RE-EXAMINATION OF MARKOV POLICIES FOR ADDITIVE DECISION PROCESS

Fujita, Toshiharu
Graduate School of Mathematics, Kyushu University

https://doi.org/10.5109/13461

出版情報：Bulletin of informatics and cybernetics. 29 (1), pp.51-65, 1997-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 29, No. 1, 1997

RE-EXAMINATION OF MARKOV POLICIES FOR ADDITIVE DECISION PROCESS

By

Toshiharu FUJITA*

Abstract

The purpose of this paper is to ensure that Markov policy is enough for the additive decision process. We show that there exists an optimal policy which is Markov for both stochastic additive decision process and deterministic one. We also verify through a multi-stage stochastic decision tree method that, among the class of general policies, there exist an optimal Markov policy. This fact is of course obtained by solving the regular recursive equation.

1. Introduction

In this paper we consider a class of sequential optimization problems whose objective function is additive. We focus our attention on the tacitly known but never clearly proved fact that Markov policy is enough in the so-called Markov decision processes (Bellman and Zadeh(1970, p. 152, *ll.* 19-22), Bertsekas and Shreve(1978, p. 6, *ll.* 20-23), Howard(1960), Stokey and Lucas Jr.(1989), Puterman(1994), Iwamoto, Tsurusaki and Fujita, and others). It has long been thought that there is no room for argument on this point. However, we must accept a recent result that an optimal policy is not Markov for a certain problem especially for minimum objective function problem (Iwamoto and Fujita(1995), Iwamoto, Tsurusaki and Fujita). These two apparently inconsistent facts enable us to reconsider whether Markov policy is enough or not.

In section 2 we discuss the stochastic maximization of additive function. We derive a recursive equation both for the class of general policies and for the class of Markov policies. Verifying that the optimal value functions in both classes are identical, we show that Markov policy is enough. In section 3 we give rigorous proofs of theorems in section 2. In section 4 we show the corresponding results for the deterministic optimization, which is a degenerate case of the stochastic one. In section 5, illustrating multi-stage stochastic decision trees, we give a two-stage stochastic decision process, a pair of optimal value functions, and an optimal Markov policy. The pair is also obtained by solving the corresponding recursive equation.

* Graduate School of Mathematics, Kyushu University 33, Fukuoka 812-81, Japan.
Research Fellow of the Japan Society for the Promotion of the Science.
e-mail: fujita@math.kyushu-u.ac.jp

Throughout the paper the following definition and notation are used :

$N \geq 2$ is an integer ; the *total number of stages*

$X = \{s_1, s_2, \ldots, s_p\}$ is a finite *state space*

$U = \{a_1, a_2, \ldots, a_k\}$ is a finite *action space*

$r_n : X \times U \to R^1$ is an *n-th reward function* $(1 \leq n \leq N)$

$r_G : X \to R^1$ is a *terminal reward function*

$f : X \times U \to X$ is a *deterministic transition law*

; $f(x,u)$ represents the successor state of x for action u

p is a *Markov transition law*

: $p(y|x,u) \geq 0 \ \ \forall (x,u,y) \in X \times U \times X, \quad \sum_{y \in X} p(y|x,u) = 1 \ \ \forall (x,u) \in X \times U$

$y \sim p(\cdot\,|x,u)$ denotes that next state y conditioned on state x and action u appears with probability $p(y|x,u)$.

2. Stochastic Maximization

We consider the stochastic maximization problem with additive function as follows :

$$\begin{array}{lll} \text{Maximize} & E[\,r_1(x_1,u_1) + r_2(x_2,u_2) + \cdots + r_N(x_N,u_N) + r_G(x_{N+1})\,] & \\ \text{subject to} & \text{(i)} \ \ x_{n+1} \sim p(\cdot\,|x_n,u_n) & \\ & \text{(ii)} \ \ u_n \in U \qquad\qquad n = 1,2,\ldots,N & \end{array} \tag{1}$$

where the sequence of states $\{x_2, x_3, \ldots, x_{N+1}\}$ together with a sequence of intermediate actions $\{u_1, u_2, \ldots, u_N\}$ is stochastically generated through an initial state x_1, the Markov transition law $x_{n+1} \sim p(\cdot\,|x_n,u_n)$ and a (general or Markov) policy.

2.1. General Policies

First we consider the problem (1) with the set of all general policies. In general, a *policy* $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$ is a sequence of *decision functions*

$$\sigma_1 : X \to U, \quad \sigma_2 : X \times X \to U, \quad \ldots \ , \quad \sigma_N : \underbrace{X \times \cdots \times X}_{N \text{ times}} \to U. \tag{2}$$

In what follows, in order to distinguish the policy, we call this policy *general* policy. Note that the size in our data specified in (2) yields k^{p^n} *n-th* decision functions σ_n $(n = 1,2,\ldots,N)$ and $k^{p+p^2+\cdots+p^N}$ general policies σ.

An application of general policy σ at an initial state x_1 generates stochastically the alternate sequence of states and actions $\{u_1, x_2, u_2, x_3, \ldots, u_N, x_{N+1}\}$ as follows :

$$\begin{array}{llll} \sigma_1(x_1) = u_1 & \to & p(\cdot\,|x_1,u_1) \sim x_2 & \to \\ \sigma_2(x_1,x_2) = u_2 & \to & p(\cdot\,|x_2,u_2) \sim x_3 & \to \\ \quad\vdots & & \quad\vdots & \\ \sigma_N(x_1,x_2,\ldots,x_N) = u_N & \to & p(\cdot\,|x_N,u_N) \sim x_{N+1}. & \end{array}$$

We call this problem *general problem.* With any general policy $\sigma = \{\sigma_n, \ldots, \sigma_N\}$ over the $(N-n+1)$-stage process starting on n-th stage and terminating at the last stage, we associate the expected value :

$$\begin{aligned} I^n(x_n;\sigma) &= \sum_{(x_{n+1},\ldots,x_N)\in X\times\cdots\times X}\sum\cdots\sum\{[r_n(x_n,u_n)+\cdots+r_N(x_N,u_N)+r_G(x_{N+1})] \\ &\qquad \times p(x_{n+1}|x_n,u_n)\cdots p(x_{N+1}|x_N,u_N)\} \end{aligned} \tag{3}$$

where $\{u_n, x_{n+1}, u_{n+1}, x_{n+2}, \ldots, u_N, x_{N+1}\}$ is stochastically generated through the general policy σ and the starting state x_n as follows :

$$\begin{array}{lll} \sigma_n(x_n) = u_n & \to\ p(\,\cdot\,|x_n,u_n) \sim x_{n+1} & \to \\ \sigma_{n+1}(x_n,x_{n+1}) = u_{n+1} & \to\ p(\,\cdot\,|x_{n+1},u_{n+1}) \sim x_{n+2} & \to \\ \vdots & \quad\vdots & \\ \sigma_N(x_n,x_{n+1},\ldots,x_N) = u_N & \to\ p(\,\cdot\,|x_N,u_N) \sim x_{N+1}. & \end{array} \tag{4}$$

We define the family of the corresponding *general subproblems* as follows :

$$\begin{array}{ll} V^{N+1}(x_{N+1}) = r_G(x_{N+1}) & x_{N+1} \in X \\ V^n(x_n) = \operatorname*{Max}_{\sigma} I^n(x_n;\sigma) & x_n \in X, \quad 1 \le n \le N. \end{array} \tag{5}$$

Note that the general problem (1) is identical to (5) with $n = 0$. Furthermore we should remark that the maximization for the subproblems stated above is taken for all general policies, namely, in problem (5)

$$\sigma_n : X \to U, \quad \sigma_{n+1} : X \times X \to U, \quad \ldots\ , \quad \sigma_N : \underbrace{X \times \cdots \times X}_{N-n+1 \text{ times}} \to U.$$

Then we have the recursive formula for the general subproblems :

THEOREM 2.1.

$$\begin{array}{ll} V^{N+1}(x) = r_G(x) & x \in X \\ V^n(x) = \operatorname*{Max}_{u\in U}[r_n(x,u) + \sum_{y\in X} V^{n+1}(y)p(y|x,u)] & x \in X, \quad 1 \le n \le N. \end{array} \tag{6}$$

2.2. Markov Policies

In this subsection we restrict the problem (1) to the set of all Markov policies. We call this problem *Markov problem.* Here a policy

$$\pi = \{\pi_1, \pi_2, \ldots, \pi_N\}$$

is called *Markov* if

$$\pi_1 : X \to U, \quad \pi_2 : X \to U, \quad \ldots\ , \quad \pi_N : X \to U. \tag{7}$$

Thus, any Markov policy π with an initial state x_1 yields the Markov chain on X and the sequence of the resulting actions as follows :

$$\begin{array}{lll}\pi_1(x_1)=u_1 & \rightarrow\ p(\,\cdot\,|x_1,u_1)\sim x_2 & \rightarrow\\ \pi_2(x_2)=u_2 & \rightarrow\ p(\,\cdot\,|x_2,u_2)\sim x_3 & \rightarrow\\ \quad\vdots & \qquad\quad\vdots & \\ \pi_N(x_N)=u_N & \rightarrow\ p(\,\cdot\,|x_N,u_N)\sim x_{N+1}. & \end{array}$$

We remark that the size in (7) yields k^p *n-th* decision functions π_n $(n=1,2,\ldots,N)$ and k^{Np} Markov policies π.

Note that any Markov policy $\pi=\{\pi_n,\ \ldots,\ \pi_N\}$ over the $(N-n+1)$-stage process is associated with its expected value $I^n(x_n;\pi)$ defined by (3), where the alternate sequence $\{u_n,x_{n+1},u_{n+1},x_{n+2},\ldots,u_N,x_{N+1}\}$ is similarly generated as in (4). Here we remark that

$$u_n=\pi_n(x_n),\quad u_{n+1}=\pi_{n+1}(x_{n+1}),\quad \cdots,u_N=\pi_N(x_N).$$

We define the corresponding *Markov subproblems* as follows :

$$\begin{array}{ll}v^{N+1}(x_{N+1})=r_G(x_{N+1}) & x_{N+1}\in X\\ v^n(x_n)=\operatorname*{Max}_{\pi} I^n(x_n;\pi) & x_n\in X,\ \ 1\le n\le N.\end{array}\tag{8}$$

Then (8) with $n=1$ reduces to the Markov problem (1). We have the recursive formula for the Markov subproblems :

THEOREM 2.2.

$$\begin{array}{ll}v^{N+1}(x)=r_G(x) & x\in X\\ v^n(x)=\operatorname*{Max}_{u\in U}[r_n(x,u)+\sum_{y\in X}v^{n+1}(y)p(y|x,u)] & x\in X,\ \ 1\le n\le N.\end{array}\tag{9}$$

THEOREM 2.3. (i) *A Markov policy yields the optimal value function $V^1(\cdot)$ for the general problem. That is, there exists an optimal Markov policy π^* for the general problem (1) :*

$$I^1(x_1;\pi^*)=V^1(x_1)\quad \textit{for all}\ \ x_1\in X.$$

In fact, letting $\pi^*_n(x)$ be a maximizer of (9) (or (6)) for each $x\in X$, $1\le n\le N$, we have the optimal Markov policy $\pi^*=\{\pi^*_1,\ldots,\pi^*_N\}$.
(ii) *The optimal value functions for the Markov subproblems (8) are equal to the optimal value functions for the general subproblems (5) :*

$$v^n(x)=V^n(x)\qquad x\in X,\ \ 1\le n\le N+1.$$

3. Proof of Theorems

In this subsection we prove Theorems 2.1 - 2.3. We remark that Theorem 2.3 (i) implies Theorem 2.3 (ii) and that a combination of Theorem 2.1 and Theorem 2.3 (ii) yields Theorem 2.2. Thus it suffices to prove Theorem 2.1 and Theorem 2.3 (i). Since there is no essential difficulty in extending the proof to the general N-stage process, we prove both theorems for the two-stage process, namely, for the case $N = 2$.

We note that

$$\begin{aligned} V^3(x_3) &= r_G(x_3) \\ V^2(x_2) &= \operatorname*{Max}_{\sigma_2} \sum_{x_3 \in X} [\,r_2(x_2,u_2) + r_G(x_3)\,]p(x_3|x_2,u_2) \qquad (10) \\ V^1(x_1) &= \operatorname*{Max}_{\sigma_1,\sigma_2} \sum\sum_{(x_2,x_3)\in X\times X} \{[\,r_1(x_1,u_1) + r_2(x_2,u_2) + r_G(x_3)\,] \qquad (11) \\ &\qquad \times p(x_2|x_1,u_1)p(x_3|x_2,u_2)\} \end{aligned}$$

where $u_2 = \sigma_2(x_2)$ in (10) and $u_1 = \sigma_1(x_1),\ u_2 = \sigma_2(x_1,x_2)$ in (11), respectively.

Thus the equality

$$V^2(x_2) = \operatorname*{Max}_{u_2 \in U}[r_2(x_2,u_2) + \sum_{x_3 \in X} V^3(x_3)p(x_3|x_2,u_2)] \quad x_2 \in X$$

is trivial. We prove

$$V^1(x_1) = \operatorname*{Max}_{u_1 \in U}[r_1(x_1,u_1) + \sum_{x_2 \in X} V^2(x_2)p(x_2|x_1,u_1)] \quad x_1 \in X. \qquad (12)$$

Let us choose an optimal (necessarily Markov) policy σ_2^* for the one-stage process :

$$V^2(x_2) = I^2(x_2;\sigma_2^*) \quad \forall x_2 \in X. \qquad (13)$$

From the definition (5), we can for each $x_1 \in X$ choose an optimal (not necessarily Markov) policy $\tilde{\sigma} = \{\tilde{\sigma}_1,\tilde{\sigma}_2\}$ for the two-stage process :

$$V^1(x_1) = I^1(x_1;\tilde{\sigma}) \quad x_1 \in X.$$

Thus we see that

$$\begin{aligned} & V^1(x_1) \\ =\ & I^1(x_1;\tilde{\sigma}_1,\tilde{\sigma}_2) \\ =\ & \sum\sum_{(x_2,x_3)\in X\times X} \{[r_1(x_1,u_1) + r_2(x_2,u_2) + r_G(x_3)]p(x_2|x_1,u_1)p(x_3|x_2,u_2)\} \qquad (14) \end{aligned}$$

where

$$u_1 = \tilde{\sigma}_1(x_1),\ u_2 = \tilde{\sigma}_2(x_1,x_2).$$

Since

$$\begin{aligned} & \sum\sum_{(x_2,x_3)\in X\times X} \{[r_1(x_1,u_1) + r_2(x_2,u_2) + r_G(x_3)]p(x_2|x_1,u_1)p(x_3|x_2,u_2)\} \\ =\ & \sum_{x_2 \in X} \{r_1(x_1,u_1) + \sum_{x_3 \in X} [r_2(x_2,u_2) + r_G(x_3)]p(x_3|x_2,u_2)\}p(x_2|x_1,u_1) \end{aligned}$$

and

$$\sum_{x_3\in X}[r_2(x_2,u_2)+r_G(x_3)]p(x_3|x_2,u_2)\leq I^2(x_2;\sigma_2^*)=V^2(x_2)\quad \forall x_2\in X,$$

we have from (14)

$$\begin{aligned} V^1(x_1) &\leq \sum_{x_2\in X}[r_1(x_1,u_1)+V^2(x_2)]p(x_2|x_1,u_1) \\ &= r_1(x_1,u_1)+\sum_{x_2\in X}V^2(x_2)p(x_2|x_1,u_1). \end{aligned}$$

Thus taking maximum over $u\in U$, we get

$$V^1(x_1)\leq \operatorname*{Max}_{u_1\in U}[r_1(x_1,u_1)+\sum_{x_2\in X}V^2(x_2)p(x_2|x_1,u_1)]\quad \forall x_1\in X. \tag{15}$$

On the other hand, let for any $x_1\in X$, $u^*=u^*(x_1)\in U$ be a maximizer of the right hand side of (15). This defines a Markov decision function

$$\pi_1^*:X\to U\quad \pi_1^*(x_1)=u^*(x_1).$$

Then we have

$$\begin{aligned} &\operatorname*{Max}_{u_1\in U}[r_1(x_1,u_1)+\sum_{x_2\in X}V^2(x_2)p(x_2|x_1,u_1)] \\ = \ & r_1(x_1,u_1)+\sum_{x_2\in X}V^2(x_2)p(x_2|x_1,u_1)\quad (\,u_1=\pi_1^*(x_1)\,). \end{aligned}\tag{16}$$

From (13), we get

$$V^2(x_2)=\sum_{x_3\in X}[r_2(x_2,u_2)+r_G(x_3)]p(x_3|x_2,u_2)\quad (\,u_2=\sigma_2^*(x_2)\,). \tag{17}$$

Thus we have from (17)

$$\begin{aligned} &r_1(x_1,u_1)+\sum_{x_2\in X}V^2(x_2)p(x_2|x_1,u_1)\quad (\,u_1=\pi_1^*(x_1)\,) \\ = \ & r_1(x_1,u_1)+\sum_{x_2\in X}[\sum_{x_3\in X}[r_2(x_2,u_2)+r_G(x_3)]p(x_3|x_2,u_2)]p(x_2|x_1,u_1) \\ = \ & \sum\sum_{(x_2,x_3)\in X\times X}\{[r_1(x_1,u_1)+r_2(x_2,u_2)+r_G(x_3)]p(x_2|x_1,u_1)p(x_3|x_2,u_2)\}. \end{aligned}\tag{18}$$

Combining (16) and (18), we obtain

$$\begin{aligned} &\operatorname*{Max}_{u_1\in U}[r_1(x_1,u_1)+\sum_{x_2\in X}V^2(x_2)p(x_2|x_1,u_1)] \\ = \ & \sum\sum_{(x_2,x_3)\in X\times X}\{[r_1(x_1,u_1)+r_2(x_2,u_2)+r_G(x_3)]p(x_2|x_1,u_1)p(x_3|x_2,u_2)\} \end{aligned}$$

$$\begin{aligned} & \qquad\qquad\qquad\qquad (u_1 = \pi_1^*(x_1)\,,\ u_2 = \sigma_2^*(x_2)\,) \\ \leq\ & \operatorname*{Max}_{\sigma_1,\sigma_2} \sum_{(x_2,x_3)\in X\times X} \{[r_1(x_1,u_1) + r_2(x_2,u_2) + r_G(x_3)] \\ & \qquad\qquad \times p(x_2|x_1,u_1)p(x_3|x_2,u_2)\} \\ =\ & V^1(x_1). \end{aligned} \tag{19}$$

Both equations (15) and (19) imply the desired equality (12). This completes the proof of Theorem 2.1.

Furthermore, the equalities in (19) imply that the optimal value function $V^1(\,\cdot\,)$ is yielded by the Markov policy $\bar{\pi} = \{\pi_1^*, \sigma_2^*\}$:

$$V^1(x_1) = I^1(x_1;\bar{\pi}) \quad x_1 \in X.$$

This completes the proof of Theorem 2.3 (i).

4. Deterministic Maximization

In this section we consider the deterministic maximization problem with additive function as follows :

$$\begin{array}{lll} \text{Maximize} & r_1(x_1,u_1) + r_2(x_2,u_2) + \cdots + r_N(x_N,u_N) + r_G(x_{N+1}) & \\ \text{subject to} & \text{(i)}\ \ f(x_n,u_n) = x_{n+1} & \\ & \text{(ii)}\ \ u_n \in U \qquad\qquad n = 1,2,\ldots,N. & \end{array} \tag{20}$$

Note that this problem is the special case of the stochastic maximization problem (1). Through this section the problem (20) with the set of all general (resp. Markov) policies

$$\sigma = \{\sigma_1,\ \sigma_2,\ \ldots,\ \sigma_N\} \qquad (\text{resp. } \pi = \{\pi_1,\ \pi_2,\ \ldots,\ \pi_N\}\ )$$

is called the general (resp. Markov) problem.

First we consider the general problem. We associate any policy $\sigma = \{\sigma_n,\ \ldots,\ \sigma_N\}$ for the $(N-n+1)$-stage process starting on n-th stage and terminating at the last stage with its value :

$$I^n(x_n;\sigma) = r_n(x_n,u_n) + \cdots + r_N(x_N,u_N) + r_G(x_{N+1}) \tag{21}$$

where $\{u_n, x_{n+1}, u_{n+1}, x_{n+2}, \ldots, u_N, x_{N+1}\}$ is uniquely determined through σ and x_n as follows :

$$\begin{array}{lcll} \sigma_n(x_n) = u_n & \longrightarrow & f(x_n,u_n) = x_{n+1} & \longrightarrow \\ \sigma_{n+1}(x_n,x_{n+1}) = u_{n+1} & \longrightarrow & f(x_{n+1},u_{n+1}) = x_{n+2} & \longrightarrow \\ \vdots & & \vdots & \\ \sigma_N(x_1,x_1,\ldots,x_N) = u_N & \longrightarrow & f(x_N,u_N) = x_{N+1}. & \end{array} \tag{22}$$

We consider the following family of subproblems :

$$\begin{array}{ll} V^{N+1}(x_{N+1}) = r_G(x_{N+1}) & x_{N+1} \in X \\ V^n(x_n) = \operatorname*{Max}_{\sigma} I^n(x_n;\sigma) & x_n \in X,\ \ 1 \leq n \leq N. \end{array} \tag{23}$$

Note that the general problem is identical to (23) with $n = 0$. Then we have the backward recursive formula :

THEOREM 4.1.

$$\begin{aligned} &V^{N+1}(x) = r_G(x) && x \in X \\ &V^n(x) = \operatorname*{Max}_{u \in U}[r_n(x,u) + V^{n+1}(f(x,u))] && x \in X, \quad 1 \le n \le N. \end{aligned} \tag{24}$$

Next we restrict the problem (20) to the set of all Markov policies. Note that any Markov policy $\pi = \{\pi_n, \ldots, \pi_N\}$ over the $(N-n+1)$-stage process is associated with its value $I^n(x_n; \pi)$ defined by (21), where the alternate sequence $\{u_n, x_{n+1}, u_{n+1}, x_{n+2}, \ldots, u_N, x_{N+1}\}$ is similarly determined through the Markov policy π and the starting state x_n as in (22).

We define the corresponding *Markov subproblems* as follows :

$$\begin{aligned} &v^{N+1}(x_{N+1}) = r_G(x_{N+1}) && x_{N+1} \in X \\ &v^n(x_n) = \operatorname*{Max}_{\pi} I^n(x_n; \pi) && x_n \in X, \quad 1 \le n \le N. \end{aligned} \tag{25}$$

Then (25) with $n = 1$ reduces to the Markov problem. We have the recursive formula for the Markov subproblems :

THEOREM 4.2.

$$\begin{aligned} &v^{N+1}(x) = r_G(x) && x \in X \\ &v^n(x) = \operatorname*{Max}_{u \in U}[r_n(x,u) + v^{n+1}(f(x,u))] && x \in X, \quad 1 \le n \le N. \end{aligned} \tag{26}$$

THEOREM 4.3. (i) *A Markov policy yields the optimal value function $V^1(\cdot)$ for the general problem. That is, there exists an optimal Markov policy π^* for the general problem (20) :*

$$I^1(x_1; \pi^*) = V^1(x_1) \quad \textit{for all} \ \ x_1 \in X.$$

In fact, letting $\pi_n^*(x)$ be a maximizer of (26) (or (24)) for each $x \in X$, $1 \le n \le N$, we have the optimal Markov policy $\pi^* = \{\pi_1^*, \ldots, \pi_N^*\}$.
(ii) *The optimal value functions for the Markov subproblems (25) are equal to the optimal value functions for the general subproblems (23) :*

$$v^n(x) = V^n(x) \qquad x \in X, \quad 1 \le n \le N+1.$$

Since Theorems 4.1 - 4.3 are special cases of Theorems 2.1 - 2.3 respectively, each theorem in this section is clear.

5. Example

We illustrate the following two-stage, three-state and two-action stochastic decision process :

$$\begin{aligned} \text{Maximize} \quad & E[\,r_1(u_1) + r_2(u_2) + r_G(x_3)\,] \\ \text{subject to} \quad & \text{(i)} \;\; x_{n+1} \sim p(\cdot\,|x_n, u_n) \qquad n = 1, 2 \\ & \text{(ii)} \;\; u_1 \in U, \ u_2 \in U \end{aligned} \tag{27}$$

where the data is

$$r_G(s_1) = 0.3 \qquad r_G(s_2) = 1.0 \qquad r_G(s_3) = 0.8 \tag{28}$$

$$r_2(a_1) = 1.0 \qquad r_2(a_2) = 0.6$$

$$r_1(a_1) = 0.7 \qquad r_1(a_2) = 1.0$$

$u_t = a_1$

$x_t \backslash x_{t+1}$	s_1	s_2	s_3
s_1	0.8	0.1	0.1
s_2	0.0	0.1	0.9
s_3	0.8	0.1	0.1

$u_t = a_2$

$x_t, \backslash x_{t+1}$	s_1	s_2	s_3
s_1	0.1	0.9	0.0
s_2	0.8	0.1	0.1
s_3	0.1	0.0	0.9

In order to solve the problem, we directly generate one- and two- stage stochastic decision trees and enumerate all the possible histories together with the related expected values. We call this brute force enumeration method a *multi-stage stochastic decision tree method*. For any given policy, this tree method traces all the resulting histories. Then it yields the value of the policy. Further, from among all general policies, it selects an optimal policy together with the sequence of optimal value functions. The multi-stage stochastic decision tree method is also applies to non-additive problems (Iwamoto and Fujita(1995), Iwamoto, Tsurusaki and Fujita).

We remark that the size yields $2^3 = 8$ first decision functions $\sigma_1 = \begin{pmatrix} \sigma_1(s_1) \\ \sigma_1(s_2) \\ \sigma_1(s_3) \end{pmatrix}$ and $2^{3\times 3} = 512$ second decision functions

$$\sigma_2 = \begin{pmatrix} \sigma_2(s_1, s_1) & \sigma_2(s_2, s_1) & \sigma_2(s_3, s_1) \\ \sigma_2(s_1, s_2) & \sigma_2(s_2, s_2) & \sigma_2(s_3, s_2) \\ \sigma_2(s_1, s_3) & \sigma_2(s_2, s_3) & \sigma_2(s_3, s_3) \end{pmatrix}.$$

As a total, there are 8 × 512 = 4096 general policies $\sigma = \{\sigma_1, \sigma_2\}$ for the problem (27).

First, we have from definition (28)

$$V^3(s_1) = 0.3, \quad V^3(s_2) = 1.0, \quad V^3(s_3) = 0.8. \tag{29}$$

Second, the decision tree in Figure 1 shows

$$V^2(s_1) = 1.53, \quad V^2(s_2) = 1.82, \quad V^2(s_3) = 1.42. \tag{30}$$

Third, the enumeration in Figures 2, 3 and 4 calculates the maximum expected values :

$$V^1(s_1) = 2.791, \quad V^1(s_2) = 2.548, \quad V^1(s_3) = 2.431. \tag{31}$$

The calculation yields, at the same time, the optimal policy $\sigma^* = \{\sigma_1^*(x_1),\ \sigma_2^*(x_1, x_2)\}$:

$$\sigma_1^*(s_1) = a_2, \quad \sigma_1^*(s_2) = a_2, \quad \sigma_1^*(s_3) = a_2$$

$$\begin{array}{lll}\sigma_2^*(s_1,s_1)=a_2, & \sigma_2^*(s_2,s_1)=a_2, & \sigma_2^*(s_3,s_1)=a_2\\ \sigma_2^*(s_1,s_2)=a_1, & \sigma_2^*(s_2,s_2)=a_1 \text{ or } a_2, & \sigma_2^*(s_3,s_2)=a_1\\ \sigma_2^*(s_1,s_3)=a_1 \text{ or } a_2, & \sigma_2^*(s_2,s_3)=a_1, & \sigma_2^*(s_3,s_3)=a_1.\end{array}$$

Thus the general policy σ^* reduces a Markov policy $\pi^* = \{\pi_1^*(x_1),\ \pi_2^*(x_2)\}$:

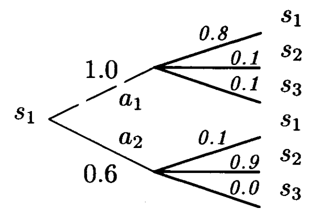
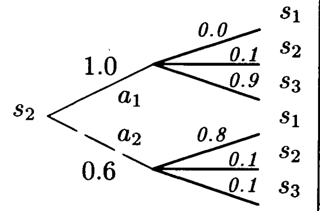
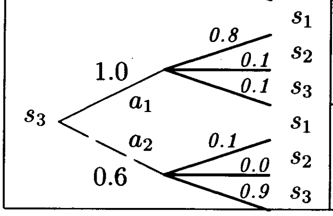
$$\pi_1^*(s_1)=a_2,\quad \pi_1^*(s_2)=a_2,\quad \pi_1^*(s_3)=a_2$$

$$\pi_2^*(s_1)=a_2,\quad \pi_2^*(s_2)=a_1,\quad \pi_2^*(s_3)=a_1.$$

Thus, the Markov policy π^* is optimal. Finally we remark that the pair of optimal value functions (29),(30),(31) and the optimal Markov policy π^* is also obtained by solving either the corresponding recursive equation (6) or (9). Solving the latter is the so-called dynamic programming method.

Figure 1 : all one-stage behaviors from s_1, s_2 and s_3, and selection of maximum branch

$$\left(V^2(x_2)=\underset{u_2}{\text{Max}}\sum_{x_3\in X}[\,r_2(u_2)+r_G(x_3)\,]p(x_3|x_2,u_2)\quad x_2=s_1,\ s_2,\ s_3\right)$$

history	ter.	path	sum	mult.	exp.
	0.3	*0.8*	1.3	1.04	1.42
	1.0	*0.1*	2.0	0.2	
	0.8	*0.1*	1.8	0.18	
	0.3	*0.1*	0.9	0.09	**1.53**
	1.0	*0.9*	1.6	1.44	
	0.8	*0.0*	1.4	0	
	0.3	*0.0*	1.3	0	**1.82**
	1.0	*0.1*	2.0	0.2	
	0.8	*0.9*	1.8	1.62	
	0.3	*0.8*	0.9	0.72	1.02
	1.0	*0.1*	1.6	0.16	
	0.8	*0.1*	1.4	0.14	
	0.3	*0.8*	1.3	1.04	**1.42**
	1.0	*0.1*	2.0	0.2	
	0.8	*0.1*	1.8	0.18	
	0.3	*0.1*	0.9	0.09	1.35
	1.0	*0.0*	1.6	0	
	0.8	*0.9*	1.4	1.26	

In Figure 1 we use the following list of simplified notations :

history = $x_2 \;\; r_2(u_2) \,/\, u_2 \;\; p(x_3 \mid x_2, u_2) \;\; x_3$
ter. = terminal value = $r_G(x_3)$
path = path probability = $p(x_3 \mid x_2, u_2)$
sum = sum of the two = $r_2(u_2) + r_G(x_3)$
mult. = path × sum
exp. = expected value.

Further, the *italic* face means probability, and the **bold** face denotes a selection of maximum of up expected value or down.

Figure 2 : all two-stage behaviors from s_1 and selection of maximum branch

$$\left(V^1(s_1) = \mathop{\mathrm{Max}}_{u_1,u_2} \sum_{(x_2,x_3)\in X\times X} \{[r_1(u_1)+r_2(u_2)+r_G(x_3)]p(x_2|s_1,u_1)p(x_3|x_2,u_2)\}\right)$$

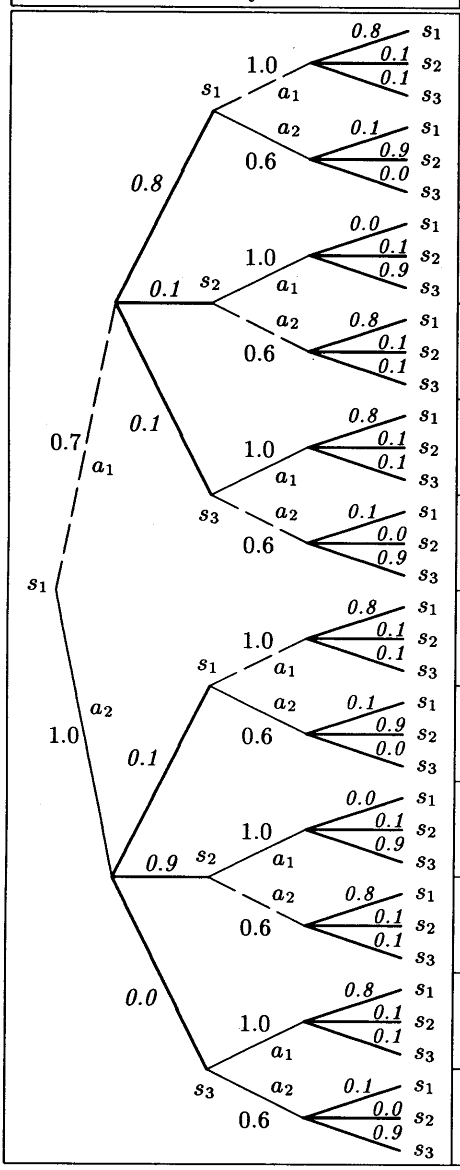
history	ter.	path	sum	mult.	sub.	total
	0.3	*0.64*	2.0	1.28	1.696	2.248
	1.0	*0.08*	2.7	0.216		
	0.8	*0.08*	2.5	0.2		
	0.3	*0.08*	1.6	0.128	**1.784**	
	1.0	*0.72*	2.3	1.656		
	0.8	*0.0*	2.1	0		
	0.3	*0.0*	2.0	0	**0.252**	
	1.0	*0.01*	2.7	0.027		
	0.8	*0.09*	2.5	0.225		
	0.3	*0.08*	1.6	0.128	0.172	
	1.0	*0.01*	2.3	0.023		
	0.8	*0.01*	2.1	0.021		
	0.3	*0.08*	2.0	0.16	**0.212**	
	1.0	*0.01*	2.7	0.027		
	0.8	*0.01*	2.5	0.025		
	0.3	*0.01*	1.6	0.016	0.205	
	1.0	*0.0*	2.3	0		
	0.8	*0.09*	2.1	0.189		
	0.3	*0.08*	2.3	0.184	0.242	**2.791**
	1.0	*0.01*	3.0	0.03		
	0.8	*0.01*	2.8	0.028		
	0.3	*0.01*	1.9	0.019	**0.253**	
	1.0	*0.09*	2.6	0.234		
	0.8	*0.0*	2.4	0		
	0.3	*0.0*	2.3	0	**2.538**	
	1.0	*0.09*	3.0	0.27		
	0.8	*0.81*	2.8	2.268		
	0.3	*0.72*	1.9	1.368	1.818	
	1.0	*0.09*	2.6	0.234		
	0.8	*0.09*	2.4	0.216		
	0.3	*0.0*	2.3	0	**0**	
	1.0	*0.0*	3.0	0		
	0.8	*0.0*	2.8	0		
	0.3	*0.0*	1.9	0	**0**	
	1.0	*0.0*	2.6	0		
	0.8	*0.0*	2.4	0		

Figure 3 : all two-stage behaviors from s_2 and selection of maximum branch

$$\left(V^1(s_2) = \underset{u_1,u_2}{\mathrm{Max}} \sum_{(x_2,x_3)\in X\times X} \{[r_1(u_1)+r_2(u_2)+r_G(x_3)]p(x_2|s_2,u_1)p(x_3|x_2,u_2)\} \right)$$

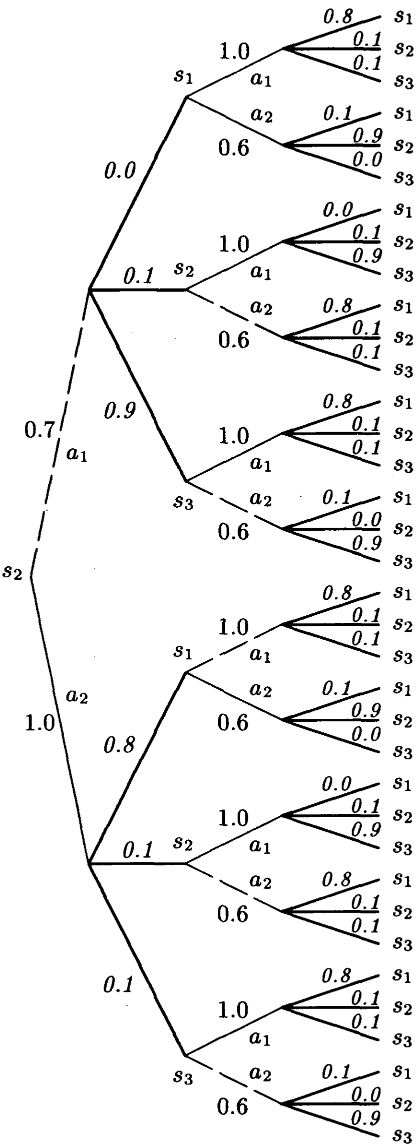
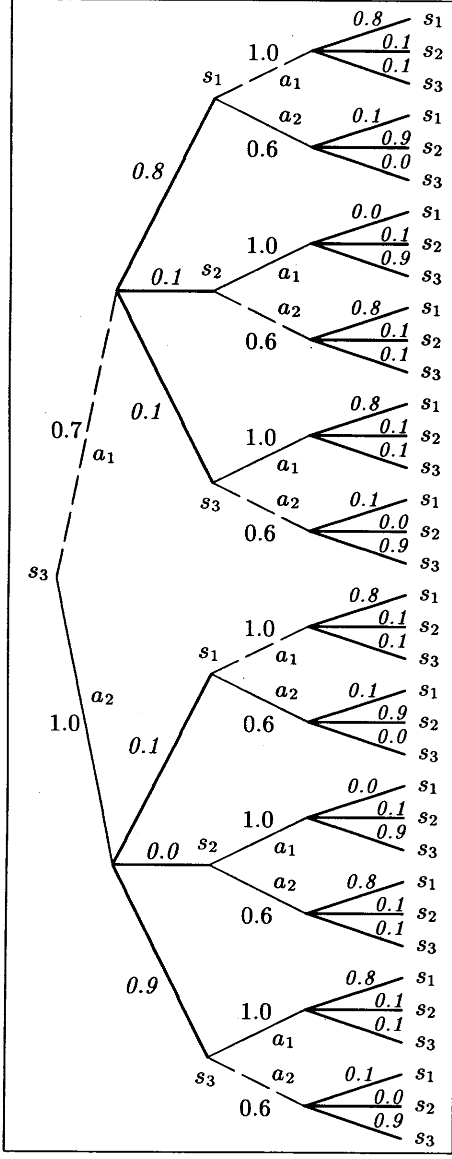
history	ter.	path	sum	mult.	sub.	total
	0.3	*0.0*	2.0	0	**0**	2.16
	1.0	*0.0*	2.7	0		
	0.8	*0.0*	2.5	0		
	0.3	*0.0*	1.6	0	**0**	
	1.0	*0.0*	2.3	0		
	0.8	*0.0*	2.1	0		
	0.3	*0.0*	2.0	0	**0.252**	
	1.0	*0.01*	2.7	0.027		
	0.8	*0.09*	2.5	0.225		
	0.3	*0.08*	1.6	0.128	0.172	
	1.0	*0.01*	2.3	0.023		
	0.8	*0.01*	2.1	0.021		
	0.3	*0.72*	2.0	1.44	**1.908**	
	1.0	*0.09*	2.7	0.243		
	0.8	*0.09*	2.5	0.225		
	0.3	*0.09*	1.6	0.144	1.845	
	1.0	*0.0*	2.3	0		
	0.8	*0.81*	2.1	1.701		
	0.3	*0.64*	2.3	1.472	1.936	**2.548**
	1.0	*0.08*	3.0	0.24		
	0.8	*0.08*	2.8	0.224		
	0.3	*0.08*	1.9	0.152	**2.024**	
	1.0	*0.72*	2.6	1.872		
	0.8	*0.0*	2.4	0		
	0.3	*0.0*	2.3	0	**0.282**	
	1.0	*0.01*	3.0	0.03		
	0.8	*0.09*	2.8	0.252		
	0.3	*0.08*	1.9	0.152	0.202	
	1.0	*0.01*	2.6	0.026		
	0.8	*0.01*	2.4	0.024		
	0.3	*0.08*	2.3	0.184	**0.242**	
	1.0	*0.01*	3.0	0.03		
	0.8	*0.01*	2.8	0.028		
	0.3	*0.01*	1.9	0.019	0.235	
	1.0	*0.0*	2.6	0		
	0.8	*0.09*	2.4	0.216		

Figure 4 : all two-stage behaviors from s_3 and selection of maximum branch

$$\left(V^1(s_3) = \operatorname*{Max}_{u_1,u_2} \sum_{(x_2,x_3)\in X\times X} \{[r_1(u_1)+r_2(u_2)+r_G(x_3)]p(x_2|s_3,u_1)p(x_3|x_2,u_2)\}\right)$$



history	ter.	path	sum	mult.	sub.	total
	0.3	*0.64*	2.0	1.28	1.696	2.248
	1.0	*0.08*	2.7	0.216		
	0.8	*0.08*	2.5	0.2		
	0.3	*0.08*	1.6	0.128	**1.784**	
	1.0	*0.72*	2.3	1.656		
	0.8	*0.0*	2.1	0		
	0.3	*0.0*	2.0	0	**0.252**	
	1.0	*0.01*	2.7	0.027		
	0.8	*0.09*	2.5	0.225		
	0.3	*0.08*	1.6	0.128	0.172	
	1.0	*0.01*	2.3	0.023		
	0.8	*0.01*	2.1	0.021		
	0.3	*0.08*	2.0	0.16	**0.212**	
	1.0	*0.01*	2.7	0.027		
	0.8	*0.01*	2.5	0.025		
	0.3	*0.01*	1.6	0.016	0.205	
	1.0	*0.0*	2.3	0		
	0.8	*0.09*	2.1	0.189		
	0.3	*0.08*	2.3	0.184	0.242	**2.431**
	1.0	*0.01*	3.0	0.03		
	0.8	*0.01*	2.8	0.028		
	0.3	*0.01*	1.9	0.019	**0.253**	
	1.0	*0.09*	2.6	0.234		
	0.8	*0.0*	2.4	0		
	0.3	*0.0*	2.3	0	**0**	
	1.0	*0.0*	3.0	0		
	0.8	*0.0*	2.8	0		
	0.3	*0.0*	1.9	0	**0**	
	1.0	*0.0*	2.6	0		
	0.8	*0.0*	2.4	0		
	0.3	*0.72*	2.3	1.656	**2.178**	
	1.0	*0.09*	3.0	0.27		
	0.8	*0.09*	2.8	0.252		
	0.3	*0.09*	1.9	0.171	2.115	
	1.0	*0.0*	2.6	0		
	0.8	*0.81*	2.4	1.944		

In Figures 2, 3 and 4 we use the following notations:

history = $x_1 \;\; r_1(u_1) \;/\; u_1 \;\; p(x_2 \mid x_1, u_1) \;\; x_2 \;\; r_2(u_2) \;/\; u_2 \;\; p(x_3 \mid x_2, u_2) \;\; x_3$
ter. = terminal value = $r_G(x_3)$
path = path probability = $p(x_2 \mid x_1, u_1)p(x_3 \mid x_2, u_2)$
sum = sum of the three = $r_1(u_1) + r_2(u_2) + r_G(x_3)$
mult. = path × sum
sub. = subtotal expected value
total = total expected value.

Acknowledgments

The author wishes to thank Professor Seiichi Iwamoto for valuable advice on this investigation. The author would like to thank the anonymous referee for careful reading of this paper and helpful comments.

References

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

Bertsekas, D.P. and Shreve, S.E. (1978), *Stochastic Optimal Control*, Academic Press, New York.

Howard, R.A. (1960), *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, Mass.

Iwamoto, S. and Fujita, T. (1995), Stochastic decision-making in a fuzzy environment, *J. Operations Res. Soc. Japan*, 38:467–482.

Iwamoto, S., Tsurusaki, K. and Fujita, T. , On Markov policies for minimax decision processes, *submitted.*

Puterman, M.L. (1994), *Markov Decision Processes : discrete stochastic dynamic programming*, Wiley & Sons, New York.

Stokey, N.L. and Lucas Jr., R.E. (1989), *Recursive Methods in Economic Dynamics*, Harvard Univ. Press, Cambridge, MA.

Received October 11, 1996
Revised January 31, 1997