MULTISTAGE MARKOV DECISION PROCESSES WITH MINIMUM CRITERIA OF RANDOM REWARDS

Otsubo, Yoshio
Department of Mathematics and Information Science, Faculty of Science, Kochi University

https://doi.org/10.5109/12598

出版情報：Bulletin of informatics and cybernetics. 38, pp.15-25, 2006-12. Research Association of Statistical Sciences
バージョン：
権利関係：

MULTISTAGE MARKOV DECISION PROCESSES WITH MINIMUM CRITERIA OF RANDOM REWARDS

by

Yoshio Ohtsubo

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.38

FUKUOKA, JAPAN
2006

Bulletin of Informatics and Cybernetics, Vol. 38, 2006

MULTISTAGE MARKOV DECISION PROCESSES WITH MINIMUM CRITERIA OF RANDOM REWARDS

By

Yoshio Ohtsubo*

Abstract

We consider multistage decision processes where criterion function is an expectation of minimum function. We formulate them as Markov decision processes with imbedded parameters. The policy depends upon a history including past imbedded parameters, and the rewards at each stage are random and depend upon current state, action and a next state. We then give an optimality equation by using operators and show that there exists a right continuous deterministic Markov policy, which depends upon a current state and an imbedded parameter.

Key Words and Phrases: Existence of optimal policy, Invariant imbedding method, Markov decision process, Minimum criteria, Optimality equation.

1. Introduction

We consider a generalization of stochastic decision-making in a fuzzy environment. Bellman and Zadeh (1970) first introduce multistage decision processes in a fuzzy environment and propose a recursive formula for deterministic processes and stochastic processes. However Iwamoto and Fujita (1995) and Iwamoto et al. (1999) point out that Bellman and Zadeh's stochastic recursive formula is a posteriori conditional decision process, whose criterion is

$$r_1(a_1) \wedge E^{a_1}[r_2(a_2) \wedge E^{a_2}[r_3(a_3) \wedge \cdots \wedge E^{a_{N-2}}[r_{N-1}(a_{N-1}) \wedge E^{a_{N-1}}[r_N(x_N)]] \cdots]],$$

where $a_1, a_2, \ldots, a_{N-1}$ are actions (controls), $r_1, r_2, \ldots, r_{N-1}$ are reward (membership) functions on an action space, r_N is a reward (goal) function on a state space, x_N is a state, E^{a_i} is a conditional expectation operator given an action a_i and $\wedge$ denotes the minimum operator. Iwamoto and Fujita (1995) investigate a regular decision process with a criterion

$$E^{\pi}_{x_1}[r_1(a_1) \wedge r_2(a_2) \wedge \cdots \wedge r_{N-1}(a_{N-1}) \wedge r_N(x_N)],$$

where x_1 is an initial state and π is a policy, and Iwamoto et al. (1999) consider a generalized decision process with a criterion

$$E^{\pi}_{x_1}[r_1(x_1, a_1) \circ r_2(x_2, a_2) \circ \cdots \circ r_{N-1}(x_{N-1}, a_{N-1}) \circ r_N(x_N)],$$

* Department of Mathematics and Information Science, Faculty of Science, Kochi University, 2-5-1 Akebono-cho, Kochi 780-8520, Japan. E-mail: ohtsubo@math.kochi-u.ac.jp

$r_1, r_2, \ldots, r_{N-1}$ are membership functions of a state and a action, and $\circ$ is an associative binary relation which includes the minimum relation $\wedge$ as a special case. Both give an optimality equation as a recursive formula and find an optimal Markov policy by using an invariant imbedding method. When $\circ = \wedge$, Iwamoto et al. (2001) show that an optimal Markov policy for an invariant imbedding approach yields an optimal general policy which depends upon the history consisting of only states and actions, and give an example in which there is no optimal Markov policy in a space of policies for general decision problem, in which an invariant imbedding approach is not used.

In such a fuzzy environment Kacprzyk (1978) and Esogbue and Bellman (1984) investigate multistage decision processes with fuzzy termination time. Also White (1993a), Wu and Lin (1999) and Ohtsubo and Toyonaga (2002) consider a minimizing problem of threshold probability in discounted Markov decision processes by using a kind of invariant imbedding method, and Ohtsubo (2003) applies them to stochastic shortest path problem.

In this paper we consider multistage decision processes with a countable state space and a minimum criterion

$$E_{x_1}^{\pi}[r_1(x_1, a_1, x_2) \wedge r_2(x_2, a_2, x_3) \wedge \cdots \wedge r_{N-1}(x_{N-1}, a_{N-1}, x_N) \wedge r_N(x_N, a_N)],$$

where $r_1, r_2, \ldots, r_{N-1}$ are random variables depending upon not only a current state and a current action but also a next state. By the way, $r_i(a_i)$ in Bellman and Zadeh (1970) and Iwamoto and Fujita (1995) and $r_i(x_i, a_i)$ in Iwamoto et al. (1999) are both deterministic but not random. However $r_i(x_i, a_i, x_{i+1})$ in our problem is random. Also, in all references as above, policies depend upon only current state and action, but our policies depend upon a past history. We then show that an optimal value function for imbedded processes satisfies an optimality equation (recursive formula) by using operators and give an optimal (imbedded) Markov policy which is right continuous with respect to imbedded parameter.

2. Notations and formulation

In this section we formulate our optimization problems as multistage Markov decision Processes $\Gamma = ((X_n), (A_n), (Y_n), p_n)$ with a finite discrete time space $\mathbf{N} = \{1, 2, \ldots N\}$. The state space S is countable, and denote the state at time $n \in \mathbf{N}$ by X_n. The action space $A = \cup_{s \in S} A(s)$ is countable, where $A(s)$ is a nonempty set of admissible finite actions when the system is in state $s \in S$, and denote the action at time $n \in \mathbf{N}$ by A_n. The reward space E is a countable set $\{y_1, y_2, \ldots\}$ where each reward y_i $(i = 1, 2, \ldots)$ is nonnegative and E is bounded, that is, $0 \leq y_i \leq M$ for some $M > 0$ and every $y_i \in E$. If $M \leq 1$, we are in a fuzzy environment. $Y_n \in E$ is a random reward function at time $n \in \mathbf{N}$ and we define time-nonstationary probability distributions by

$$\begin{aligned}
&q_n^a(s'|s) = P(X_{n+1} = s'|X_n = s, A_n = a), \quad 1 \leq n \leq N-1, \\
&\hat{q}_n^a(y|s, s') = P(Y_n = y|X_n = s, X_{n+1} = s', A_n = a), \quad 1 \leq n \leq N-1, \\
&\tilde{p}_N^a(y|s) = P(Y_N = y|X_N = s, A_N = a)
\end{aligned}$$

and set

$$p_n^a(s', y|s) = q_n^a(s'|s)\hat{q}_n^a(y|s, s') = P(X_{n+1} = s', Y_n = y|X_n = s, A_n = a)$$

for $s, s' \in S, a \in A(s)$ and $y \in E$, where $\sum_{s' \in S} q_n^a(s'|s) = \sum_{y \in E} \hat{q}_n^a(y|s, s') = \sum_{y \in E} \tilde{p}_N^a(y|s) = 1$. For an additive case, it is known in Chapter 7 of Howard (1960) and Chapter 1 of White (1993b) that the right-side of an optimality equation is

$$\sum_{s' \in S} \sum_{y \in E} p_n^a(s', y|s)(y + v(s')) = \sum_{s' \in S} \sum_{y \in E} p_n^a(s', y|s)y + \sum_{s' \in S} q_n^a(s'|s)v(s')$$

for some function v, which implies that one-step expected reward $\hat{r}_s^a = \sum_{s' \in S} \sum_{y \in E} p_n^a(s', y|s)y$ depend upon a state s and an action a. Hence in an additive case it is enough to give one-step reward $\hat{r}_s^a$, which is not random. However in our minimum case we generally have

$$\sum_{s' \in S} \sum_{y \in E} p_n^a(s', y|s)(y \wedge v(s')) \neq \sum_{s' \in S} \sum_{y \in E} p_n^a(s', y|s)y \wedge \sum_{s' \in S} q_n^a(s'|s)v(s').$$

Thus our decision problem is a generalization of Bellman and Zadeh (1970), Iwamoto and Fujita (1995) and Iwamoto et al. (2001).

We define the random reward as a criterion function by

$$Z = \min_{1 \leq n \leq N} Y_n.$$

Then our problem is to maximize the expected reward $E_s^\pi[Z]$ with respect to all policies π. In order to analysis our problem we define the random reward for a subproblem by

$$Z_n = \min_{n \leq k \leq N} Y_k, \quad 1 \leq n \leq N.$$

Further we define another random sequence as an imbedded parameter by

$$\Lambda_1 = \lambda, \quad \Lambda_{n+1} = \min(Y_n, \Lambda_n), \quad 1 \leq n \leq N,$$

where λ is a given initial value in $[0, M]$.

We use $S_1 = S \times [0, M]$ as a new state space. Let $H_1 = S_1$ and $H_{n+1} = H_n \times A \times S_1$ for each $1 \leq n \leq N-1$. Then H_n represents the set of all possible histories $h_n = (s_1, \lambda_1, a_1, s_2, \lambda_2, \ldots, a_{n-1}, s_n, \lambda_n)$ of the system when the nth action must be chosen, and we denote by θ_n the history at time $n \in \mathbf{N}$. A decision rule δ_n for time $n \in \mathbf{N}$ is a conditional probability given $\theta_n = (X_1, \Lambda_1, A_1, X_2, \Lambda_2, \ldots, A_{n-1}, X_n, \Lambda_n)$: $\delta_n(a_n|h_n) = P(A_n = a_n|\theta_n = h_n)$. It is assumed that $\delta_n(A_n \in A(s_n)|h_n) = 1$ for every history $h_n = (s_1, \lambda_1, a_1, s_2, \lambda_2, \ldots, s_n, \lambda_n) \in H_n$ and $\delta_n(a_n|\cdot)$ is Lebesgue-Stieltjes measurable function on H_n. We denote by $\Delta(n)$ the set of such decision rules δ_n. A policy π is a finite sequence of decision rules $(\delta_n, n \in \mathbf{N}) = (\delta_1, \delta_2, \ldots, \delta_N)$, where $\delta_n \in \Delta(n)$. We denote by C the set of all such policies. We also denote by C_n^N the set of all sequences $(\delta_n, \delta_{n+1}, \ldots, \delta_{N-1})$ of decision rules $\delta_k \in \Delta(k-n+1)$ $(k = n, n+1, \ldots, N)$. C_n^N are used in subproblems. We then note that $C_1^N = C$.

A policy $\pi = (\delta_n, n \in \mathbf{N})$ is said to be Markov (for an imbedded decision process) when the decision rule δ_n is a function of $(X_n, \Lambda_n) = (s_n, \lambda_n)$ for every $n \in \mathbf{N}$. We denote the set of such decision rules by Δ_M and the set of all Markov policies by C_M. Also, a policy $\pi = (\delta_n, n \in \mathbf{N})$ is called a deterministic Markov policy if π is Markov and $\delta_n(a|s, \lambda) = 1$ for every $(s, \lambda) \in S_1$ and some $a \in A(s)$. We write $\delta_n(s, \lambda) = a$ for such a decision rule δ_n and we denote by Δ_D the set of such decision rules. We also

denote the set of all deterministic Markov policies by C_D. $\pi \in C_n^N \cap C_D$ means that $\pi = (\delta_n, \delta_{n+1}, \dots, \delta_N) \in C_n^N$ and $\delta_k \in \Delta_D$ $(k = n, n+1, \dots, N)$.

A decision rule $\delta \in \Delta_D$ is said to be right continuous (on S_1) if for each $(s, \lambda) \in S_1$ there is a positive real number μ such that $\delta(s, \lambda) = \delta(s, \lambda + u)$ for all $u : 0 \le u < \mu$. A policy $\pi = (\delta_n, n \in \mathbf{N}) \in C_D$ is said to be right continuous if the decision rule δ_n is right continuous for every $n \in \mathbf{N}$.

We consider an imbedded decision subproblem in which we maximize an expectation

$$F_n^\pi(s, \lambda) = E_s^\pi[\lambda \wedge Z_n], \quad n \in \mathbf{N}$$

with respect to all policies π for any imbedded parameter $\lambda \in [0, M]$ where $x \wedge y = \min(x, y)$. When $N = 3$, the explicit form of the expectation $F_1^\pi(s_1, \lambda)$ is

$$\begin{aligned} E_{s_1}^\pi[\lambda \wedge Z_1] \quad = \quad & \sum_{a_1 \in A(s_1)} \sum_{y_1 \in E} \sum_{s_2 \in S} \sum_{a_2 \in A(s_2)} \sum_{y_2 \in E} \sum_{s_3 \in S} \sum_{a_3 \in A(s_3)} \sum_{y_3 \in E} (\lambda \wedge y_1 \wedge y_2 \wedge y_3) \\ & \times \tilde{p}_3^{a_3}(y_3|s_3)\delta_3(a_3|s_1, \lambda, a_1, s_2, \lambda \wedge y_1, a_2, s_3, \lambda \wedge y_1 \wedge y_2) \\ & \times p_2^{a_2}(s_3, y_2|s_2)\delta_2(a_2|s_1, \lambda, a_1, s_2, \lambda \wedge y_1) \\ & \times p_1^{a_1}(s_2, y_1|s_1)\delta_1(a_1|s_1, \lambda) \end{aligned}$$

for $(s_1, \lambda) \in S_1$ and $\pi = (\delta_1, \delta_2, \delta_3) \in C_1^3$. We define optimal value functions F_n^* for the imbedded subproblem by

$$F_n^*(s, \lambda) = \sup_{\pi \in C_n^N} F_n^\pi(s, \lambda), \quad n \in \mathbf{N}$$

for each $(s, \lambda) \in S_1$. Then we notice that optimal value in the original problem is

$$F_1^*(s, M) = \sup_{\pi \in C_1^N} F_1^\pi(s, M) = \sup_{\pi \in C} E_s^\pi[Z],$$

since $Z \le M$. A policy π is said to be optimal if $F_1^*(s, M) = F_1^\pi(s, M)$ for every $s \in S$.

We define the following sets of functions: let $\mathcal{F}$ be the set of functions F from S_1 into an interval $[0, M]$ such that $F(s, 0) = 0$ for each $s \in S$ and $F(s, \cdot)$ is measurable on $[0, M]$, and let $\mathcal{F}_c$ be the set of functions $F \in \mathcal{F}$ such that $F(s, \cdot)$ is nondecreasing and continuous on $[0, M]$ for each $s \in S$. In Theorem 3.1 it is shown that $F_n^* \in \mathcal{F}_c$.

We define operators T_n^a, T_n^δ and T_n from $\mathcal{F}$ into itself as follows. For $F \in \mathcal{F}$, $(s, \lambda) \in S_1$, $a \in A(s)$, $\delta \in \Delta_M$ and $1 \le n \le N - 1$,

$$\begin{aligned} T_n^a F(s, \lambda) &= \sum_{s' \in S} \sum_{y \in E} F(s', \lambda \wedge y) p_n^a(s', y|s), \\ T_n^\delta F(s, \lambda) &= \sum_{a \in A(s)} T_n^a F(s, \lambda)\delta(a|s, \lambda), \\ T_n F(s, \lambda) &= \sup_{\delta \in \Delta_M} T_n^\delta F(s, \lambda) = \max_{a \in A(s)} T_n^a F(s, \lambda) \end{aligned}$$

and

$$T_N^a F(s, \lambda) = \sum_{y \in E} F(s, \lambda \wedge y)\tilde{p}_N^a(y|s),$$

$$T_N^\delta F(s,\lambda) = \sum_{a\in A(s)} T_N^a F(s,\lambda)\delta(a|s,\lambda),$$
$$T_N F(s,\lambda) = \sup_{\delta\in\Delta_M} T_N^\delta F(s,\lambda) = \max_{a\in A(s)} T_N^a F(s,\lambda).$$

In all argument, for $F, G \in \mathcal{F}$, $F \geq G$ means that $F(s,\lambda) \geq G(s,\lambda)$ for all $(s,\lambda) \in S_1$.

3. Optimal value and optimal policy

In this section we give an optimality equation by using operators T_n and show that there exist a right continuous deterministic Markov policy.

We first give fundamental lemmas for operators T_n^a, T_n^δ and T_n.

LEMMA 3.1. (i) *For $F, G \in \mathcal{F}$, $\delta \in \Delta$ and $n \in \mathbf{N}$, $T_n^\delta F - T_n^\delta G = T_n^\delta(F-G)$.*

(ii) *Let $n \in \mathbf{N}$. If $F, G \in \mathcal{F}$ and $F \geq G$, then $T_n^a F \geq T_n^a G$ for each $a \in A(\cdot)$, $T_n^\delta F \geq T_n^\delta G$ for each $\delta \in \Delta$ and $T_n F \geq T_n G$.*

(iii) *Let $n \in \mathbf{N}$. If $F \in \mathcal{F}_c$, then $T_n^a F \in \mathcal{F}_c$ for any $a \in A(\cdot)$ and $T_n F \in \mathcal{F}_c$.*

PROOF. The statements (i) and (ii) are immediate results of definitions.

(iii) Let $F \in \mathcal{F}_c$ and let $s \in S$ be arbitrary. Then $F(s',0) = 0$ for every $s' \in S$. Since $y \in [0,M]$, we have $T_n^a F(s,0) = 0$ for every $a \in A(s)$ and hence $T_n F(s,0) = 0$. It easily follows that $T_n^a F(s,\cdot)$ and $T_n F(s,\cdot)$ are nondecreasing on $[0,M]$, since $F(s',\cdot)$ is nondecreasing. Also, by the dominated convergence theorem we see that $T_n^a F(s,\cdot)$ is continuous on $[0,M]$ for each $a \in A(s)$, since $F(s',\cdot)$ is continuous. Thus since $A(s)$ is finite, $T_n F(s,\cdot)$ is also continuous on $[0,M]$. Therefore the proof is complete.

LEMMA 3.2. *Let $n \in \mathbf{N}$. For each $F \in \mathcal{F}_c$, there exists a right continuous decision rule $\delta \in \Delta_D$ satisfying $T_n F = T_n^\delta F$.*

PROOF. Let $F \in \mathcal{F}_c$ and $(s,\lambda) \in S_1$ be arbitrarily fixed. From Lemma 3.1 (iii), $T_n^a F(s,\cdot)$ is continuous on $[0,M]$ for each $a \in A(s)$. Since $A(s)$ is finite, we see that there exist $\mu > 0$ and $a \in A(s)$ such that $TF(s,u) = T^a F(s,u)$ for all u satisfying $\lambda \leq u < \lambda + \mu$. For such an action a, if we define $\delta \in \Delta_D$ by $\delta(s,u) = a$ for every u so that $\lambda \leq u < \lambda + \mu$, then δ is right continuous and $T_n F(s,\lambda) = T_n^\delta F(s,\lambda)$. Therefore the proof is complete.

For any $\pi = (\delta_n, \delta_{n+1}, \ldots, \delta_N) \in C_n^N$ and a given history $(s,\lambda,a) \in S_1 \times A$, the cut-head policy of π to (s,λ,a) is defined by ${}^1\pi^{(s,\lambda,a)} = (\delta_{n+1}^{(s,\lambda,a)}, \delta_{n+2}^{(s,\lambda,a)}, \ldots, \delta_N^{(s,\lambda,a)})$ where $\delta_{k+1}^{(s,\lambda,a)}(\cdot|h_k) = \delta_{k+1}(\cdot|(s,\lambda,a),h_k)$ for every $h_k \in H_{k-n+1}$ and each $k = n, n+1, \ldots, N-1$. Then we see that ${}^1\pi^{(s,\lambda,a)} \in C_{n+1}^N$ for a fixed (s,λ,a). For the sake of simplicity we use a notation:

$$T_n^{\delta_n} F_{n+1}^{{}^1\pi}(s,\lambda) = \sum_{a\in A(s)} \delta_n(a|s,\lambda) \sum_{s',y} F_{n+1}^{{}^1\pi^{(s,\lambda,a)}}(s', \lambda\wedge y) p_n^a(s',y|s)$$

for $\pi = (\delta_n, \delta_{n+1}, \ldots, \delta_N) \in C_n^N$ and $(s,\lambda) \in S_1$.

LEMMA 3.3. (i) *For each $n \in \mathbf{N}$ and any $\pi \in C_n^N$, $F_n^\pi \in \mathcal{F}$.*

(ii) *Let $n = 1, 2, \ldots, N-1$ and let $\pi = (\delta_n, \delta_{n+1}, \ldots, \delta_N) \in C_n^N$ be arbitrary. Then $F_n^\pi = T_n^{\delta_n} F_{n+1}^{{}^1\pi}$.*

PROOF. To show that $F_n^\pi \in \mathcal{F}$, it suffices to prove that $F_n^\pi(s,\cdot)$ is measurable on $[0,1]$ for each $s \in S$, since $F_n^\pi(s,0) = E_s^\pi[0 \wedge Z_n] = 0$ for every $s \in S$. For any $\pi = (\delta_N) \in C_N^N$ we have

$$F_N^\pi(s,\lambda) = E_s^\pi[\lambda \wedge Z_N] = \sum_{a \in A(s)} \delta_N(a|s,\lambda) \sum_y (\lambda \wedge y) \tilde{p}_N^a(y|s).$$

Hence $F_N^\pi(s,\cdot)$ is measurable. We assume that $F_{n+1}^{{}^1\pi}(s,\cdot)$ is measurable for $n \leq N-1$ and any ${}^1\pi \in C_{n+1}^N$. Then for any $\pi = (\delta_n, \cdots, \delta_N) \in C_n^N$ we have

$$T_n^{\delta_n} F_{n+1}^{{}^1\pi}(s,\lambda) = \sum_{a \in A(s)} \delta_n(a|s,\lambda) \sum_{s',y} F_{n+1}^{{}^1\pi(s,\lambda,a)}(s', \lambda \wedge y) p_n^a(s', y|s).$$

Thus $T_n^{\delta_n} F_{n+1}^{{}^1\pi}(s,\dot{)}$ is measurable. However it follows from Markov property that

$$\begin{aligned} T_n^{\delta_n} F_{n+1}^{{}^1\pi}(s,\lambda) &= \sum_{a \in A(s)} \delta_n(a|s,\lambda) \sum_{s',y} E_{s'}^{{}^1\pi(s,\lambda,a)}[\lambda \wedge y \wedge Z_{n+1}] p_n^a(s', y|s) \\ &= E_s^\pi[\lambda \wedge Z_n] \\ &= F_n^\pi(s,\lambda). \end{aligned}$$

Hence $F_n^\pi(s,\dot{)}$ is measurable and we also have $T_n^{\delta_n} F_{n+1}^{{}^1\pi}(s,\lambda) = F_n^\pi(s,\lambda)$. Therefore the proof is complete.

We next give a main theorem for optimal value functions and optimal policies in subproblems.

THEOREM 3.1. (i) *For each $n \in \mathbf{N}$, $F_n^* \in \mathcal{F}_c$ and $\{F_n^*, n \in \mathbf{N}\}$ satisfies optimality equations:*

$$F_n^* = T_n F_{n+1}^*, \ 1 \leq n \leq N-1,$$

with $F_N^*(s,\lambda) = \max_{a \in A(s)} \sum_y (\lambda \wedge y) \tilde{p}_N^a(y|s)$.

(ii) *For each $n \in \mathbf{N}$, there exists a right continuous policy $\pi \in C_n^N \cap C_D$ such that $F_n^* = F_n^\pi$.*

PROOF. We prove the statements (i) and (ii) by backward induction. When $n = N$, we see that

$$\begin{aligned} F_N^*(s,\lambda) &= \sup_{\pi \in C_N^N} E_s^\pi[\lambda \wedge Z_N] \\ &= \sup_{\delta_N} \sum_{a \in A(s)} \delta_N(a|s,\lambda) \sum_y (\lambda \wedge y) \tilde{p}_N^a(y|s) \\ &= \max_{a \in A(s)} \sum_y (\lambda \wedge y) \tilde{p}_N^a(y|s). \end{aligned}$$

Since $\sum_y (\lambda \wedge y) \tilde{p}_N^a(y|s)$ is right continuous in λ, there is a right continuous decision rule $\delta_N \in \Delta_D$ such that

$$F_N^*(s,\lambda) = \sum_y (\lambda \wedge y) \tilde{p}_N^{\delta_N(s,\lambda)}(y|s) = F_N^{\pi^*}(s,\lambda)$$

where $\pi^* = (\delta_N) \in C_N^N \cap C_D$, and hence $F_N^* \in \mathcal{F}_c$. We assume that $F_k^* \in \mathcal{F}_c$ and there exist a right continuous policy $\sigma \in C_k^N \cap C_D$ such that $F_k^* = F_k^\sigma$. Then it follows from Lemma 3.2 that there is a right continuous decision rule $\hat{\delta}_{k-1} \in \Delta_D$ such that $T_{k-1}F_k^* = T_{k-1}^{\hat{\delta}_{k-1}}F_k^*$, which implies that $\pi = (\hat{\delta}_{k-1}, \sigma) \in C_{k-1}^N \cap C_D$ is a right continuous policy. It also follows from Lemma 3.3(ii) that

$$F_{k-1}^*(s,\lambda) \geq F_{k-1}^\pi(s,\lambda) = T_{k-1}^{\hat{\delta}_{k-1}}F_k^\sigma(s,\lambda) = T_{k-1}^{\hat{\delta}_{k-1}}F_k^*(s,\lambda) = T_{k-1}F_k^*(s,\lambda)$$

for each $(s,\lambda) \in S_1$. Conversely, we see from Lemma 3.3(ii) again that for any policy $\tau = (\delta_{k-1}, \delta_k, \ldots, \delta_N) \in C_{k-1}^N$,

$$F_{k+1}^\tau(s,\lambda) = T_{k-1}^{\delta_{k-1}}F_k^{{}^1\tau}(s,\lambda) \leq T_{k-1}^{\delta_{k-1}}F_k^*(s,\lambda) \leq T_{k-1}F_k^*(s,\lambda),$$

since ${}^1\tau \in C_k^N$. Taking supremum over C_{k-1}^N, we obtain $F_{k-1}^*(s,\lambda) \leq T_{k-1}F_k^*(s,\lambda)$. Thus, combining with the previous inequality, we have $T_{k-1}F_k^* = F_{k-1}^* = F_{k-1}^\pi$. Hence, π satisfies $F_{k-1}^* = F_{k-1}^\pi$, and from Lemma 3.1(iii), we have $F_{k-1}^* \in \mathcal{F}_c$. By backward induction, the proof of the statements (i) and (ii) is complete.

We finally find optimal function and optimal policy in the original problem.

THEOREM 3.2. Optimal value $F_1^*(s,M)$ is given by

$$F_1^* = T_1T_2\cdots T_{N-1}F_N^*$$

with $F_N^*(s,r) = \max_{a\in A(s)}\sum_y(\lambda \wedge y)\tilde{p}_N^a(y|s)$, and a right continuous optimal policy $\pi = (\delta_1, \delta_2, \ldots, \delta_n) \in C_D$ is obtained by

$$T_nF_{n+1}^* = T_n^{\delta_n}F_{n+1}^*,\ 1 \leq n \leq N-1,$$
$$F_N^*(s,\lambda) = \sum_y(\lambda \wedge y)\tilde{p}_N^{\delta_N(s,\lambda)}(y|s) = \max_{a\in A(s)}\sum_y(\lambda \wedge y)\tilde{p}_N^a(y|s),\ (s,\lambda) \in S_1.$$

PROOF. These are immediate results of Theorem 3.1 and its proof.

4. A numerical example

We give a simple numerical example of two stages, two states and two actions in this section.

Let state space be $S = \{s_1, s_2\}$ and let action space be $A(s_1) = A(s_2) = \{a_1, a_2\}$ and let $N = 2$. We assume that probability distributions $p_1^a(s', y|s)$ and $\tilde{p}_2^a(y|s)$ are determined by

$$p_1^{a_1}(s_1, y|s_1) = \begin{cases} \frac{1}{4} & \text{if } y = 1 \\ \frac{1}{4} & \text{if } y = 3 \end{cases}, \qquad p_1^{a_1}(s_2, 2|s_1) = \tfrac{1}{2},$$

$$p_1^{a_2}(s_1, \tfrac{8}{5}|s_1) = \tfrac{2}{5}, \qquad p_1^{a_2}(s_2, y|s_1) = \begin{cases} \frac{1}{5} & \text{if } y = \frac{6}{5} \\ \frac{2}{5} & \text{if } y = \frac{3}{2} \end{cases},$$

$$p_1^{a_1}(s_1, \tfrac{5}{4}|s_2) = 1, \qquad p_1^{a_2}(s_2, 3|s_2) = 1,$$

$$\tilde{p}_2^{a_1}(y|s_1) = \begin{cases} \frac{2}{5} & \text{if } y = 3 \\ \frac{3}{5} & \text{if } y = 2 \end{cases}, \qquad \tilde{p}_2^{a_1}(y|s_2) = \begin{cases} \frac{1}{2} & \text{if } y = 1 \\ \frac{1}{2} & \text{if } y = 2 \end{cases},$$

$$\tilde{p}_2^{a_2}(y|s_1) = \begin{cases} \frac{3}{10} & \text{if } y = \frac{3}{2} \\ \frac{7}{10} & \text{if } y = 3 \end{cases}, \qquad \tilde{p}_2^{a_2}(y|s_2) = \begin{cases} \frac{4}{5} & \text{if } y = 1 \\ \frac{1}{5} & \text{if } y = 4 \end{cases}.$$

Then we shall first give optimal value F_2^* and optimal decision δ_2 at $N=2$. Since $F_2^*(s,\lambda)=\max_{a\in A(s)}\sum_y(\lambda\wedge y)\tilde{p}_2^a(y|s)$ by Theorem 3.1, we have

$$\begin{aligned} F_2^*(s_1,\lambda) &= \max(\frac{2}{5}(\lambda\wedge 3)+\frac{3}{5}(\lambda\wedge 2), \frac{3}{10}(\lambda\wedge\frac{3}{2})+\frac{7}{10}(\lambda\wedge 3)) \\ &= \begin{cases} \lambda & \text{if } \lambda<2 \\ \frac{2}{5}\lambda+\frac{6}{5} & \text{if } 2\le\lambda<\frac{5}{2} \\ \frac{7}{10}\lambda+\frac{9}{20} & \text{if } \frac{5}{2}\le\lambda<3 \\ \frac{51}{20} & \text{if } \lambda\ge 3 \end{cases}, \end{aligned}$$

and

$$\begin{aligned} F_2^*(s_2,\lambda) &= \max(\frac{1}{2}(\lambda\wedge 1)+\frac{1}{2}(\lambda\wedge 2), \frac{4}{5}(\lambda\wedge 1)+\frac{1}{5}(\lambda\wedge 4)) \\ &= \begin{cases} \lambda & \text{if } \lambda<1 \\ \frac{1}{2}\lambda+\frac{1}{2} & \text{if } 1\le\lambda<2 \\ \frac{3}{2} & \text{if } 2\le\lambda<\frac{7}{2} \\ \frac{1}{5}\lambda+\frac{4}{5} & \text{if } \frac{7}{2}\le\lambda<4 \\ \frac{8}{5} & \text{if } \lambda\ge 4 \end{cases}. \end{aligned}$$

Hence we have

$$\delta_2(s_1,\lambda)=\begin{cases} a_1 \text{ or } a_2 & \text{if } \lambda<\frac{3}{2} \\ a_1 & \text{if } \frac{3}{2}\le\lambda<\frac{5}{2} \\ a_2 & \text{if}\lambda\ge\frac{5}{2} \end{cases}, \qquad \delta_2(s_2,\lambda)=\begin{cases} a_1 \text{ or } a_2 & \text{if } \lambda<1 \\ a_1 & \text{if } 1\le\lambda<\frac{7}{2} \\ a_2 & \text{if}\lambda\ge\frac{7}{2} \end{cases}.$$

Next we consider optimal F_1^* and δ_1 at the first stage. From Theorem 3.1, we have

$$\begin{aligned} F_1^*(s,\lambda) &= T_1F_2^*(s,\lambda)=\max_{a\in A(s)}T_1^aF_2^*(s,\lambda) \\ &= \max_{a\in A(s)}\sum_{s'\in S}\sum_{y\in E}F_2^*(s',\lambda\wedge y)p_1^a(s',y|s). \end{aligned}$$

By the way, we easily see that

$$\begin{aligned} T_1^{a_1}F_2^*(s_1,\lambda) &= \frac{1}{4}F_2^*(s_1,\lambda\wedge 1)+\frac{1}{4}F_2^*(s_1,\lambda\wedge 3)+\frac{1}{2}F_2^*(s_2,\lambda\wedge 2) \\ &= \begin{cases} \lambda & \text{if } \lambda<1 \\ \frac{1}{2}\lambda+\frac{1}{2} & \text{if } 1\le\lambda<2 \\ \frac{1}{10}\lambda+\frac{13}{10} & \text{if } 2\le\lambda<\frac{5}{2} \\ \frac{7}{40}\lambda+\frac{89}{80} & \text{if } \frac{5}{2}\le\lambda<3 \\ \frac{131}{80} & \text{if } \lambda\ge 3 \end{cases}, \end{aligned}$$

and

$$T_1^{a_2}F_2^*(s_1,\lambda) = \frac{2}{5}F_2^*(s_1,\lambda\wedge\frac{8}{5})+\frac{2}{5}F_2^*(s_2,\lambda\wedge\frac{3}{2})+\frac{1}{5}F_2^*(s_2,\lambda\wedge\frac{6}{5})$$
$$= \begin{cases} \lambda & \text{if } \lambda<1 \\ \frac{7}{10}\lambda+\frac{3}{10} & \text{if } 1\le\lambda<\frac{6}{5} \\ \frac{3}{5}\lambda+\frac{21}{50} & \text{if } \frac{6}{5}\le\lambda<\frac{3}{2} \\ \frac{2}{5}\lambda+\frac{18}{25} & \text{if } \frac{3}{2}\le\lambda<\frac{8}{5} \\ \frac{34}{25} & \text{if } \lambda\ge\frac{8}{5} \end{cases}.$$

Hence we have optimal value in the imbedded problem as follows :

$$F_1^*(s_1,\lambda) = \max_a T_1^a F_2^*(s_1,\lambda)$$
$$= \begin{cases} T_1^{a_1}F_2^*(s_1,\lambda)=T_1^{a_2}F_2^*(s_1,\lambda) & \text{if } \lambda<1 \\ T_1^{a_2}F_2^*(s_1,\lambda) & \text{if } 1\le\lambda<\frac{43}{25} \\ T_1^{a_1}F_2^*(s_1,\lambda) & \text{if } \lambda\ge\frac{43}{25} \end{cases}.$$
$$= \begin{cases} \lambda & \text{if } \lambda<1 \\ \frac{7}{10}\lambda+\frac{3}{10} & \text{if } 1\le\lambda<\frac{6}{5} \\ \frac{3}{5}\lambda+\frac{21}{50} & \text{if } \frac{6}{5}\le\lambda<\frac{3}{2} \\ \frac{2}{5}\lambda+\frac{18}{25} & \text{if } \frac{3}{2}\le\lambda<\frac{8}{5} \\ \frac{34}{25} & \text{if } \frac{8}{5}\le\lambda<\frac{43}{25} \\ \frac{1}{2}\lambda+\frac{1}{2} & \text{if } \frac{43}{25}\le\lambda<2 \\ \frac{1}{10}\lambda+\frac{13}{10} & \text{if } 2\le\lambda<\frac{5}{2} \\ \frac{7}{40}\lambda+\frac{89}{80} & \text{if } \frac{5}{2}\le\lambda<3 \\ \frac{131}{80} & \text{if } \lambda\ge3 \end{cases}.$$

and optimal decision at state s_1 is

$$\delta_1(s_1,\lambda) = \begin{cases} a_1 \text{ or } a_2 & \text{if } \lambda<1 \\ a_2 & \text{if } 1\le\lambda<\frac{43}{25} \\ a_1 & \text{if } \lambda\ge\frac{43}{25} \end{cases}.$$

Similarly, optimal value at state s_2 is

$$F_1^*(s_2,\lambda) = \begin{cases} \lambda & \text{if } \lambda<\frac{5}{4} \\ \frac{5}{4} & \text{if } \frac{5}{4}\le\lambda<\frac{3}{2} \\ \frac{1}{2}\lambda+\frac{1}{2} & \text{if } \frac{3}{2}\le\lambda<2 \\ \frac{3}{2} & \text{if } \lambda\ge2 \end{cases},$$

since we easily have

$$T_1^{a_1}F_2^*(s_2,\lambda)=\begin{cases}\lambda & \text{if } \lambda<\frac{5}{4}\\ \frac{5}{4} & \text{if } \lambda\geq\frac{5}{4}\end{cases}, \qquad T_1^{a_2}F_2^*(s_2,\lambda)=\begin{cases}\lambda & \text{if } \lambda<1\\ \frac{1}{2}\lambda+\frac{1}{2} & \text{if } 1\leq\lambda<2\\ \frac{3}{2} & \text{if } \lambda\geq 2\end{cases},$$

and optimal decision is

$$\delta_1(s_2,\lambda)=\begin{cases}a_1 \text{ or } a_2 & \text{if } \lambda<1\\ a_1 & \text{if } 1\leq\lambda<\frac{3}{2}\\ a_2 & \text{if}\lambda\geq\frac{3}{2}\end{cases}.$$

Therefore, since $M=4$, we have optimal value in the original problem as follows:

$$F_1^*(s_1,M)=\frac{131}{80}, \quad F_1^*(s_2,M)=\frac{3}{2}.$$

Acknowledgement

The author is grateful to the referee for valuable remarks and helpful comments.

References

Bellman, R.E. and Zadeh, L.A (1970). Decision-making in a fuzzy environment, *Management Science.* **17**, B141-B164.

Iwamoto, S. and Fujita, T. (1995). Stochastic decision-making in a fuzzy environment, *J. Operations Research Society of Japan.* **38**, 467-482.

Iwamoto, S., Tsurusaki, K. and Fujita, T. (1999). Conditional decision-making in fuzzy environment, *J. Operations Research Society of Japan.* **42**, 198-218.

Iwamoto, S., Tsurusaki, K. and Fujita, T. (2001). On Markov policies for minimax decision processes, *J. Math. Anal. Appl.* **253**, 58-78.

Kacprzyk, J. (1978). Decision-making in a fuzzy environment with fuzzy termination time, *Fuzzy Sets and Systems.* **1**, 169-179.

Esogbue, A.O. and Bellman, R.E. (1984). Fuzzy dynamic programming and its extensions, *TIMS/Studies in Management Sciences.* **20**, 147-167.

White, D.J. (1993a). Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.* **173**, 634-646.

Wu, C. and Lin, Y. (1999). Minimizing risk models in Markov decision processes with policies depending on target values, *J. Math. Anal. Appl.* **231**, 47-67.

Ohtsubo, Y. and Toyonaga, K. (2002). Optimal policy for minimizing risk models in Markov decision processes, *J. Math. Anal. Appl.* **271**, 66-81.

Ohtsubo, Y. (2003). Minimizing risk models in stochastic shortest path problems, *Math. Meth. Oper. Res.* **57**, 79-88.

Howard, R.A. (1960). *Dynamic Programming and Markov Processes*, Wiley, London.

White, D.J. (1993b). *Markov Decision Processes*, Wiley, New York.

Received August 26, 2004
Revised April 27, 2005