallenberg and Lee [3], Filar and Lee [4], Kawai [7], Chung [8, 9], Sobel [13, 14] and White [16].

This paper discusses the moment optimal problem for the first-passage model on the basis of [11]. The first-passage model is also of practical interest. In particular, the model can be applied to solve optimal control problems of reliability and queueing systems and other controlled stochastic systems.

A k-moment is defined in Section 2. Some formulas for k-moments are given by Theorem 2.1 in Section 2. Sufficient and necessary conditions for a policy π to be a k-moment optimal policy are given by Theorem 2.6. Theorems 2.7 and 2.8 state that the problems of the existence and calculation of a k-moment optimal policy (or a moment-optimal policy) in the space of general policies can be changed into the same problems in the space of deterministic stationary policies. Theorem 2.9 states that there exists a stationary policy which is moment optimal if A is nonempty and

[1]Institute of Applied Mathematics, Academia Sinica, Beijing 100080, PRC

https://doi.org/10.1017/S0334270000000850 Published online by Cambridge University Press

finite. An algorithm of policy-improvement type is given in Section 3 for finding the k-moment optimal stationary policy.

The first-passage model with denumerable state space is $\{S, A, q, r, V_k\}$, where the state space S and action set A are nonempty and countable. Let $S = \{0, 1, 2, \dots\}$, $S_0 = \{1, 2, 3, \dots\}$. A one-step reward r satisfies $|r(i, a)| \le M$ and $r(0, a) = 0$, $i \in S, a \in A$. The symbol q denotes the family of stationary one-step transition laws: when the system is in state i and we take an action a, the system moves to a new state j selected according to the conditional probability $q(j|i, a)$, where q satisfies $q(0|0, a) = 1, a \in A$. A definition of criterion V_k is given in Section 2.

The set of general policies $\pi = (\pi_0, \pi_1, \pi_2, \dots)$ is denoted by Π. A mapping $f : S \to A$ is called a deterministic decision rule. Let F denote the set of all deterministic decision rules f. For $f \in F$, $f^\infty = (f, f, \dots)$ is called a stationary policy. Π_s^d denotes the set of all stationary policies. Obviously $\Pi_s^d \subset \Pi$.

At any stage $t (\ge 0)$, X_t and Δ_t denote respectively a state of the system and an action taken in that state.

ASSUMPTION A. There exists a real number $\alpha > 0$ and a positive integer N such that $P_\pi\{x_N = 0|x_0 = i\} \ge \alpha$ for $\forall\pi \in \Pi, \forall i \in S_0$.

In the following, we assume that Assumption A is always true.

Let $X_0 = i_0$, $\Delta_0 = a_0$, $X_1 = i_1$, $\Delta_1 = a_1$, ..., $X_n = i_n$. The sequence $h_n = (i_0, a_0, i_1, a_1, \dots, i_n)$ is called a history up to stage n and $H_n (n \ge 0)$ denotes the set of all h_n.

Let $\pi = (\pi_0, \pi_1, \pi_2, \dots) \in \Pi$, $h_n = (i_0, a_0, i_1, a_1, \dots, i_n) \in H_n (n \ge 1)$. The policy $\pi' = (\pi'_0, \pi'_1, \dots) \in \Pi$ is defined as follows. For $\forall t \ge 0$, $\forall h_t \in H_t$, define

$$\pi'_t(a|h_t) = \pi_{n+t}(a|i_0, a_0, i_1, a_1, \dots, a_{n-1}, h_t), \qquad a \in A.$$

Write $\pi' = \pi(i_0, a_0, \dots, i_{n-1}, a_{n-1})$ or $\pi' = \pi(\bar{h}_n)$.

The following facts stated here without proof are derived in [11].

LEMMA 1.1. *Let $n \ge N$, $i_0 \in S_0$, $\pi \in \Pi$, then*

$$\sum_{i \in S_0} P_\pi\{X_n = i|X_0 = i_0\} \le (1 - \alpha)^{[n/N]},$$

where $[X]$ denotes the greatest integer which does not exceed X.

LEMMA 1.2.

$$\sum_{t=0}^{\infty} \sum_{j \in S_0} P_\pi\{X_t = j|X_0 = i\} \le \frac{N}{\alpha} \qquad \textit{for } \forall i \in S_0,\ \forall\pi \in \Pi.$$

https://doi.org/10.1017/S0334270000000850 Published online by Cambridge University Press

PROOF. This follows immediately from the proof of Lemma 2.2 in [11].

Suppose $X_0 = i$ and let τ denote the smallest integer t such that $X_t = 0$. Let

$$V(\pi, i) = E_\pi\left[\sum_{t=0}^{\tau} r(X_t, \Delta_t) | X_0 = i\right], \qquad \pi \in \Pi,\ i \in S.$$

$V(\pi, i)$ is the expected total reward obtained using the policy π starting from i. Let $V^*(i) = \sup_{\pi\in\Pi} V(\pi, i)$, $i \in S$.

THEOREM 1.1 (Optimality equation).

$$V^*(i) = \sup_{a\in A}\left\{r(i, a) + \sum_{j\in S_0} q(j|i, a)V^*(j)\right\}, \qquad i \in S.$$

Let $\pi \in \Pi$, $h_n = (i_0, a_0, i_1, a_1, \dots, i_n) \in H_n$. If $P_\pi\{X_0 = i_0, \Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \dots, X_n = i_n | X_0 = i_0\} > 0$, then h_n is called a realizable history under the policy π.

Let

$$A^*(i) = \left\{a \in A | r(i, a) + \sum_{j\in S_0} q(j|i, a)V^*(j) = V^*(i)\right\}, \qquad i \in S.$$

THEOREM 1.2. *Let $i \in S$, $\pi \in \Pi$. Then a necessary and sufficient condition that $V(\pi, i) = V^*(i)$ is that for $\forall n \geq 0$, if $h_n = (i, a_0, \dots, i_n)$ is a realizable history under the policy π and $\pi_n(a|h_n) > 0$, then $a \in A^*(i_n)$.*

PROOF. Similar to the proof of Theorem 2.4 in [11].

By Theorem 1.2 we have

COROLLARY 1.1. (1) *If $f(i) \in A^*(i)$ for all $i \in S$, then $V(f^\infty, i) = V^*(i)$ for all $i \in S$.*

(2) *Let $i \in S$, $\pi = (\pi_0, \pi_1, \dots) \in \Pi$ and $V(\pi, i) = V^*(i)$. If $\pi_0(a|i) > 0$, then $a \in A^*(i)$.*

COROLLARY 1.2 (Bellman's optimality principle). *Let $i \in S$, $\pi \in \Pi$ and $V(\pi, i) = V^*(i)$. If $h_n = (i, a_0, i_1, a_1, \dots, i_n)$ $(n \geq 1)$ is a realizable history under the policy π, then $V(\pi(\overline{h}_n), i_n) = V^*(i_n)$.*

PROOF. Let $\pi(\overline{h}_n) = (\pi'_0, \pi'_1, \pi'_2, \dots)\ \forall m \geq 0$. Let $\tilde{h}_m = (i_n, \tilde{a}_0, \tilde{i}_1, \tilde{a}_1, \dots, \tilde{i}_m) \in H_m$ be a realizable history under the policy $\pi(\overline{h}_n)$ and $\pi'_m(a|\tilde{h}_m) > 0$. It is easy to see, $(i, a_0, i_1, a_1, \dots, i_n, \tilde{a}_0, \tilde{i}_1, \tilde{a}_1, \dots, \tilde{i}_m)$ is a realizable history under policy π. By the definition of $\pi(\overline{h}_n)$,

$$\pi_{n+m}(a|i, a_0, i_1, a_1, \dots, i_{n-1}, a_{n-1}, \tilde{h}_m) = \pi'_m(a|\tilde{h}_m) > 0,$$

by Theorem 1.2 (necessity), $a \in A^*(\tilde{i}_m)$. So, by Theorem 1.2 (sufficiency), $V(\pi(\overline{h}_n), i_n) = V^*(i_n)$.

THEOREM 1.3. *If f^∞ is optimal in Π_s^d (that is, $V(f^\infty, i) \geq V(g^\infty, i)$ for $\forall i \in S$, $\forall g^\infty \in \Pi_s^d$.), then f^∞ is also optimal in Π (that is, $V(f^\infty, i) \geq V(\pi, i)$ for $\forall i \in S$, $\forall \pi \in \Pi$).*

LEMMA 1.3. *Let S be finite, $f \in F$. If a set of numbers $\{V(i) : i \in S_0\}$ satisfies*

$$V(i) = \sum_{j \in S_0} q(j|i, f(i))V(j), \qquad i \in S_0,$$

then $V(i) \equiv 0$, $i \in S_0$.

Let $V_1, V_2 \in R^n (n \geq 1), V_i = (V_i(1), V_i(2), \dots, V_i(n))$, $i = 1, 2$. Define

$$V_1 \geq V_2 \Longleftrightarrow V_1(i) \geq V_2(i) \qquad \text{for} \quad i = 1, 2, \dots, n.$$
$$V_1 > V_2 \Longleftrightarrow V_1 \geq V_2 \qquad \text{and} \qquad V_1 \neq V_2.$$

2. The moment optimal problem

By the Cauchy criterion, we know that $\sum_{n=N+1}^{\infty} n^p (1-\alpha)^{[n-1/N]}$ is convergent for $p = 1, 2, \dots$. Let

$$D(\alpha, N, p) = \left[\sum_{n=N+1}^{\infty} n^p (1-\alpha)^{[n-1/N]}\right] + N^p, \qquad p = 1, 2, \dots.$$

LEMMA 2.1. *Let $i \in S_0$, $\pi \in \Pi$, $p = 1, 2, \dots$. Then*

$$E_\pi[\tau^p | X_0 = i] \leq D(\alpha, N, p).$$

PROOF. By Lemma 1.1,

$$\begin{aligned}
E_\pi[\tau^p|X_0=i] &= \sum_{n=1}^{\infty} n^p P_\pi\{\tau=n|X_0=i\} \\
&= \sum_{n=1}^{N} n^p P_\pi\{\tau=n|X_0=i\} + \sum_{n=N+1}^{\infty} n^p P_\pi\{\tau=n|X_0=i\} \\
&\le N^p \sum_{n=1}^{N} P_\pi\{\tau=n|X_0=i\} + \sum_{n=N+1}^{\infty} n^p P_\pi\{X_{n-1}\neq 0|X_0=i\} \\
&\le N^p P_\pi\{\tau\le N|X_0=i\} + \sum_{n=N+1}^{\infty} n^p (1-\alpha)^{[n-1/N]} \\
&\le D(\alpha, N, p).
\end{aligned}$$

So, by Lemma 2.1, when $i \in S_0, \pi \in \Pi,\ p = 1, 2, \ldots,$

$$\begin{aligned}
E_\pi\left[\left|\sum_{t=0}^{\tau} r(X_t, \Delta_t)\right|^p \ |X_0=i\right] &\le E_\pi[M^p(\tau+1)^p\ |X_0=i] \\
&\le (2M)^p E_\pi[\tau^p\ |X_0=i] \\
&\le (2M)^p D(\alpha, N, p). \qquad (2.1)
\end{aligned}$$

DEFINITION 2.1. Let

$$V_k(\pi, i) = E_\pi\left\{\left[\sum_{t=0}^{\tau} r(X_t, \Delta_t)\right]^k \ |X_0=i\right\},\quad i \in S, \pi \in \Pi,\ k = 1, 2, \ldots.$$

Let $V_0(\pi, i) \equiv 1,\ i \in S, \pi \in \Pi$.

It is easy to see, $V_k(\pi, 0) = 0, \pi \in \Pi,\ k = 1, 2, \ldots$.
Because $r(0, a) = 0$ and $q(0|0, a) = 1$, we have

$$V_k(\pi, i) = E_\pi\left\{\left[\sum_{t=0}^{\infty} r(X_t, \Delta_t)\right]^k \ |X_0=i\right\},\quad i \in S, \pi \in \Pi,\ k = 1, 2, \ldots.$$

THEOREM 2.1. *Let* $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$, $i \in S$, $k = 1, 2, \ldots$. *Then*

$$V_k(\pi, i) = \sum_{a\in A} \pi_0(a|i)\left\{R_k(i, a, \pi) + \sum_{j\in S} q(j|i, a) V_k(\pi(i, a), j)\right\},$$

https://doi.org/10.1017/S0334270000000850 Published online by Cambridge University Press

where

$$R_k(i,a,\pi) = \sum_{p=0}^{k-1} C_k^p r^{k-p}(i,a) \sum_{j\in S} q(j|i,a) V_p(\pi(i,a), j),$$
$$r^{k-p}(i,a) \equiv [r(i,a)]^{k-p}.$$

The definition of $\pi(i,a)$ can be found in Section 1.

PROOF. Let $i \in S_0$, $k = 1, 2, \ldots$. By the total mathematical expectation formula,

$$\begin{aligned}
V_k(\pi,i) &= E_\pi\left\{\left[\sum_{t=0}^{\infty} r(X_t,\Delta_t)\right]^k |X_0 = i\right\} \\
&= \sum_{a\in A}\pi_0(a|i)E_\pi\left\{\left[\sum_{t=0}^{\infty} r(X_t,\Delta_t)\right]^k |X_0 = i,\, \Delta_0 = a\right\} \\
&= \sum_{a\in A}\pi_0(a|i)E_\pi\left\{\left[r(i,a) + \sum_{t=1}^{\infty} r(X_t,\Delta_t)\right]^k |X_0 = i,\, \Delta_0 = a\right\} \\
&= \sum_{a\in A}\pi_0(a|i)\left[\sum_{p=0}^{k-1} C_k^p r^{k-p}(i,a)\sum_{j\in S} q(j|i,a)V_p(\pi(i,a),j) + \right. \\
&\qquad \left. \sum_{j\in S} q(j|i,a)V_k(\pi(i,a),j)\right] \\
&= \sum_{a\in A}\pi_0(a|i)\left[R_k(i,a,\pi) + \sum_{j\in S} q(j|i,a)V_k(\pi(i,a),j)\right].
\end{aligned}$$

The proposition is obviously true for i=0.

Let $M_l(\pi) = (-1)^{l+1}V_l(\pi), \pi \in \Pi, l = 0, 1, 2, \ldots$, where $V_l(\pi)$ is a vector and its i-th component is $V_l(\pi, i)$, $i \in S$.

Let $M^k(\pi) = (M_0(\pi), M_1(\pi), \ldots, M_k(\pi)), \pi \in \Pi,\ \ k = 1, 2, \ldots$.

DEFINITION 2.2. Let $k \geq 1$, $\pi_1, \pi_2 \in \Pi$. $M^k(\pi_1) > M^k(\pi_2) \Longleftrightarrow \exists n$, $1 \leq n \leq k$, such that $M_l(\pi_1) = M_l(\pi_2)$ for $l < n$ and $M_n(\pi_1) > M_n(\pi_2)$.

$$M^k(\pi_1) \geq M^k(\pi_2) \Longleftrightarrow M^k(\pi_1) > M^k(\pi_2) \quad \text{or} \quad M^k(\pi_1) = M^k(\pi_2).$$

DEFINITION 2.3. Let $k \geq 1$, $\pi^* \in \Pi$. If $M^k(\pi^*) \geq M^k(\pi)$ for $\forall\pi \in \Pi$, then π^* is called a k-moment optimal policy in Π.

If π^* is a k-moment optimal policy in Π for all $k \geq 1$, then π^* is called a moment-optimal policy in Π.

The set of the k-moment optimal policies in Π is denoted by $\Pi(k)(k \geq 1)$. Let $\Pi(0) = \Pi$. The set of the moment optimal policy in Π is denoted by $\Pi(\infty)$. Obviously, $\Pi(\infty) = \bigcap_{k=1}^{\infty} \Pi(k)$. It is easy to see by the definition that $\Pi(k) \subset \Pi(k-1)$, $k \geq 1$.

DEFINITION 2.4. Let $M_0^*(i) \equiv -1$, $\Pi(0,i) \equiv \Pi$, $i \in S$ and define $M_n^*(i)$ and $\Pi(n,i)(i \in S, n \geq 1)$ as follows. If $\Pi(n-1,i) \neq \emptyset$, then

$$M_n^*(i) = \sup_{\pi \in \Pi(n-1,i)} M_n(\pi, i),$$

$$\Pi(n,i) = \{\pi \in \Pi(n-1,i) | M_n(\pi,i) = M_n^*(i)\},$$

where $M_n(\pi,i) = (-1)^{n+1} V_n(\pi,i)$.

It is easy to see that $\Pi(n,0) \equiv \Pi$, $n = 0, 1, 2, \ldots$. By (2.1),

$$|M_n^*(i)| \leq (2M)^n D(\alpha, N, n), \qquad i \in S,\ n = 1, 2, \ldots. \tag{2.2}$$

DEFINITION 2.5. Let

$$R_n(i,a) = (-1)^{n+1} \sum_{k=0}^{n-1} C_n^k (-1)^{k+1} r^{n-k}(i,a) \sum_{j \in S} q(j|i,a) M_k^*(j),$$
$$i \in S,\quad a \in A,\quad n = 1, 2, \ldots.$$

Let $A_0^*(i) \equiv A$, $i \in S$ and define $A_n^*(i)$ $(i \in S, n \geq 1)$ as follows. If $A_{n-1}^*(i) \neq \emptyset$ and $\Pi(n-1,j) \neq \emptyset$ for all $j \in S$, then

$$A_n^*(i) = \left\{ a \in A_{n-1}^*(i) | R_n(i,a) + \sum_{j \in S} q(j|i,a) M_n^*(j) \right.$$
$$\left. = \sup_{\tilde{a} \in A_{n-1}^*(i)} \left[R_n(i,\tilde{a}) + \sum_{j \in S} q(j|i,\tilde{a}) M_n^*(j) \right] \right\}.$$

It is easy to see that $R_n(0,a) \equiv 0$, $a \in A$, $n = 1, 2, \ldots$; and $A_n^*(0) \equiv A$, $n = 0, 1, 2, \ldots$.

THEOREM 2.2. *Let* $k \geq 1$.

(1) *Let* $A_{k-1}^*(i) \neq \emptyset$ *for all* $i \in S$, *then*

$$\sup_{a \in A_{k-1}^*(i)} \left\{ R_k(i,a) + \sum_{j \in S} q(j|i,a) M_k^*(j) \right\} = M_k^*(i) \qquad \textit{for all} \quad i \in S.$$

https://doi.org/10.1017/S0334270000000850 Published online by Cambridge University Press

(2) *If* $f(i) \in A_k^*(i)$ *for all* $i \in S$, *then* $f^\infty \in \bigcap_{i \in S} \Pi(k, i)$.

(3) *Let* $A_{k-1}^*(j) \neq \emptyset$ *for all* $j \in S$. *Let* $i \in S, \pi \in \Pi(k, i)$. *If* $\pi_0(a|i) > 0$, *then* $a \in A_k^*(i)$.

(4) *Let* $A_{k-1}^*(j) \neq \emptyset$ *for all* $j \in S$. *Let* $i \in S, \pi \in \Pi(k, i)$. *If* $h_n = (i, a_0, i_1, a_1, \dots, i_n) \in H_n (n \geq 1)$ *is a realizable history under the policy* π, *then* $\pi(\overline{h}_n) \in \Pi(k, i_n)$.

PROOF. (Apply induction to k). We know that proposition (Theorem 2.2) is true for $k = 1$ by Theorem 1.1, Corollary 1.1 and Corollary 1.2.

Inductive hypothesis I: the proposition (Theorem 2.2) is true for $1 \leq k \leq l - 1$.

(1) Let $A_{l-1}^*(i) \neq \emptyset$ for all $i \in S$. We take $f(i) \in A_{l-1}^*(i)$ for $\forall i \in S$. By the inductive hypothesis I and (2) in Theorem 2.2, $f^\infty \in \bigcap_{i \in S} \Pi(l-1, i)$. So $\Pi(l-1, i) \neq \emptyset$ for all $i \in S$.

For $\forall i \in S, \forall \pi \in \Pi(l - 1, i)$, by Theorem 2.1,

$$M_l(\pi, i) = \sum_{a \in A} \pi_0(a|i) \left\{ (-1)^{l+1} R_l(i, a, \pi) + \sum_{j \in S} q(j|i, a) M_l(\pi(i, a), j) \right\}.$$

By the inductive hypothesis I and (4) in Theorem 2.2, $\pi(i, a) \in \Pi(l - 1, j)$ when $\pi_0(a|i)q(j|i, a) > 0$. So

$$\begin{aligned}
&\sum_{\substack{a \in A \\ \pi_0(a|i) > 0}} \pi_0(a|i)(-1)^{l+1} R_l(i, a, \pi) \\
&= \sum_{\substack{a \in A \\ \pi_0(a|i) > 0}} \pi_0(a|i)(-1)^{l+1} \sum_{p=0}^{l-1} C_l^p r^{l-p}(i, a) \sum_{\substack{j \in S \\ q(j|i,a) > 0}} q(j|i, a) M_p(\pi(i, a), j)(-1)^{p+1} \\
&= \sum_{\substack{a \in A \\ \pi_0(a|i) > 0}} \pi_0(a|i)(-1)^{l+1} \sum_{p=0}^{l-1} C_l^p (-1)^{p+1} r^{l-p}(i, a) \sum_{\substack{j \in S \\ q(j|i,a) > 0}} q(j|i, a) M_p^*(j) \\
&= \sum_{\substack{a \in A \\ \pi_0(a|i) > 0}} \pi_0(a|i) R_l(i, a),
\end{aligned}$$

and

$$\sum_{\substack{a \in A \\ \pi_0(a|i) > 0}} \pi_0(a|i) \sum_{\substack{j \in S \\ q(j|i,a) > 0}} q(j|i, a) M_l(\pi(i, a), j) \leq \sum_{\substack{a \in A \\ \pi_0(a|i) > 0}} \pi_0(a|i) \sum_{\substack{j \in S \\ q(j|i,a) > 0}} q(j|i, a) M_l^*(j).$$

That is,

$$M_l(\pi, i) \leq \sum_{a \in A} \pi_0(a|i) \left\{ R_l(i, a) + \sum_{j \in S} q(j|i, a) M_l^*(j) \right\}.$$

By the inductive hypothesis I and (3) in Theorem 2.2, $a \in A^*_{l-1}(i)$ when $\pi_0(a|i) > 0$. Therefore we have

$$M_l(\pi, i) \le \sup_{a \in A^*_{l-1}(i)} \left\{ R_l(i,a) + \sum_{j \in S} q(j|i,a) M^*_l(j) \right\}.$$

By definition,

$$M^*_l(i) \le \sup_{a \in A^*_{l-1}(i)} \left\{ R_l(i,a) + \sum_{j \in S} q(j|i,a) M^*_l(j) \right\}, \qquad i \in S. \tag{2.3}$$

For each $\epsilon > 0$, we take $f(i) \in A^*_{l-1}(i)$ for $\forall i \in S$ such that

$$\begin{aligned} R_l(i, f(i)) + \sum_{j \in S} q(j|i, f(i)) M^*_l(j) &\ge \sup_{a \in A^*_{l-1}(i)} \left\{ R_l(i,a) + \sum_{j \in S} q(j|i,a) M^*_l(j) \right\} - \frac{\epsilon\alpha}{N} \\ &\ge M^*_l(i) - \frac{\epsilon\alpha}{N}. \end{aligned} \tag{2.4}$$

PROPOSITION A1. Let $i \in S_0$. Then

$$\begin{aligned} &\sum_{n=0}^{m-1} \sum_{i_n \in S_0} P_{f^\infty}\{X_n = i_n | X_0 = i\} R_l(i_n, f(i_n)) + \sum_{i_m \in S_0} P_{f^\infty}\{X_m = i_m | X_0 = i\} M^*_l(i_m) \\ &\quad \ge M^*_l(i) - \frac{\epsilon\alpha}{N} \sum_{n=0}^{m-1} \sum_{i_n \in S_0} P_{f^\infty}\{X_n = i_n | X_0 = i\}, \qquad m = 1, 2, \ldots. \end{aligned}$$

PROOF OF PROPOSITION A1. This follows immediately on applying induction to m (or see the proof of (2.2) in [11]).

PROPOSITION A2. If $g(i) \in A^*_{l-1}(i)$ for all $i \in S$, then

$$\begin{aligned} M_l(g^\infty, i) &= \sum_{n=0}^{m-1} \sum_{i_n \in S} P_{g^\infty}\{X_n = i_n | X_0 = i\} R_l(i_n, g(i_n)) \\ &\quad + \sum_{i_m \in S} P_{g^\infty}\{X_m = i_m | X_0 = i\} M_l(g^\infty, i_m) \qquad i \in S, \quad m = 1, 2, \ldots. \end{aligned}$$

PROOF OF PROPOSITION A2. By inductive hypothesis I and (2) in Theorem 2.2, $g^\infty \in \bigcap_{i \in S} \Pi(l-1, i)$. By Theorem 2.1,

$$\begin{aligned} M_l(g^\infty, i) &= (-1)^{l+1} R_l(i, g(i), g^\infty) + \sum_{j \in S} q(j|i, g(i)) M_l(g^\infty, j) \\ &= R_l(i, g(i)) + \sum_{j \in S} q(j|i, g(i)) M_l(g^\infty, j), \qquad i \in S. \end{aligned} \tag{2.5}$$

By (2.5), we can prove that Proposition A2 is true by applying induction to m.

By Propositions A1, A2 and Lemma 1.2,

$$M_l(f^\infty, i) \geq M_l^*(i) - \epsilon + \sum_{i_m \in S_0} P_{f^\infty}\{X_m = i_m | X_0 = i\}(M_l(f^\infty, i_m) - M_l^*(i_m)),$$
$$i \in S_0, \quad m = 1, 2, \ldots .$$

By Lemma 1.1 and (2.1), (2.2)

$$M_l(f^\infty, i) \geq M_l^*(i) - \epsilon - 2(1-\alpha)^{[m/N]}(2M)^l D(\alpha, N, l), \quad i \in S_0, m = N, N+1, \ldots .$$

Let $m \to \infty$. We have $M_l(f^\infty, i) \geq M_l^*(i) - \epsilon, i \in S_0$.

So, by (2.5), (2.4)

$$\begin{aligned} M_l^*(i) \geq M_l(f^\infty, i) &= R_l(i, f(i)) + \sum_{j \in S} q(j|i, f(i)) M_l(f^\infty, j) \\ &\geq R_l(i, f(i)) + \sum_{j \in S} q(j|i, f(i))[M_l^*(j) - \epsilon] \\ &= R_l(i, f(i)) + \sum_{j \in S} q(j|i, f(i)) M_l^*(j) - \epsilon \\ &\geq \sup_{a \in A_{l-1}^*(i)} \left\{ R_l(i, a) + \sum_{j \in S} q(j|i, a) M_l^*(j) \right\} - \frac{\epsilon\alpha}{N} - \epsilon, \quad i \in S. \end{aligned}$$

That is,

$$M_l^*(i) \geq \sup_{a \in A_{l-1}^*(i)} \left\{ R_l(i, a) + \sum_{j \in S} q(j|i, a) M_l^*(j) \right\} - \frac{\epsilon\alpha}{N} - \epsilon, \qquad i \in S.$$

If we let $\epsilon \to 0$, we see that (1) is true for $k = l$ combining (2.3).

(2) Let $f(i) \in A_l^*(i)$ for all $i \in S$. Obviously $f(i) \in A_{l-1}^*(i)$ for all $i \in S$. By the definition of $A_l^*(i)$,

$$R_l(i, f(i)) + \sum_{j \in S} q(j|i, f(i)) M_l^*(j) = \sup_{a \in A_{l-1}^*(i)} \left\{ R_l(i, a) + \sum_{j \in S} q(j|i, a) M_l^*(j) \right\}, \quad i \in S.$$

We have from the above proof of (1) that

$$M_l(f^\infty, i) \geq M_l^*(i), \qquad i \in S. \tag{2.6}$$

By inductive hypothesis I and (2) in Theorem 2.2, $f^\infty \in \bigcap_{i \in S} \Pi(l-1, i)$. So $M_l(f^\infty, i) \leq M_l^*(i)$, $i \in S$. From (2.6) we have $f^\infty \in \bigcap_{i \in S} \Pi(l, i)$, that is, (2) is true for $k = l$.

https://doi.org/10.1017/S0334270000000850 Published online by Cambridge University Press

(3) Let $A^*_{l-1}(j) \neq \emptyset$ for all $j \in S$. Let $i \in S, \pi \in \Pi(l,i)$. Obviously $\pi \in \Pi(l-1,i)$. By inductive hypothesis I and (3) in Theorem 2.2, $a \in A^*_{l-1}(i)$ when $\pi_0(a|i) > 0$. So

$$\pi_0(a|i)\left\{R_l(i,a) + \sum_{j\in S} q(j|i,a)M^*_l(j)\right\} \leq$$
$$\pi_0(a|i) \sup_{\bar{a}\in A^*_{l-1}(i)} \left\{R_l(i,\bar{a}) + \sum_{j\in S} q(j|i,\bar{a})M^*_l(j)\right\}, \quad a \in A. \tag{2.7}$$

We know from the above proof of (1) that

$$\begin{aligned} M^*_l(i) = M_l(\pi,i) &\leq \sum_{a\in A} \pi_0(a|i)\left\{R_l(i,a) + \sum_{j\in S} q(j|i,a)M^*_l(j)\right\} \\ &\leq \sum_{a\in A} \pi_0(a|i) \sup_{\bar{a}\in A^*_{l-1}(i)} \left\{R_l(i,\bar{a}) + \sum_{j\in S} q(j|i,\bar{a})M^*_l(j)\right\} \\ &= \sup_{a\in A^*_{l-1}(i)} \left\{R_l(i,a) + \sum_{j\in S} q(j|i,a)M^*_l(j)\right\} = M^*_l(i). \end{aligned}$$

So

$$\sum_{a\in A} \pi_0(a|i)\left\{R_l(i,a) + \sum_{j\in S} q(j|i,a)M^*_l(j)\right\}$$
$$= \sum_{a\in A} \pi_0(a|i) \sup_{\bar{a}\in A^*_{l-1}(i)} \left\{R_l(i,\bar{a}) + \sum_{j\in S} q(j|i,\bar{a})M^*_l(j)\right\}. \tag{2.8}$$

By (2.8) and (2.7),

$$\pi_0(a|i)\left\{R_l(i,a) + \sum_{j\in S} q(j|i,a)M^*_l(j)\right\}$$
$$= \pi_0(a|i) \sup_{\bar{a}\in A^*_{l-1}(i)} \left\{R_l(i,\bar{a}) + \sum_{j\in S} q(j|i,\bar{a})M^*_l(j)\right\}, \qquad a \in A.$$

Therefore, when $\pi_0(a|i) > 0$, we have $a \in A^*_{l-1}(i)$ and

$$R_l(i,a) + \sum_{j\in S} q(j|i,a)M^*_l(j) = \sup_{\bar{a}\in A^*_{l-1}(i)} \left\{R_l(i,\bar{a}) + \sum_{j\in S} q(j|i,\bar{a})M^*_l(j)\right\},$$

that is, $a \in A^*_l(i)$. So (3) is true for $k = l$.

(4) Let $A^*_{l-1}(j) \neq \emptyset$ for all $j \in S$. Let $i \in S, \pi \in \Pi(l, i)$ and $h_n = (i, a_0, i_1, a_1, \dots, i_n)$ $(n \geq 1)$ be a realizable history under the policy π. We shall prove that $\pi(\overline{h}_n) \in \Pi(l, i_n)$.

(Applying induction to n). Let $n = 1$ and $h_1 = (i, a_0, i_1)$ be a realizable history under the policy π. Obviously $\pi \in \Pi(l-1, i)$. We have from the above proofs of (1) and (3),

$$\begin{aligned} M^*_l(i) = M_l(\pi, i) &= \sum_{a\in A} \pi_0(a|i) \left\{ R_l(i,a) + \sum_{j\in S} q(j|i,a) M_l(\pi(i,a), j) \right\} \\ &\leq \sum_{a\in A} \pi_0(a|i) \left\{ R_l(i,a) + \sum_{j\in S} q(j|i,a) M^*_l(j) \right\} \\ &\leq M^*_l(i). \end{aligned}$$

Therefore

$$\sum_{a\in A} \pi_0(a|i) \sum_{j\in S} q(j|i,a) M_l(\pi(i,a), j) = \sum_{a\in A} \pi_0(a|i) \sum_{j\in S} q(j|i,a) M^*_l(j). \quad (2.9)$$

By inductive hypothesis I and (4) in Theorem 2.2, $\pi(i, a) \in \Pi(l-1, j)$ when $\pi_0(a|i)q(j|i,a) > 0$. So

$$\pi_0(a|i)q(j|i,a)M_l(\pi(i,a), j) \leq \pi_0(a|i)q(j|i,a)M^*_l(j), \quad a \in A, j \in S. \quad (2.10)$$

By (2.9) and (2.10),

$$\pi_0(a|i)q(j|i,a)M_l(\pi(i,a), j) = \pi_0(a|i)q(j|i,a)M^*_l(j), \quad a \in A, j \in S.$$

So, when $\pi_0(a_0|i)q(i_1|i, a_0) > 0$, we have $\pi(i, a_0) \in \Pi(l-1, i_1)$ and $M_l(\pi(i, a_0), i_1) = M^*_l(i_1)$, that is, $\pi(\overline{h}_1) \in \Pi(l, i_1)$. The proposition is true for $n = 1$.

Suppose the proposition is true for n. Let $h_{n+1} = (i, a_0, i_1, a_1, \dots, i_{n+1})$ be a realizable history under the policy π. It is easy to see that $h_n = (i, a_0, i_1, a_1, \dots, i_n)$ is also a realizable history under the policy π. By the supposition that $\pi(\overline{h}_n) \in \Pi(l, i_n)$, it is also easy to see that $\pi_n(a_n|h_n)q(i_{n+1}|i_n, a_n) > 0$, that is, (i_n, a_n, i_{n+1}) is a realizable history under the policy $\pi(\overline{h}_n)$. Applying the result for $n = 1$, we have $\pi(\overline{h}_{n+1}) = \pi(\overline{h}_n)(i_n, a_n) \in \Pi(l, i_{n+1})$, that is, the proposition is also true for $n + 1$. So (4) is true for $k = l$.

COROLLARY 2.1. *Let $k \geq 1$, $A^*_{k-1}(j) \neq \emptyset$ for all $j \in S$. Let $i \in S$, $\Pi(k, i) \neq \emptyset$. Then $A^*_k(i) \neq \emptyset$.*

PROOF. This follows immediately from Theorem 2.2(3).

https://doi.org/10.1017/S0334270000000850 Published online by Cambridge University Press

COROLLARY 2.2. *Let $k \geq 1$. If $A_k^*(i) \neq \emptyset$ for all $i \in S$, then $\bigcap_{i \in S} \Pi(k, i) \neq \emptyset$.*

PROOF. This follows immediately from Theorem 2.2(2).

COROLLARY 2.3. *Let $n \geq 1$. Then*

$$\Pi(n, j) \neq \emptyset \text{ for all } j \in S \Longleftrightarrow A_n^*(j) \neq \emptyset \text{ for all } j \in S.$$

PROOF. ($\Leftarrow$) This follows immediately from Corollary 2.2.

($\Rightarrow$) (Apply induction to n). The proposition is true for $n = 1$ by Corollary 2.1.

Suppose it is true for n. Let $\Pi(n+1, j) \neq \emptyset$ for all $j \in S$. Obviously $\Pi(n, j) \neq \emptyset$ for all $j \in S$. So $A_n^*(j) \neq \emptyset$ for all $j \in S$. By Corollary 2.1, $A_{n+1}^*(j) \neq \emptyset$ for all $j \in S$. That is, the proposition is also true for $n + 1$.

THEOREM 2.3. *Let $k \geq 0$, $A_k^*(i) \neq \emptyset$ for all $i \in S$. Then $\forall \epsilon > 0$, $\exists f^\infty$ such that $f(i) \in A_k^*(i)$ for all $i \in S$ and*

$$M_{k+1}(f^\infty, i) \geq M_{k+1}^*(i) - \epsilon, \qquad i \in S.$$

PROOF. The case for $k = 0$ corresponds to Theorem 2.2 in [11]. We know that the proposition is true for $k \geq 1$ from the proof of Theorem 2.2(1).

THEOREM 2.4. *Let $k \geq 1$, $A_{k-1}^*(j) \neq \emptyset$ for all $j \in S$. Let $i \in S$. Then $\pi \in \Pi(k, i) \Longleftrightarrow \forall n \geq 0$, if $h_n = (i, a_0, i_1, a_1, \ldots, i_n)$ is a realizable history under the policy π and $\pi_n(a|h_n) > 0$, then $a \in A_k^*(i_n)$.*

PROOF. ($\Rightarrow$) Let $n \geq 1$. By Theorem 2.2(4), $\pi(\overline{h}_n) \in \Pi(k, i_n)$. Let $\pi(\overline{h}_n) = (\pi_0', \pi_1', \pi_2', \ldots)$. It is easy to see that $\pi_0'(a|i_n) = \pi_n(a|h_n)$, $a \in A$. By Theorem 2.2(3), $a \in A_k^*(i_n)$ when $\pi_n(a|h_n) > 0$.

Let $n = 0$. By Theorem 2.2(3), $a \in A_k^*(i)$ when $\pi_0(a|i) > 0$.

($\Leftarrow$) (Apply induction to k). The proposition is true for $k = 1$ by Theorem 1.2. Suppose the proposition is true for $1 \leq k \leq l - 1$. We consider the case that $k = l$.

Let $A_{l-1}^*(j) \neq \emptyset$ for all $j \in S$ and let $i \in S$. By the inductive hypothesis and the sufficiency supposition, $\pi \in \Pi(l-1, i)$. We have from the proof of Theorem 2.2(1) that

$$M_l(\pi, i) = \sum_{a \in A} \pi_0(a|i) \left\{ R_l(i, a) + \sum_{j \in S} q(j|i, a) M_l(\pi(i, a), j) \right\}. \tag{2.11}$$

Let $m \geq 0$. By Theorem 2.2(4), when $P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \ldots, X_{m+1} = i_{m+1} | X_0 = i\} > 0$, we have $\pi(i, a_0, i_1, a_1, \ldots, i_m, a_m) \in \Pi(l-1, i_{m+1})$. So, by (2.11), when $P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \ldots, X_{m+1} = i_{m+1} | X_0 = i\} > 0$, we have

$$M_l(\pi(i, a_0, i_1, a_1, \ldots, i_m, a_m), i_{m+1}) = \sum_{a_{m+1}\in A} \pi_{m+1}(a_{m+1}|i, a_0, i_1, a_1, \ldots, i_{m+1}) \Big\{ R_l(i_{m+1}, a_{m+1}) + \sum_{i_{m+2}\in S} q(i_{m+2}|i_{m+1}, a_{m+1}) M_l(\pi(i, a_0, i_1, a_1, \ldots, i_m, a_m)(i_{m+1}, a_{m+1}), i_{m+2}) \Big\}.$$

Therefore, we have

$$\begin{aligned} &\sum_{\substack{a_0\in A, i_1\in S,\\ a_1\in A,\ldots, i_{m+1}\in S}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \ldots, X_{m+1} = i_{m+1}|X_0 = i\} \times M_l(\pi(i, a_0, i_1, a_1, \ldots, i_m, a_m), i_{m+1}) \\ &= \sum_{\substack{a_0\in A, i_1\in S,\\ a_1\in A,\ldots, i_{m+1}\in S}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \ldots, X_{m+1} = i_{m+1}|X_0 = i\} \times \Big\{ \sum_{a_{m+1}\in A} \pi_{m+1}(a_{m+1}|i, a_0, i_1, a_1, \ldots, i_{m+1}) R_l(i_{m+1}, a_{m+1}) \\ &\quad + \sum_{\substack{a_{m+1}\in A,\\ i_{m+2}\in S}} \pi_{m+1}(a_{m+1}|i, a_0, i_1, a_1, \ldots, i_{m+1}) q(i_{m+2}|i_{m+1}, a_{m+1}) \times M_l(\pi(i, a_0, i_1, a_1, \ldots, i_m, a_m, i_{m+1}, a_{m+1}), i_{m+2}) \Big\} \\ &= \sum_{\substack{i_{m+1}\in S\\ a_{m+1}\in A}} P_\pi\{X_{m+1} = i_{m+1}, \Delta_{m+1} = a_{m+1}|X_0 = i\} R_l(i_{m+1}, a_{m+1}) \\ &\quad + \sum_{\substack{a_0\in A, i_1\in S,\\ a_1\in A,\ldots, i_{m+2}\in S}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \ldots, X_{m+2} = i_{m+2}|X_0 = i\} \times M_l(\pi(i, a_0, i_1, a_1, \ldots, i_{m+1}, a_{m+1}), i_{m+2}), \quad m \geq 0. \end{aligned} \tag{2.12}$$

By (2.11) and (2.12), it is easy to prove by induction that

$$\begin{aligned} M_l(\pi, i) &= \sum_{n=0}^{m} \sum_{i_n\in S, a_n\in A} P_\pi\{X_n = i_n, \Delta_n = a_n|X_0 = i\} R_l(i_n, a_n) \\ &\quad + \sum_{\substack{a_0\in A, i_1\in S,\\ a_1\in A,\ldots, i_{m+1}\in S}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \ldots, X_{m+1} = i_{m+1}|X_0 = i\} \times M_l(\pi(i, a_0, i_1, a_1, \ldots, i_m, a_m), i_{m+1}), \qquad m = 0, 1, 2, \ldots. \end{aligned}$$

https://doi.org/10.1017/S0334270000000850 Published online by Cambridge University Press

By the sufficiency supposition, $a \in A_l^*(i)$ when $\pi_0(a|i) > 0$. So, by Theorem 2.2(1),

$$M_l^*(i) = \sum_{a\in A} \pi_0(a|i) \left\{ R_l(i,a) + \sum_{j\in S} q(j|i,a) M_l^*(j) \right\}, \tag{2.13}$$

Let $m \geq 0$. By the sufficiency supposition, when $P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \dots, X_{m+1} = i_{m+1}|X_0 = i\} > 0$, if $\pi_{m+1}(a_{m+1}|i, a_0, i_1, a_1, \dots, i_{m+1}) > 0$, then $a_{m+1} \in A_l^*(i_{m+1})$. So, by Theorem 2.2(1), when $P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \dots, X_{m+1} = i_{m+1}|X_0 = i\} > 0$, we have

$$M_l^*(i_{m+1}) = \sum_{a_{m+1}\in A} \pi_{m+1}(a_{m+1}|i, a_0, i_1, a_1, \dots, i_{m+1}) \left\{ R_l(i_{m+1}, a_{m+1}) + \sum_{j\in S} q(j|i_{m+1}, a_{m+1}) M_l^*(j) \right\}.$$

Therefore, we have

$$\begin{aligned}
&\sum_{\substack{a_0\in A, i_1\in S,\\ a_1\in A,\dots,i_{m+1}\in S}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \dots, X_{m+1} = i_{m+1}|X_0 = i\} M_l^*(i_{m+1}) \\
&\quad = \sum_{\substack{a_0\in A, i_1\in S,\\ a_1\in A,\dots,i_{m+1}\in S}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \dots, X_{m+1} = i_{m+1}|X_0 = i\} \\
&\qquad + \sum_{a_{m+1}\in A} \pi_{m+1}(a_{m+1}|i, a_0, i_1, a_1, \dots, i_{m+1}) \left\{ R_l(i_{m+1}, a_{m+1}) + \sum_{j\in S} q(j|i_{m+1}, a_{m+1}) M_l^*(j) \right\} \\
&\quad = \sum_{\substack{i_{m+1}\in S\\ a_{m+1}\in A}} P_\pi\{X_{m+1} = i_{m+1}, \Delta_{m+1} = a_{m+1}|X_0 = i\} R_l(i_{m+1}, a_{m+1}) \\
&\qquad + \sum_{\substack{a_0\in A, i_1\in S,\\ a_1\in A,\dots,i_{m+2}\in S}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \dots, X_{m+2} = i_{m+2}|X_0 = i\} M_l^*(i_{m+2}), \\
&\qquad m \geq 0.
\end{aligned} \tag{2.14}$$

By (2.13) and (2.14), it is easy to prove by induction that

$$\begin{aligned}
M_l^*(i) &= \sum_{n=0}^{m} \sum_{i_n\in S, a_n\in A} P_\pi\{X_n = i_n, \Delta_n = a_n|X_0 = i\} R_l(i_n, a_n) \\
&\quad + \sum_{\substack{a_0\in A, i_1\in S,\\ a_1\in A,\dots,i_{m+1}\in S}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \dots, X_{m+1} = i_{m+1}|X_0 = i\} M_l^*(i_{m+1}), \\
&\qquad m = 0, 1, 2, \dots.
\end{aligned}$$

So, when $i \in S_0$, by (2.1), (2.2) and Lemma 1.1,

$$\begin{aligned} |M_l(\pi,i) - M_l^*(i)| \le & \sum_{\substack{a_0\in A, i_1\in S,\\ a_1\in A,\dots,i_m\in S, a_m\in A, i_{m+1}\in S_0}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \dots, \\ & X_{m+1} = i_{m+1} | X_0 = i\} 2(2M)^l D(\alpha, N, l) \\ & = \sum_{i_{m+1}\in S_0} P_\pi\{X_{m+1} = i_{m+1} | X_0 = i\} 2(2M)^l D(\alpha, N, l) \\ & \le (1-\alpha)^{[m+1/N]} 2(2M)^l D(\alpha, N, l), \qquad m = N, N+1, \dots . \end{aligned}$$

Let $m \to \infty$. We have $M_l(\pi, i) = M_l^*(i)$. So $\pi \in \Pi(l, i)$ (if $i = 0$, then $\pi \in \Pi = \Pi(l, 0)$ obviously). The proposition is also true for $k = l$.

Obviously Theorem 2.4 is an extension of Theorem 1.2.

THEOREM 2.5. *Let* $k \ge 0$. *Then* $\Pi(k) = \bigcap_{i\in S} \Pi(k, i)$.

PROOF. (Apply induction to k.) The proposition is true for $k = 0$ obviously. Suppose the proposition is true for $0 \le k \le l - 1$.

Let $\pi \in \Pi(l)$. It is easy to see that $\pi \in \Pi(l - 1)$. By the inductive hypothesis, $\pi \in \bigcap_{i\in S} \Pi(l - 1, i)$. By Corollary 2.3, $A_{l-1}^*(i) \ne \emptyset$ for all $i \in S$. By Theorem 2.3, $\forall \epsilon > 0, \exists f^\infty$ such that $f(i) \in A_{l-1}^*(i)$ for all $i \in S$ and

$$M_l(f^\infty, i) \ge M_l^*(i) - \epsilon, \qquad i \in S.$$

By Theorem 2.2(2) and the inductive hypothesis, $f^\infty \in \Pi(l - 1)$. Since $\pi, f^\infty \in \Pi(l - 1)$, therefore $M^{l-1}(\pi) = M^{l-1}(f^\infty)$. Since $\pi \in \Pi(l)$, therefore $M^l(\pi) \ge M^l(f^\infty)$. Hence $M_l(\pi, i) \ge M_l(f^\infty, i)$ for all $i \in S$, that is,

$$M_l(\pi, i) \ge M_l^*(i) - \epsilon, \qquad i \in S.$$

Let $\epsilon \to 0$. We have $M_l(\pi, i) = M_l^*(i)$ for all $i \in S$. So $\pi \in \bigcap_{i\in S} \Pi(l, i)$, that is, $\Pi(l) \subset \bigcap_{i\in S} \Pi(l, i)$.

Let $\pi \in \bigcap_{i\in S} \Pi(l, i)$. It is easy to see that $\pi \in \bigcap_{i\in S} \Pi(l - 1, i)$. By the inductive hypothesis, $\pi \in \Pi(l - 1)$. Choose any $\tilde\pi \in \Pi$. Obviously $M^{l-1}(\pi) \ge M^{l-1}(\tilde\pi)$. If $M^{l-1}(\pi) > M^{l-1}(\tilde\pi)$, then

$$M^l(\pi) > M^l(\tilde\pi). \tag{2.15}$$

If $M^{l-1}(\pi) = M^{l-1}(\tilde\pi)$, then $\tilde\pi \in \Pi(l - 1)$. By the inductive hypothesis, $\tilde\pi \in \bigcap_{i\in S} \Pi(l - 1, i)$. Since $\pi \in \bigcap_{i\in S} \Pi(l, i)$, we have $M_l(\pi) \ge M_l(\tilde\pi)$. Hence

$$M^l(\pi) \ge M^l(\tilde\pi). \tag{2.16}$$

https://doi.org/10.1017/S0334270000000850 Published online by Cambridge University Press

By (2.15) and (2.16), $M^l(\pi) \geq M^l(\tilde{\pi})$. Therefore $\pi \in \Pi(l)$, that is, $\bigcap_{i\in S}\Pi(l,i) \subset \Pi(l)$.

To sum up, we know that the proposition is true for $k = l$.

THEOREM 2.6. *Let $k \geq 1$. Then $\pi \in \Pi(k) \Longleftrightarrow \forall n \geq 0$ if $h_n = (i_0, a_0, i_1, a_1, \ldots, i_n)$ is a realizable history under the policy π and $\pi_n(a|h_n) > 0$, then $a \in A_k^*(i_n)$.*

PROOF. ($\Rightarrow$) Let $\pi \in \Pi(k)$. By Theorem 2.5, $\pi \in \bigcap_{i\in S}\Pi(k,i)$. By Corollary 2.3, $A_k^*(i) \neq \emptyset$ for all $i \in S$. Obviously $\pi \in \Pi(k, i_0)$. By Theorem 2.4,if $h_n = (i_0, a_0, i_1, a_1, \ldots, i_n)$ $(n \geq 0)$ is a realizable history under the policy π and $\pi_n(a|h_n) > 0$, then $a \in A_k^*(i_n)$.

($\Leftarrow$) Choose any $i \in S$. We take $a \in A$ such that $\pi_0(a|i) > 0$. By the sufficiency supposition, $a \in A_k^*(i)$. So $A_k^*(j) \neq \emptyset$ for all $j \in S$. By the sufficiency supposition and Theorem 2.4, $\pi \in \Pi(k, i)$ for all $i \in S$. By Theorem 2.5, $\pi \in \Pi(k)$.

Obviously this theorem is an extension of Theorem 2.4 in [11].

COROLLARY 2.4. *$\pi \in \Pi(\infty) \Longleftrightarrow \forall n \geq 0$, if $h_n = (i_0, a_0, i_1, a_1, \ldots, i_n)$ is a realizable history under the policy π and $\pi_n(a|h_n) > 0$, then $a \in \bigcap_{k=1}^{\infty} A_k^*(i_n)$.*

PROOF. This follows immediately from Theorem 2.6.

THEOREM 2.7. (1) *Let $k \geq 1$. If $\Pi(k) \neq \emptyset$, then $\exists f^\infty \in \Pi(k)$.*
(2) *If $\Pi(\infty) \neq \emptyset$, then $\exists f^\infty \in \Pi(\infty)$.*

PROOF. (1) By Theorem 2.5 and Corollary 2.3, $A_k^*(i) \neq \emptyset$ for all $i \in S$. We take $f(i) \in A_k^*(i)$ for all $i \in S$. By Theorem 2.2(2) and Theorem 2.5, $f^\infty \in \Pi(k)$.

(2) We take $\pi \in \Pi(\infty)$ and $\forall i \in S$ take $a \in A$ such that $\pi_0(a|i) > 0$. By Corollary 2.4, $a \in \bigcap_{k=1}^{\infty} A_k^*(i)$. That is, $\bigcap_{k=1}^{\infty} A_k^*(i) \neq \emptyset$ for all $i \in S$. We take $f(i) \in \bigcap_{k=1}^{\infty} A_k^*(i)$ for all $i \in S$. By Corollary 2.4, $f^\infty \in \Pi(\infty)$.

THEOREM 2.8. (1) *Let $k \geq 1$. If f^∞ is a k-moment optimal policy in Π_s^d (that is, $M^k(f^\infty) \geq M^k(g^\infty)$ for all $g^\infty \in \Pi_s^d$), then $f^\infty \in \Pi(k)$.*
(2) *If f^∞ is a moment optimal policy in Π_s^d, then $f^\infty \in \Pi(\infty)$.*

PROOF. (1) (Apply induction to k.) The proposition is true for $k = 1$ by Theorem 1.3 and Theorem 2.5. Suppose the proposition is true for $1 \leq k \leq l - 1$.

Let f^∞ be a l-moment optimal policy in Π_s^d. It is easy to see that f^∞ is a $(l-1)$-moment optimal policy in Π_s^d. By the inductive hypothesis and Theorem 2.5,

$f^\infty \in \Pi(l-1) = \bigcap_{i\in S}\Pi(l-1,i)$. By Corollary 2.3, $A^*_{l-1}(i) \neq \emptyset$ for all $i \in S$. By Theorem 2.3, $\forall\epsilon > 0$, $\exists g^\infty$ such that $g(i) \in A^*_{l-1}(i)$ for all $i \in S$ and

$$M_l(g^\infty, i) \geq M^*_l(i) - \epsilon, \qquad i \in S.$$

By Theorem 2.2(2) and Theorem 2.5, $g^\infty \in \Pi(l-1)$. So $M^{l-1}(g^\infty) = M^{l-1}(f^\infty)$. By the supposition, $M^l(f^\infty) \geq M^l(g^\infty)$. So $M_l(f^\infty, i) \geq M_l(g^\infty, i)$, $i \in S$. Hence

$$M_l(f^\infty, i) \geq M^*_l(i) - \epsilon, \qquad i \in S.$$

Let $\epsilon \to 0$. We have $M_l(f^\infty, i) = M^*_l(i)$, $i \in S$. By Theorem 2.5, $f^\infty \in \bigcap_{i\in S}\Pi(l,i) = \Pi(l)$. That is, the proposition is true for $k = l$. The proof of (1) is complete.

(2) This follows immediately from (1).

Theorems 2.7 and 2.8 state that the problems of the existence and calculation of a k-moment optimal policy (or a moment optimal policy) in Π can be changed into the same problems in Π^d_s.

THEOREM 2.9. *If A is nonempty and finite, then $\exists f^\infty \in \Pi(\infty)$.*

PROOF. Let A be nonempty and finite. By the definition of $A^*_k(i)$ and Corollary 2.3, $A^*_k(i) \neq \emptyset$ for $\forall i \in S$, $\forall k \geq 1$. Because A is finite and $A^*_k(i) \subset A^*_{k-1}(i)$, $i \in S$, $k \geq 1$, it is easy to see that $\bigcap_{k=1}^{\infty} A^*_k(i) \neq \emptyset$ for all $i \in S$. We take $f(i) \in \bigcap_{k=1}^{\infty} A^*_k(i)$ for all $i \in S$. By Corollary 2.4, $f^\infty \in \Pi(\infty)$.

THEOREM 2.10. *For $k \geq 1$, let $f^\infty \in \Pi(k-1)$. If*

$$M_k(f^\infty, i) = \sup_{a\in A^*_{k-1}(i)} \left\{ R_k(i,a) + \sum_{j\in S} q(j|i,a) M_k(f^\infty, j) \right\} \text{ for all } i \in S,$$

then $f^\infty \in \Pi(k)$.

PROOF. By Theorem 2.5 and Corollary 2.3, $A^*_{k-1}(i) \neq \emptyset$ for all $i \in S$. By Theorem 2.3, $\forall\epsilon > 0$, $\exists g^\infty$ such that $g(i) \in A^*_{k-1}(i)$ for all $i \in S$ and

$$M_k(g^\infty, i) \geq M^*_k(i) - \epsilon, \quad i \in S.$$

By the supposition,

$$R_k(i, g(i)) + \sum_{j\in S} q(j|i, g(i)) M_k(f^\infty, j) \leq M_k(f^\infty, i), \qquad i \in S.$$

Imitating the proof of Theorem 2.2(1), we have

$$M_k(f^\infty, i) \geq M_k(g^\infty, i), \qquad i \in S,$$

that is,

$$M_k(f^\infty, i) \geq M_k^*(i) - \epsilon, \qquad i \in S.$$

Let $\epsilon \to 0$. We have

$$M_k(f^\infty, i) \geq M_k^*(i), \qquad i \in S.$$

By Theorem 2.5, $f^\infty \in \bigcap_{i \in S} \Pi(k-1, i)$. So, by Theorem 2.5, $f^\infty \in \bigcap_{i \in S} \Pi(k, i) = \Pi(k)$.

3. Algorithm

We shall now give an algorithm of policy-improvement type for finding a k-moment optimal stationary policy. In this section we suppose that S and A are finite. By Theorem 2.9, there exists a f^∞ which is a moment-optimal policy. Obviously, f^∞ is also a $k(\geq 1)$-moment optimal policy.

THEOREM 3.1. *Let $k \geq 1$, $f^\infty \in \Pi(k-1)$. The equation*

$$R_k(i, f(i)) + \sum_{j \in S_0} q(j|i, f(i))V(j) = V(i), \qquad i \in S_0, \tag{3.1}$$

possesses a unique solution $V(i) = M_k(f^\infty, i)$, $i \in S_0$.

PROOF. By Theorem 2.1 and 2.5, $\{M_k(f^\infty, i) : i \in S_0\}$ is a solution of (3.1). By Lemma 1.3, the solution of (3.1) is unique.

By solving (3.1), we can find $M_k(f^\infty, i)$, $i \in S$.

THEOREM 3.2 (Policy improvement). *For $k \geq 1$, let $f^\infty \in \Pi(k-1)$. If $g(i) \in A_{k-1}^*(i)$ for all $i \in S$ and*

$$R_k(i, g(i)) + \sum_{j \in S} q(j|i, g(i))M_k(f^\infty, j) \geq M_k(f^\infty, i) \text{ for all } i \in S,$$

then $M_k(g^\infty) \geq M_k(f^\infty)$.

PROOF. The proof is similar to that of Theorem 2.2(1). Note that, by Theorem 2.5 and Corollary 2.3, $A_{k-1}^*(i) \neq \emptyset$ for all $i \in S$.

Let $k \geq 1$. By Theorem 2.9, $\exists f^\infty \in \Pi(k-1)$. We take $f_0^\infty \in \Pi(k-1)$. By Theorem 2.5 and Corollary 2.3, $A_{k-1}^*(i) \neq \emptyset$ for all $i \in S$. $f_n^\infty(n = 1, 2, \dots)$ is defined as follows: $\forall i \in S$, we take $f_n(i) \in A_{k-1}^*(i)$ such that

$$\max_{a \in A_{k-1}^*(i)} \left\{ R_k(i, a) + \sum_{j \in S} q(j|i, a) M_k(f_{n-1}^\infty, j) \right\} = R_k(i, f_n(i)) + \sum_{j \in S} q(j|i, f_n(i)) M_k(f_{n-1}^\infty, j). \quad (3.2)$$

THEOREM 3.3. *Let $k \geq 1$. For f_n^∞ $(n = 0, 1, 2, \dots)$ defined above, we have*

(1) $M_k(f_n^\infty) \geq M_k(f_{n-1}^\infty)$, $n = 1, 2, \dots$.
(2) $\exists n_0 \geq 0$ *such that* $M_k(f_{n_0}^\infty) = M_k(f_{n_0+1}^\infty)$.
(3) *If* $M_k(f_{n_0}^\infty) = M_k(f_{n_0+1}^\infty)$, *then* $f_{n_0}^\infty \in \Pi(k)$.

PROOF. (1) By Theorem 2.2(2) and Theorem 2.5, $f_n^\infty \in \Pi(k-1)$, $n \geq 0$. By Theorem 2.6, $f_n(i) \in A_{k-1}^*(i)$, $i \in S$, $n \geq 0$. By Theorem 3.1 and 3.2, (1) is true.
(2) Because S and A are finite, Π_s^d is finite. Condition (2) is true from (1).
(3) From Theorem 3.1 and Theorem 2.10, (3) is true.

Let $k \geq 1$. An iteration algorithm for finding a k-moment optimal stationary policy is stated as follows:

(1) $l \Leftarrow 1$. Choose any $f_0^\infty \in \Pi_s^d$.
(2) By (3.2), with the policy improvement iteration starting from f_0^∞(replace k by l in (3.2)), we can find $g^\infty \in \Pi(l)$ (see Theorem 3.3). By Theorem 2.5, $M_l(g^\infty, i) = M_l^*(i)$, $i \in S$.
(3) If $l = k$, then stop. We have $g^\infty \in \Pi(k)$. If $l < k$, then go to (4).
(4) By the definition of $A_l^*(i)$, we find $A_l^*(i)$, $i \in S$. Obviously $A_l^*(i) \neq \emptyset$, $i \in S$.
(5) $l \Leftarrow l + 1$. Let $f_0 = g$. Go to (2).

By the above algorithm, we can find $A_k^*(i)$, $i \in S$, $k \geq 1$. We take $f(i) \in \bigcap_{k=1}^{\infty} A_k^*(i)$ for all $i \in S$, then $f^\infty \in \Pi(\infty)$ (see the proof of Theorem 2.9).

Acknowledgement

This research is supported by the National Natural Science Foundation of China and the Youth Fund of Yunnan Province of China. The authors are very grateful to the referee for his suggestions of various improvements to the paper and to Prof. Gui Xiang-yun for her help to the authors.

References

[1] M. Baykal-Gürsoy and K. W. Ross, "Variability sensitive Markov decision processes", *Math. Oper. Res.* **17** (1992) 558–571.

[2] Dong ze qing, "An accelerated successive approximation method of discounted Markovian decision programming and the least variance problem in optimal policies (Chinese)", *Acta Math. Sinica* **21** (1978) 135–150.

[3] J. A. Filar, L. C. M. Kallenberg and Lee Huey-Miin, "Variance-penalized Markov decision processes", *Math. Oper. Res.* **14** (1989) 147–161.

[4] J. A. Filar and Lee Huey-Miin, "Gain/variability tradeoffs in undiscounted Markov decision processes", in *Proc. 1985 IEEE Conf. (24th Conf.), Decision and Control*, 1106–1112.

[5] S. C. Jaquette, "Markov decision processes with a new optimality criterion: Small interest rates", *Ann. Statist.* **43** (1972) 1894–1901.

[6] S. C. Jaquette, "Markov decision processes with a new optimality criterion: Discrete time", *Ann. Statist.* **1** (1973) 496–505.

[7] H. A. Kawai, "A variance minimization problem for a Markov decision process", *EurJour. Oper. Res.* **31** (1987) 140–145.

[8] Kun-Jen Chung, "A note on maximal mean/standard deviation ratio in an undiscounted Markov decision process", *Oper. Res. Letters* **8** (1989) 201–203.

[9] Kun-Jen Chung, "Mean-variance tradeoffs in an undiscounted Markov decision process: The unichain case", *Oper. Res.* **42** (1994) 184–188.

[10] Lin Jian-xing, "The moment optimal model in which the discount factor is dependent on history(chinese)", M. Sc. Thesis, Department of Appl. Math., Qinghua University.

[11] Liu Jian-yong and Liu Ke, "Markov decision programming- the first-passage model with denumerable state space", *Syst. Sci. and Math. Sci.* **5** (1992) 340–351.

[12] G. Quelle, "Dynamic programming of expectation and variance", *J. Math. Anal. Appl.* **55** (1976) 239–252.

[13] M. L. Sobel, "The variance of discounted Markov decision process", *J. Appl. Prob.* **19** (1982) 794–802.

[14] M. L. Sobel, "Maximal mean/standard deviation ratio in an undiscounted Markov decision process", *Oper. Res. Letters* **4** (1985) 157–158.

[15] M. L. Sobel, "Mean-variance tradeoffs in an undiscounted Markov decision process", *Oper. Res.* **42** (1994) 175–183.

[16] D. J. White, "Variance and probabilistic criteria in finite markov decision processes: A review", *J. Opti. Theory Appl.* **56** (1988) 1–29.