A dual approach in optimizing threshold probabilities

岩本, 誠一
九州大学大学院経済学研究院 経済工学部門 : 教授

植野, 貴之
長崎県立大学経済学部経済学科 : 講師

https://doi.org/10.15017/10601

出版情報：經濟學研究. 73 (1), pp.19-33, 2006-08-25. 九州大学経済学会
バージョン：
権利関係：

A dual approach in optimizing threshold probabilities

Seiichi Iwamoto and Takayuki Ueno

Abstract

We consider a threshold probability optimization problem over controlled Markov chains. The problem is which class of policies we optimize the threshold probability in and how we find an optimal policy. This paper formulates the optimization problem in general (large) class and presents a pair of primal and dual methods. A primal method is based upon state-expansion with cumulative rewards up to date and a dual is with threshold levels for the remaining process. We derive duality theorem and consistency theorem, which show that optimal solutions characterize each other. Further a typical model with Bellman and Zadeh's data is illustrated.

1 Introduction

Since Bellman and Zadeh [3], there has been a broad study on fuzzy decision theory and its applications ([4,6,17–19] and others). They have originated deterministic, stochastic and fuzzy systems on multistage decision processes in fuzzy environment [3, §4,5]. We focus on the stochastic system.

It is well known that there exists an optimal policy which is *Markov* for *additive* criterion over Markov decision processes ([8,21] and others). In fact, an optimal solution is easily derived by solving the recursive equation, which turns out to be valid because of *linearity* of expectation operator for additive criterion. However, as for *minimum* criterion – a *nonadditive* one –, there does not always exist an optimal policy in *Markov (small) class*. The linearity is of no use for nonadditive criteria. Then the problem is how we obtain a *valid* recursive formula for nonadditive criteria, which constitutes a wide class of *associative* criteria.

Recently Iwamoto has shown that an optimal policy exists for minimum criterion in *general (large) class* [9–16]. The method is an invariant imbedding [1,2,20,22,23] based upon state expansion in stochastic optimization.

In this paper, we consider the "threshold probability" criterion over controlled Markov chains. Our problem is twofold: (1) which class of policies we optimize the threshold probability in and (2) how we find optimal policy. This paper formulates the optimization problem in general class and proposes a pair primal and dual methods. A key idea of primal method is an observation that the reward accumulation is additive. The primal method translates the reward-accumulation process into one additional state-transformation. It is also important to observe that the threshold-level process is subtractive. The dual method translates the threshold-level process into the other.

In Section 2 we describe the problem and a formulation. In Section 3 a primal method is presented. We show how the original controlled Markov chain is reduced to an equivalent controlled Markov chain on expanded state spaces with past values. In Section 4 we propose a dual method based upon state-expansion with future ones. In Section 5 we give duality theorem and consistency theorem, which show that optimal solution of one problem characterizes completely that of the other. Both methods yield the desired (common) optimal solution for the original problem. Finally, in Section 6, a three-state two-decision two-stage model with Bellman and Zadeh's data is illustrated.

2 Decision Process with Threshold Probability

Throughout the paper, the following data is given :

$N \geq 2$ is an integer; the *total number of stages*
$X = \{s_1, s_2, \ldots, s_l\}$ 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* $(0 \leq n \leq N-1)$
$k : X \to R^1$ is a *terminal function*
$\underline{c} \in R^1$ is a *lower level*
$p = \{p(\cdot|\cdot,\cdot)\}$ 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 $p(y|x,u)$.

Let an N-stage controlled Markov chain $\{(X_n, U_n)\}$ on finite state space X and finite decision space U be under a Markov transition law p. We maximize the threshold probability that the total additive reward (random variable)

$$r_0(X_0, U_0) + r_1(X_1, U_1) + \cdots + r_{N-1}(X_{N-1}, U_{N-1}) + k(X_N)$$

is greater than or equal to a given lower level (constant) $\underline{c}$.

The problem is to how to find (optimal) policy which maximizes the threshold probability $P(r_0 + r_1 + \cdots + r_{N-1} + k \geq \underline{c})$. We focus our attention on two view points. One is concerned with policy classes where the optimization should be taken. The other is a dual method which solve the threshold probability problems.

Now let us introduce a large class of policies, which depend not only on today's state but also on state-to-date. Let $X^n := X \times X \times \cdots \times X$ be direct product of n state spaces X. A mapping $\sigma_n : X^{n+1} \to U$ is called *n-th general decision function*, whose sequence $\sigma = \{\sigma_0, \sigma_1, \ldots, \sigma_{N-1}\}$ is a *general policy*. The set of all general policies Π_g is called *general class*. When each general decision function σ_n depends only on the last (= current) state, the general policy reduces to a Markov policy $\pi = \{\pi_0, \pi_1, \ldots, \pi_{N-1}\}$. Let Π denote

the set of all Markov policies. Then $\Pi \subset \Pi_g$ holds. We choose the general class Π_g as policy class where optimization is taken, because of nonexistence of optimal policy in Markov class Π.

Now we consider the maximization problem of thereshold probability over general class Π_g :

$$\begin{array}{lll} & \text{Maximize} & P_{x_0}^{\sigma}(r_0 + r_1 + \cdots + r_{N-1} + k \geq \underline{c}) \\ \mathrm{Q}_0(x_0) & \text{subject to} & (\mathrm{i})_{\mathrm{n}} \quad X_{n+1} \sim p(\cdot\,|x_n, u_n) \qquad n = 0, 1, \ldots, N-1 \\ & & (\mathrm{ii})_{\mathrm{n}} \quad u_n \in U \end{array}$$

where $P_{x_0}^{\sigma}$ is the (discrete) probability measure on history space

$$H_N := X \times U \times X \times U \times \cdots \times U \times X \quad (2N+1)\text{-factors}$$

determined uniquely through an initial state x_0, the Markov transition law p and a general policy $\sigma(\in \Pi_g)$. (For a different formulation of threshold probability optimization, see [5, 24].)

Any general policy $\sigma(\in \Pi_g)$ determines the threshold probability in $\mathrm{Q}_0(x_0)$, which is a "partial" multiple sum :

$$\begin{aligned} & P_{x_0}^{\sigma}(r_0 + r_1 + \cdots + r_{N-1} + k \geq \underline{c}) \qquad (1) \\ = & \sum \sum \cdots \sum_{(x_1, x_2, \ldots, x_N) \in (*)} p(x_1|x_0, u_0) p(x_2|x_1, u_1) \cdots p(x_N|x_{N-1}, u_{N-1}) \end{aligned}$$

where the domain $(*)$ is the set of all $(x_1, x_2, \ldots, x_N) \in X^N$ satisfying

$$r_0(x_0, u_0) + r_1(x_1, u_1) + \cdots + r_{N-1}(x_{N-1}, u_{N-1}) + k(x_N) \geq \underline{c} \qquad (2)$$

Here the sequence of decisions $\{u_0, u_1, \ldots, u_{N-1}\}$ in (1),(2) is uniquely determined through general policy $\sigma = \{\sigma_0, \ldots, \sigma_{N-1}\}$:

$$u_0 = \sigma_0(x_0), \; u_1 = \sigma_1(x_0, x_1), \; \ldots, \; u_{N-1} = \sigma_{N-1}(x_0, x_1, \ldots, x_{N-1}).$$

Thus our optimization problem is to find the *optimum value function* $v_0 : X \to R^1$ defined by

$$v_0(x_0) := \operatorname*{Max}_{\sigma \in \Pi_g} P_{x_0}^{\sigma}(r_0 + \cdots + r_{N-1} + k \geq \underline{c}) \qquad (3)$$

and an *optimal policy* $\hat{\sigma} \in \Pi_g$ which attains the optimum value for all initial state :

$$v_0(x_0) = P_{x_0}^{\hat{\sigma}}(r_0 + \cdots + r_{N-1} + k \geq \underline{c}) \qquad x_0 \in X. \qquad (4)$$

3 Primal Approach

In this section, we show that an optimal policy is obtained in general class through optimization in one expanded Markov class. We transform the original problem $Q_0(x_0)$ with the fixed level $\underline{c}$ into an controlled Markov chain on new state space augmented with past values [11].

First let us introduce the sequence of additional one-dimensional state variables $\{\lambda_n\}_0^N$ called *cumulative rewards* :

$$\begin{aligned} \lambda_0 &:= 0 \\ \lambda_n &:= r_0 + r_1 + \cdots + r_{n-1}. \end{aligned}$$

This is equivalent to the sequential dynamics :

$$(\mathrm{i})'_{\mathrm{n}} \quad \lambda_{n+1} = \lambda_n + r_n(x_n, u_n) \quad n = 0, \ldots, N\text{–}1, \quad \lambda_0 = 0.$$

Then we have the *terminal* expression

$$r_0 + r_1 + \cdots + r_{N-1} + k = \lambda_N + k.$$

This yields the same probability

$$P(r_0 + r_1 + \cdots + r_{N-1} + k \geq \underline{c}) \;=\; P(\,\lambda_N + k(X_N) \geq \underline{c}).$$

under appropriate conditions. Thus the introduction of cumulative rewards has converted the *additive* threshold probability in $Q_0(x_0)$ into the same *terminal* one.

Second we define the sequence of *cumulative reward sets* $\{\Lambda_n\}$:

$$\begin{aligned} \Lambda_0 &\triangleq \{\lambda_0 \,|\, \lambda_0 = 0\} \\ \Lambda_n &\triangleq \left\{ \lambda_n \;\middle|\; \begin{array}{l} \lambda_n = r_0(x_0, u_0) + \cdots + r_{n-1}(x_{n-1}, u_{n-1}) \\ \quad (x_0, u_0, \ldots, x_{n-1}, u_{n-1}) \in X\times U\times\cdots\times X\times U \end{array} \right\} \\ & \qquad\qquad n = 1, \ldots, N. \end{aligned} \tag{5}$$

Thus Λ_n denotes the set of all possible cumulative rewards up to n-th stage i.e., for the stage-interval $[0, n)$. Then we have the forward recursive formula :

Lemma 3.1

$$\begin{aligned} \Lambda_0 &= \{0\} \\ \Lambda_{n+1} &= \{\lambda + r_n(x, u) \,|\, \lambda \in \Lambda_n,\ (x, u) \in X\times U\} \quad 0 \leq n \leq N\text{–}1. \end{aligned} \tag{6}$$

Finally we introduce a new controlled Markov chain on the expanded state spaces $\{X\times\Lambda_n\}_0^N$. Here the state variables $\{(X_n; \lambda_n)\}$ behave such that the first component $\{X_n\}$ obeys the original Markov transition law p and the second $\{\lambda_n\}$ follows the deterministic dynamics $\lambda_{n+1} := \lambda_n + r_n(x_n, u_n)$. When the decision-maker chooses a decision $u_n(\in U)$ on $(x_n; \lambda_n)(\in X\times\Lambda_n)$ at n-th stage, the next state random variable $(X_{n+1}; \lambda_{n+1})$ will take

$(x_{n+1}; \lambda_{n+1})$ with probability $p(x_{n+1}|x_n, u_n)$ at $(n+1)$-st stage, where $\lambda_{n+1} = \lambda_n + r_n(x_n, u_n)$. Thus this is expressed by a coupled dynamics $(\mathrm{i})_\mathrm{n}$, $(\mathrm{i})'_\mathrm{n}$ $0 \leq n \leq N-1$.

Now we consider the problem of maximizing the *terminal* threshold probability on the expanded Markov chain :

$$\mathrm{P}_0(x_0, 0) \qquad \begin{array}{ll} \text{Maximize} & P^{\gamma}_{x_0,0}(\lambda_N + k(X_N) \geq \underline{c}) \\ \text{subject to} & (\mathrm{i})_\mathrm{n} \quad X_{n+1} \sim p(\cdot\,|x_n, u_n) \\ & (\mathrm{i})'_\mathrm{n} \quad \lambda_{n+1} = \lambda_n + r_n(x_n, u_n),\ \lambda_0 = 0 \qquad n = 0, \ldots, N-1 \\ & (\mathrm{ii})_\mathrm{n} \quad u_n \in U \end{array}$$

We take Markov decision functions on the expanded state spaces

$$\gamma_n : X \times \Lambda_n \to U \quad 0 \leq n \leq N-1.$$

The sequence of Markov decision functions $\gamma = \{\gamma_0, \ldots, \gamma_{N-1}\}$ is called an *expanded Markov policy* based upon *cumulative rewards*. We denote the set of all expanded Markov policies by $\widetilde{\Pi}$. For any expanded Markov policy $\gamma(\in \widetilde{\Pi})$, threshold probability in $\mathrm{P}_0(x_0, 0)$ is expressed as the partial multiple sum :

$$\begin{aligned} & P^{\gamma}_{x_0,0}(\lambda_N + k(X_N) \geq \underline{c}) \\ &= \sum_{(x_1,x_2,\ldots,x_N)\,;\,\lambda_N + k(x_N) \geq c} \sum \cdots \sum p(x_1|x_0, u_0)p(x_2|x_1, u_1) \cdots p(x_N|x_{N-1}, u_{N-1}) \end{aligned} \tag{7}$$

where the sequence of decisions in (7) is determined through γ :

$$u_0 = \gamma_0(x_0; \lambda_0),\ u_1 = \gamma_1(x_1; \lambda_1),\ \ldots,\ u_{N-1} = \gamma_{N-1}(x_{N-1}; \lambda_{N-1}).$$

Now we consider subprocess starting at n-th stage and terminated at N-th, which is governed by an expanded Markov policy $\gamma = \{\gamma_n, \ldots, \gamma_{N-1}\}$. $\widetilde{\Pi}(n)$ denotes the set of all such policies. In particular, $\widetilde{\Pi}(0) = \widetilde{\Pi}$. We note that the conditional terminal probability of the event $\lambda_N + k(X_N) \geq \underline{c}$ under the constraint $X_N = x_N$ becomes

$$P(\lambda_N + k \geq \underline{c}\,|\,X_N = x_N) = \begin{cases} 1 & \lambda_N + k(x_N) \geq \underline{c} \\ 0 & \text{otherwise} \end{cases} \qquad x_N \in X. \tag{8}$$

Lemma 3.2 *We have for any* $0 \leq n \leq N-1$, $(x_n; \lambda_n) \in X \times \Lambda_n$ *and* $\gamma = \{\gamma_n, \ldots, \gamma_{N-1}\} \in \widetilde{\Pi}(n)$

$$P^{\gamma}_{x_n,\lambda_n}(\lambda_N + k(X_N) \geq \underline{c}) \;=\; \sum_{x_{n+1} \in X} P^{\gamma'}_{x_{n+1},\lambda_{n+1}}(\lambda_N + k(X_N) \geq \underline{c})p(x_{n+1}|x_n, u_n)$$

where

$$\lambda_{n+1} = \lambda_n + r_n, \quad r_n = r_n(x_n, u_n), \quad u_n = \gamma_n(x_n; \lambda_n), \quad \gamma' = \{\gamma_{n+1}, \ldots, \gamma_{N-1}\},$$

and $P^{\gamma'}_{x_N,\lambda_N} := P$ *in (8) for* $\gamma = \{\gamma_{N-1}\}$.

Proof　It suffices to note that

$$\sum\sum\cdots\sum_{(x_{n+1},x_{n+2},\cdots,x_N)\,;\,\lambda_N+k(x_N)\geq \underline{c}} p(x_{n+1}|x_n,u_n)p(x_{n+2}|x_{n+1},u_{n+1})\cdots p(x_N|x_{N-1},u_{N-1})$$

$$=\sum_{x_{n+1}\in X}\left[\sum\sum\cdots\sum_{(x_{n+2},\cdots,x_N)\,;\,\lambda_N+k(x_N)\geq \underline{c}} p(x_{n+2}|x_{n+1},u_{n+1})\cdots p(x_N|x_{N-1},u_{N-1})\right]p(x_{n+1}|x_n,u_n)$$

holds, where

$$u_m=\gamma_m(x_m;\lambda_m),\ \lambda_{m+1}=\lambda_m+r_m(x_m,u_m)\quad n\leq m\leq N\text{–}1.$$

Now we imbed the expanded Markov problem $\mathrm{P}_0(x_0;0)$ into the family of subproblems $\{\mathrm{P}_n(x_n;\lambda_n)\}$, where $\mathrm{P}_n(x_n;\lambda_n)$ denotes the controlled Markov chain starting at state $(x_n;\lambda_n)$ from n-th stage on :

$$\mathrm{P}_n(x_n;\lambda_n)\qquad\begin{array}{lll}\text{Maximize} & P^{\gamma}_{x_n,\lambda_n}(\lambda_N+k(X_N)\geq \underline{c}) & \\ \text{subject to} & (\mathrm{i})_\mathrm{m}\ \ X_{m+1}\sim p(\cdot\,|x_m,u_m) & \\ & (\mathrm{i})'_\mathrm{m}\ \ \lambda_{m+1}=\lambda_m+r_m(x_m,u_m) & m=n,\ldots,N\text{–}1\\ & (\mathrm{ii})_\mathrm{m}\ \ u_m\in U. & \end{array}$$

Here the maximization is taken for all expanded Markov policies $\gamma=\{\gamma_n,\ldots,\gamma_{N-1}\}\in\widetilde{\Pi}(n)$. We note that $\mathrm{P}_n(x_n;\lambda_n)$ has an equivalent nonterminal (additive) form :

$$\begin{array}{lll}\text{Maximize} & P^{\gamma}_{x_n,\lambda_n}(\lambda_n+r_n+\cdots+r_{N-1}+k\geq \underline{c}) & \\ \text{subject to} & (\mathrm{i})_\mathrm{m},\ (\mathrm{ii})_\mathrm{m} & m=n,\ldots,N\text{–}1\end{array}$$

where the additional dynamics $\{(\mathrm{i})'_\mathrm{m}\}$ is converted into the reward accumulation. Let $w^n(x_n;\lambda_n)$ denote the maximum value of $\mathrm{P}_n(x_n;\lambda_n)$, where

$$w^N(x_N;\lambda_N):=P(\lambda_N+k\geq \underline{c}\,|\,X_N=x_N).$$

Then we have the backward recursive equation :

Theorem 3.1

$$w^n(x;\lambda)=\operatorname*{Max}_{u\in U}\sum_{y\in X}w^{n+1}(y;\lambda+r_n(x,u))p(y|x,u)\qquad(9)$$

$$(x;\lambda)\in X\times\Lambda_n,\quad 0\leq n\leq N\text{–}1$$

$$w^N(x;\lambda)=\begin{cases}1 & \textit{if}\ \ \lambda+k(x)\geq \underline{c}\\ 0 & \textit{otherwise}\end{cases}\qquad(x;\lambda)\in X\times\Lambda_N.\qquad(10)$$

Let $\gamma^*_n(x;\lambda)$ denote a maximizer in (9). Then we have an optimal oplicy $\gamma^*=\{\gamma^*_0,\gamma^*_1,\ldots,\gamma^*_{N-1}\}$ in expanded Markov class $\widetilde{\Pi}$. Further γ^* generates a general policy

$\sigma^* = \{\sigma_0^*, \sigma_1^*, \ldots, \sigma_{N-1}^*\}$, where $\sigma_n^*(x_0, x_1, \ldots, x_n)$ is specified as follows :

$$\begin{aligned} & u_0 := \sigma_1^*(x_0; 0), \quad \lambda_1 := 0 + r_0(x_0, u_0) \\ & u_1 := \sigma_1^*(x_1; \lambda_1), \quad \lambda_2 := \lambda_1 + r_1(x_1, u_1) \\ & \qquad \vdots \\ & u_{n-1} := \sigma_{n-1}^*(x_{n-1}; \lambda_{n-1}), \quad \lambda_n := \lambda_{n-1} + r_{n-1}(x_{n-1}, u_{n-1}) \\ & \sigma_n^*(x_0, x_1, \ldots, x_n) := \sigma_n^*(x_n; \lambda_n). \end{aligned} \tag{11}$$

Then we have the following result:

Theorem 3.2 *([15, Theorem 6.1])*
(i) *The policy σ^* is optimal in general class Π_g.*
(ii) *The maximum value of expanded Markov class $\widetilde{\Pi}$ is equal to the maximum value of general class Π_g :*

$$w^0(x_0; 0) = v_0(x_0). \tag{12}$$

4 Dual Approach

In this section, we show that an optimal (general) policy is also obtained through optimization in the other expanded Markov class. We transform the original problem $\mathrm{Q}_0(x_0)$ with the given lower level $\underline{c}$ into an controlled Markov chain on new state space augmented with threshold levels.

First let us introduce the sequence of additional one-dimensional state variables $\{c_n\}_0^N$ called *threshold levels* :

$$\begin{aligned} c_0 &:= \underline{c} \\ c_n &:= \underline{c} - r_0 - r_1 - \cdots - r_{n-1}. \end{aligned}$$

This is equivalent to the sequential dynamics :

$$\text{(i)}_n'' \quad c_{n+1} = c_n - r_n(x_n, u_n) \quad n = 0, \ldots, N\text{–}1, \quad c_0 = \underline{c}.$$

It turns out that under the sequential constraint

$$r_0 + r_1 + \cdots + r_{N-1} + k \geq \underline{c}$$

if and only if

$$k \geq c_N.$$

This yields the same probability

$$P(r_0 + r_1 + \cdots + r_{N-1} + k \geq \underline{c}) \;=\; P(\,k(X_N) \geq c_N).$$

under appropriate conditions. Thus the introduction of threshold levels has reduced the *original* threshold probability in $\mathrm{Q}_0(x_0)$ to the same *terminal* one.

Second we define *threshold level sets* $\{C_n\}$:

$$\begin{aligned} C_0 &\triangleq \{c_0 \,|\, c_0 = \underline{c}\} \quad \text{where } \underline{c} \text{ is the given lower level} \\ C_n &\triangleq \left\{ c_n \;\middle|\; \begin{array}{c} c_n = \underline{c} - r_0(x_0,u_0) - \cdots - r_{n-1}(x_{n-1},u_{n-1}) \\ (x_0,u_0,\ldots,x_{n-1},u_{n-1}) \in X\times U\times\cdots\times X\times U \end{array} \right\} \\ & \qquad\qquad n = 1,\ldots,N. \end{aligned} \tag{13}$$

Thus C_n denotes the set of all possible threshold levels for the future process on stage-interval $[n, N]$. Then we have the forward recursive formula :

Lemma 4.1

$$\begin{aligned} C_0 &= \{\underline{c}\} \\ C_{n+1} &= \{c - r_n(x,u) \,|\, c \in C_n,\ (x,u) \in X\times U\} \quad 0 \le n \le N-1. \end{aligned} \tag{14}$$

Finally we introduce a new controlled Markov chain on the expanded state spaces $\{X\times C_n\}_0^N$. Here the state variables $\{(X_n; c_n)\}$ behave such that the first component $\{X_n\}$ obeys the original Markov transition law p and the second follows the deterministic dynamics $c_{n+1} := c_n - r_n(x_n, u_n)$. When the decision-maker chooses a decision $u_n(\in U)$ on $(x_n; c_n)(\in X\times C_n)$ at n-th stage, the next state random variable $(X_{n+1}; c_{n+1})$ will take $(x_{n+1}; c_{n+1})$ with probability $p(x_{n+1}|x_n, u_n)$ at $(n+1)$-st stage, where $c_{n+1} = c_n - r_n(x_n, u_n)$. Thus this is the coupled dynamics $(\mathrm{i})_\mathrm{n}$, $(\mathrm{i})''_\mathrm{n}$ $0 \le n \le N-1$.

Now we maximize the *terminal* threshold probability on the expanded Markov chain :

$$\mathrm{D}_0(x_0, \underline{c}) \qquad \begin{array}{ll} \text{Maximize} & P^{\tau}_{x_0,\underline{c}}(\, k(X_N) \ge c_N) \\ \text{subject to} & (\mathrm{i})_\mathrm{n}\ \ X_{n+1} \sim p(\,\cdot\,|x_n, u_n) \\ & (\mathrm{i})''_\mathrm{n}\ \ c_{n+1} = c_n - r_n(x_n,u_n),\ c_0 = \underline{c} \qquad n = 0,\ldots,N\text{–}1 \\ & (\mathrm{ii})_\mathrm{n}\ \ u_n \in U \end{array}$$

Let $\tau = \{\tau_0, \ldots, \tau_{N-1}\}$ be a sequence of Markov decision functions

$$\tau_n : X\times C_n \to U \quad 0 \le n \le N-1$$

on the expanded state spaces. Then τ is called an *expanded Markov policy* based upon *threshold levels*. The set of all expanded Markov policies is denoted by $\widehat{\Pi}$. For any $\tau(\in \widehat{\Pi})$, threshold probability is the partial multiple sum :

$$P^{\tau}_{x_0,\underline{c}}(\, k(X_N) \ge c_N\,) = \sum_{(x_1,x_2,\ldots,x_N)\,;\,k(x_N)\ge c_N} \sum \cdots \sum p(x_1|x_0,u_0)p(x_2|x_1,u_1)\cdots p(x_N|x_{N-1},u_{N-1}) \tag{15}$$

where the sequence of decisions in (15) is determined through τ :

$$u_0 = \tau_0(x_0; c_0),\ u_1 = \tau_1(x_1; c_1),\ \ldots,\ u_{N-1} = \tau_{N-1}(x_{N-1}; c_{N-1}).$$

Now we consider the subprocess from n-th stage to N-th governed by an expanded Markov policy $\tau = \{\tau_n, \ldots, \tau_{N-1}\}$. $\widehat{\Pi}(n)$ denotes the set of all such policies. Hence, $\widehat{\Pi}(0) = \widehat{\Pi}$. We note that the terminal probability of $k(X_N) \geq c_N$ under the condition $X_N = x_N$ becomes

$$P(k \geq c_N \,|\, X_N = x_N) = \begin{cases} 1 & k(x_N) \geq c_N \\ 0 & \text{otherwise} \end{cases} \qquad x_N \in X. \tag{16}$$

Lemma 4.2 *We have for any* $0 \leq n \leq N$–1, $(x_n; c_n) \in X \times C_n$ *and* $\tau = \{\tau_n, \ldots, \tau_{N-1}\} \in \widehat{\Pi}(n)$

$$P^{\tau}_{x_n, c_n}(\, k(X_N) \geq c_N) \;=\; \sum_{x_{n+1} \in X} P^{\tau'}_{x_{n+1}, c_{n+1}}(\, k(X_N) \geq c_N) p(x_{n+1} | x_n, u_n)$$

where

$$c_{n+1} = c_n - r_n, \quad r_n = r_n(x_n, u_n), \quad u_n = \tau_n(x_n; c_n), \quad \tau' = \{\tau_{n+1}, \ldots, \tau_{N-1}\},$$

and $P^{\tau'}_{x_N, c_N} := P$ *in (16) for* $\tau = \{\tau_{N-1}\}$.

Proof It suffices to verify the equality

$$\sum\sum \cdots \sum_{(x_{n+1}, x_{n+2}, \ldots, x_N)\,;\, k(x_N) \geq c_N} p(x_{n+1}|x_n, u_n) p(x_{n+2}|x_{n+1}, u_{n+1}) \cdots p(x_N | x_{N-1}, u_{N-1})$$

$$= \sum_{x_{n+1} \in X} \left[\sum\sum \cdots \sum_{(x_{n+2}, \ldots, x_N)\,;\, k(x_N) \geq c_N} p(x_{n+2}|x_{n+1}, u_{n+1}) \cdots p(x_N | x_{N-1}, u_{N-1}) \right] p(x_{n+1}|x_n, u_n),$$

where

$$u_m = \tau_m(x_m; c_m), \;\; c_{m+1} = c_m - r_m(x_m, u_m) \;\; n \leq m \leq N-1.$$

Now let us imbed $\mathrm{D}_0(x_0; c)$ into the family of subproblems $\{\mathrm{D}_n(x_n; c_n)\}$, where $\mathrm{D}_n(x_n; c_n)$ is the controlled Markov chain starting at state $(x_n; c_n)$ from n-th stage on :

$$\mathrm{D}_n(x_n; c_n) \qquad \begin{array}{lll} \text{Maximize} & P^{\tau}_{x_n, c_n}(\, k(X_N) \geq c_N) & \\ \text{subject to} & (\mathrm{i})_\mathrm{m} \;\; X_{m+1} \sim p(\,\cdot\,|x_m, u_m) & \\ & (\mathrm{i})''_\mathrm{m} \;\; c_{m+1} = c_m - r_m(x_m, u_m) & m = n, \ldots, N\text{–}1 \\ & (\mathrm{ii})_\mathrm{m} \;\; u_m \in U. & \end{array}$$

Here the maximization is over $\widehat{\Pi}(n)$. We see that the restoration of $\{(\mathrm{i})''_\mathrm{m}\}$ to reward accumulation yields an equivalent nonterminal (additive) thershold probability form :

$$\begin{array}{lll} \text{Maximize} & P^{\tau}_{x_n, c_n}(\, r_n + \cdots + r_{N-1} + k \geq c_n) & \\ \text{subject to} & (\mathrm{i})_\mathrm{m},\; (\mathrm{ii})_\mathrm{m} & m = n, \ldots, N\text{–}1. \end{array}$$

Let $f_n(x_n; c_n)$ denote the maximum value of $\mathrm{D}_n(x_n; c_n)$, where

$$f_N(x_N; c_N) := P(k \geq c_N \,|\, X_N = x_N).$$

Then we have the backward recursive equation :

Theorem 4.1

$$f_n(x;c) = \underset{u\in U}{\mathrm{Max}} \sum_{y\in X} f_{n+1}(y; c - r_n(x,u))p(y|x,u) \qquad (17)$$
$$(x;c) \in X\times C_n, \quad 0 \le n \le N\text{–}1$$

$$f_N(x;c) = \begin{cases} 1 & \text{if } k(x) \ge \underline{c} \\ 0 & \text{otherwise} \end{cases} \qquad (x;c) \in X\times C_N. \qquad (18)$$

Let $\bar{\tau}_n(x;c)$ denote a maximizer in (17). Then we have an optimal policy $\bar{\tau} = \{\bar{\tau}_0, \bar{\tau}_1, \ldots, \bar{\tau}_{N-1}\}$ in expanded Markov class $\widehat{\Pi}$. Further $\bar{\tau}$ generates a general policy $\bar{\sigma} = \{\bar{\sigma}_0, \bar{\sigma}_1, \ldots, \bar{\sigma}_{N-1}\}$, where $\bar{\sigma}_n(x_0, x_1, \ldots, x_n)$ is defined as follows :

$$\begin{aligned} &u_0 := \bar{\tau}_0(x_0;\underline{c}), \quad c_1 := \underline{c} - r_0(x_0,u_0) \\ &u_1 := \bar{\tau}_1(x_1;c_1), \quad c_2 := c_1 - r_1(x_1,u_1) \\ &\qquad\vdots \\ &u_{n-1} := \bar{\tau}_{n-1}(x_{n-1};c_{n-1}), \quad c_n := c_{n-1} - r_{n-1}(x_{n-1},u_{n-1}) \\ &\bar{\sigma}_n(x_0,x_1,\ldots,x_n) := \bar{\tau}_n(x_n;c_n). \end{aligned} \qquad (19)$$

Then we have the following result:

Theorem 4.2
(i) *The policy $\bar{\sigma}$ is optimal in general class Π_g.*
(ii) *The maximum value of expanded Markov class $\widehat{\Pi}$ is equal to the maximum value of general class Π_g :*

$$f_0(x_0;\underline{c}) = v_0(x_0). \qquad (20)$$

Proof This is inductively shown in the similar fashion as for Theorem 3.2 (see also [15, Theorem 6.1]).

5 Duality and Consistency

Now let us compare the primal approch with the dual. We have a complementary duality between one expanded Markov problem based upon the cumulative rewards and the other upon the threshold levels. Here we remark that, as for dynamic programming problem itself, an *optimal solution* denotes a pair of optimal value functions and optimal policy.

Theorem 5.1 (Complementary duality theorem)
(i) *For any $\lambda_n \in \Lambda_n$ there exits $c_n \in C_n$ with $\underline{c}$-sum property :*

$$\lambda_n + c_n = \underline{c}. \qquad (21)$$

Conversely for any $c_n \in C_n$ there exits $\lambda_n \in \Lambda_n$ with $\underline{c}$-sum property.
(ii) *Further the optimal solution coincides each other :*

$$w^n(x_n;\lambda_n) = f_n(x_n;c_n), \quad \gamma_n^*(x_n;\lambda_n) = \bar{\tau}_n(x_n;c_n) \qquad x_n \in X. \qquad (22)$$

(iii) *The optimal solution coicides each other in the sence of $\underline{c}$-sum property :*

$$w^n(x_n; \underline{c} - c_n) = f_n(x_n; c_n), \quad \gamma_n^*(x_n; \underline{c} - c_n) = \bar{\tau}_n(x_n; c_n) \quad (x_n; c_n) \in X \times C_n, \quad 0 \leq n \leq N. \tag{23}$$

That is

$$f_n(x_n; \underline{c} - \lambda_n) = w^n(x_n; \lambda_n), \quad \bar{\tau}_n(x_n; \underline{c} - \lambda_n) = \gamma_n^*(x_n; \lambda_n) \quad (x_n; \lambda_n) \in X \times \Lambda_n, \quad 0 \leq n \leq N. \tag{24}$$

Theorem 5.2 (Consistency theorem)
The general policy σ^* generated through the optimal policy γ^* for one expanded Markov problem based upon the cumulative rewards coincides with the general policy $\bar{\sigma}$ generated through the optimal policy $\bar{\tau}$ for the other based upon the threshold levels :

$$\sigma^* = \bar{\sigma}. \tag{25}$$

6 Bellman and Zadeh's Model

In this section, we illustrate two approaches on a three-state, two-action and two-stage process with Bellman and Zadeh's data:

$$X = \{s_1, s_2, s_3\} \qquad U = \{a_1, a_2\} \qquad N = 2$$

$$k(s_1) = 0.3 \qquad k(s_2) = 1.0 \qquad k(s_3) = 0.8$$

$$r_1(a_1) = 1.0 \qquad r_1(a_2) = 0.6$$

$$r_0(a_1) = 0.7 \qquad r_0(a_2) = 1.0$$

$u_t = a_1$

$x_t \backslash x_{t+1}$	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

We maximize the threshold probability that the total additive value is greater than or equal to the lower level $\underline{c} = 2.5$:

$$\begin{aligned} &\text{Maximize} && P_{x_0}^{\sigma}(r_0(U_0) + r_1(U_1) + k(X_2) \geq 2.5) \\ &\text{subject to} && \text{(i)}_n \quad X_{n+1} \sim p(\cdot \,|x_n, u_n) \\ &&& \text{(ii)}_n \quad u_n \in \{a_1, a_2\} \end{aligned} \qquad n = 0, 1 \tag{26}$$

This problem, as an expectation problem of generized additive utility, is solved by two dynamic programming methods and a stochastic multi-stage decision tree-table method [13]. However this paper considers in the framework of threshold probability.

6.1 Primal approach

First, we have from (5) (or (6)) the cumulative sets

$$\Lambda_0 = \{0\}, \quad \Lambda_1 = \{0.7, 1.0\}, \quad \Lambda_2 = \{1.3, 1.6, 1.7, 2.0\}$$

while the cumulative rewards are

$$\lambda_0 = 0, \quad \lambda_1 = r_0(u_0), \quad \lambda_2 = r_0(u_0) + r_1(u_1).$$

Then our problem is converted to

$$\begin{array}{lll} \text{Maximize} & P^{\gamma}_{x_0,\lambda_0}(\lambda_2 + k(X_2) \geq 2.5) & (\lambda_0 = 0) \\ \text{subject to} & \text{(i)} \quad X_{n+1} \sim p(\cdot|x_n, u_n) & \\ & \text{(i)}' \quad \lambda_{n+1} = \lambda_n + r_n(u_n) & n = 0, 1 \\ & \text{(ii)} \quad u_n \in \{a_1, a_2\} & \end{array}$$

where $\gamma = \{\gamma_0, \gamma_1\}$.

Second, by solving the recursive equation

$$\begin{aligned} w^2(x_2; \lambda_2) &= 1(\lambda_2 + k(x_2)) \\ w^1(x_1; \lambda_1) &= \operatorname*{Max}_{u_1} \sum_{x_2} w^2(x_2; \lambda_1 + r_1(u_1)) p(x_2|x_1, u_1) \\ w^0(x_0; \lambda_0) &= \operatorname*{Max}_{u_0} \sum_{x_1} w^1(x_1; \lambda_0 + r_0(u_0)) p(x_1|x_0, u_0) \end{aligned}$$

we have the optimal solution (Table 1) :

	$w^2(x_2;\lambda_2)$				$w^1(x_1;\lambda_1)$	$\gamma_1^*(x_1;\lambda_1)$	$w^0(x_0;0)$	$\gamma_0^*(x_0;0)$
$x_n \backslash \lambda_n$	1.3	1.6	1.7	2.0	0.7	1.0	0	
s_1	0	0	0	0	0.2	a_1	0.9 a_2	0.99 a_2
s_2	0	1	1	1	1.0	a_1	1.0 a_1	0.84 a_2
s_3	0	0	1	1	0.2	a_1	0.2 a_1	0.28 a_1

Table 1: Primal optimal solution

Further we see from (11) that the optimal Markov policy $\gamma^* = \{\gamma_0^*, \gamma_1^*\}$ generates the general policy $\sigma^* = \{\sigma_0^*, \sigma_1^*\}$ where

$$\sigma_0^*(s_1) = a_2, \quad \sigma_0^*(s_2) = a_2, \quad \sigma_0^*(s_3) = a_1$$

$$\begin{array}{lll} \sigma_1^*(s_1, s_1) = a_2, & \sigma_1^*(s_1, s_2) = a_1, & \sigma_1^*(s_1, s_3) = a_1 \\ \sigma_1^*(s_2, s_1) = a_2, & \sigma_1^*(s_2, s_2) = a_1, & \sigma_1^*(s_2, s_3) = a_1 \\ \sigma_1^*(s_3, s_1) = a_1, & \sigma_1^*(s_3, s_2) = a_1, & \sigma_1^*(s_3, s_3) = a_1 \end{array}$$

Thus we have obtained an optimal policy σ^* through one invariant imbedding approach in stochastic problem. We note that σ^* is not Markov because of $\sigma_1^*(s_2, s_1) \neq \sigma_1^*(s_3, s_1)$.

6.2 Dual approach

First, we have from (13) (or (14)) threshold level sets

$$C_0 = \{2.5\}, \quad C_1 = \{1.8, 1.5\}, \quad C_2 = \{1.2, 0.9, 0.8, 0.5\}$$

where threshold levels are

$$c_0 = 2.5, \quad c_1 = 2.5 - r_0(u_0), \quad c_2 = 2.5 - r_0(u_0) - r_1(u_1).$$

Then our problem

$$\begin{array}{lll} \text{Maximize} & P^{\tau}_{x_0,c_0}(\,k(X_2) \geq c_2) & (\,c_0 = 2.5\,) \\ \text{subject to} & (\text{i})_{\text{n}} \quad X_{n+1} \sim p(\,\cdot\,|x_n, u_n) & \\ & (\text{i})''_{\text{n}} \quad c_{n+1} = c_n - r_n(u_n) & n = 0, 1 \\ & (\text{ii})_{\text{n}} \quad u_n \in \{a_1, a_2\} & \end{array}$$

is reduced to the recursive equation

$$\begin{aligned} f_2(x_2; c_2) &= \begin{cases} 1 & \text{if} \quad k(x_2) \geq c_2 \\ 0 & \text{otherwise} \end{cases} \\ f_1(x_1; c_1) &= \underset{u_1}{\text{Max}} \sum_{x_2} f_2(x_2; c_1 - r_1(u_1)) p(x_2|x_1, u_1) \\ f_0(x_0; c_0) &= \underset{u_0}{\text{Max}} \sum_{x_1} f_1(x_1; c_0 - r_0(u_0)) p(x_1|x_0, u_0). \end{aligned}$$

Second, the recursive equation yields the optimal solution (Table 2)

	$f_2(x_2; c_2)$				$f_1(x_1; c_1)$	$\bar{\tau}_1(x_1; c_1)$	$f_0(x_0; 2.5)$	$\bar{\tau}_0(x_0; 2.5)$
$x_n \backslash c_n$	1.2	0.9	0.8	0.5	1.8	1.5	2.5	
s_1	0	0	0	0	0.2	a_1	0.9 a_2	0.99 a_2
s_2	0	1	1	1	1.0	a_1	1.0 a_1	0.84 a_2
s_3	0	0	1	1	0.2	a_1	0.2 a_1	0.28 a_1

Table 2: Dual optimal solution

Further we see from (19) that the optimal Markov policy $\bar{\tau} = \{\bar{\tau}_0, \bar{\tau}_1\}$ generates the general policy $\bar{\sigma} = \{\bar{\sigma}_0, \bar{\sigma}_1\}$ where

$$\begin{array}{lll} \bar{\sigma}_0(s_1) = a_2, & \bar{\sigma}_0(s_2) = a_2, & \bar{\sigma}_0(s_3) = a_1 \\ \bar{\sigma}_1(s_1, s_1) = a_2, & \bar{\sigma}_1(s_1, s_2) = a_1, & \bar{\sigma}_1(s_1, s_3) = a_1 \\ \bar{\sigma}_1(s_2, s_1) = a_2, & \bar{\sigma}_1(s_2, s_2) = a_1, & \bar{\sigma}_1(s_2, s_3) = a_1 \\ \bar{\sigma}_1(s_3, s_1) = a_1, & \bar{\sigma}_1(s_3, s_2) = a_1, & \bar{\sigma}_1(s_3, s_3) = a_1 \end{array}$$

Thus we have the optimal policy (nonMarkov!!) σ^* through the other invariant imbedding.

The general policy σ^* through the primal method coincides with the general policy $\bar{\sigma}$ through the dual :

$$\sigma^* = \bar{\sigma}.$$

References

[1] R.E. Bellman, *Dynamic Programming*, (Princeton Univ. Press, NJ, 1957).

[2] R.E. Bellman and E.D. Denman: *Invariant Imbedding, Lecture Notes in Operation Research and Mathematical Systems, Vol.52*, (Springer-Verlag, Berlin, 1971).

[3] R.E. Bellman and L.A. Zadeh: Decision-making in a fuzzy enviro nment, *Management Science*, **17** (1970), B141-B164.

[4] J.F. Baldwin and B.W. Pilsworth: Dynamic programming for fuzzy ystems with fuzzy environment, *Journal of Mathematical Analysis and Applications*, **85** (1982), 1-23.

[5] M. Bouakiz and Y. Kebir: Target-level criterion in Markov decision processes, *Journal of Optimization of Theory and Applications*, **86** (1995), 1-15.

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

[7] T. Fujita and S. Iwamoto: An optimistic decision-making in fuzzy environment, Ed. J.L. Casti, *Proceedings of The Seventh BELLMAN Continuum (International Workshop on Computation, Optimization and Control)*, The Santa Fe Institute, May, 1999; *Applied Mathematics and Computation* **120** (2001), no.1/3, 91-108.

[8] R.A. Howard: *Dynamic Programming and Markov Processes*, (MIT Press, Mass., 1960).

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

[10] S. Iwamoto: Associative dynamic programs, *Journal of Mathematical Analysis and Applications*, **201** (1996), 195-211.

[11] S. Iwamoto: Maximizing threshold probability through invariant imbedding, Eds. H.-F. Wang and U.-P. Wen, *Proceedings of The Eighth BELLMAN Continuum*, Hsinchu, ROC, Dec., 2000, pp.17-22.

[12] S. Iwamoto: Fuzzy decision-making through three dynamic programming approaches, Eds. H.-F. Wang and U.-P. Wen, *Proceedings of The Eighth BELLMAN Continuum*, Hsinchu, ROC, Dec., 2000, pp.23-27.

[13] S. Iwamoto: A class of dual fuzzy dynamic programs, Ed. J.L. Casti, *Proceedings of The Seventh BELLMAN Continuum (International Workshop on Computation, Optimization and Control)*, The Santa Fe Institute, May, 1999; *Applied Mathematics and Computation* **120** (2001), 91-108.

[14] S. Iwamoto, K. Tsurusaki and T. Fujita: On Markov policies for minimax decision processes, *Journal of Mathematical Analysis and Applications*, **253** (2001), 58-78.

[15] S. Iwamoto, T. Ueno and T. Fujita: Controlled Markov chains with utility functions, Eds. H. Zhenting, J.A. Filar and A. Chen, *Proceedings of International Workshop on "Markov Processes and Controlled Markov Chains"*, Changsha, China, Aug., 1999; Kluwer, 2002, pp.135-148.

[16] S. Iwamoto and M. Sniedovich: Sequential decision making in fuzzy environment, *Journal of Mathematical Analysis and Applications*, **222** (1998), 208-224.

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

[18] J. Kacprzyk and A.O. Esogbue: Fuzzy dynamic programming: Main developments and applications, *Fuzzy Sets and Systems*, **81** (1996), 31-45.

[19] J. Kacprzyk and P. Staniewski: A new approach to the control of stochastic systems in a fuzzy environment, *Archiwum Automatyki i Telemechaniki*, **25** (1980), 443-444.

[20] E.S. Lee: *Quasilinearization and Invariant Imbedding*, (Academic Press, New York, 1968).

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

[22] M.R. Scott: *Invariant Imbedding and its Applications to Ordinary Differential Equations : An introduction*, (Addison-Wesly, London, 1973).

[23] M. Sniedovich: *Dynamic Programming*, (Marcel Dekker Inc. New York, 1992).

[24] C. Wu and Y. Lin: Minimizing risk models in Markov decision processed with policies depending on target values, *Journal of Mathematical Analysis and Applitions*, **231** (1999), 47-67.

Seiichi Iwamoto
Professor, Graduate School of Economics
Kyushu University

Takayuki Ueno
Lecturer, Facultyl of Economics
Nagasaki Prefectural University