89-F-3

# ON DISCRETE-TIME CONTROL PROCESSES UNDER UNCERTAINTY: A DECISION THEORETIC APPROACH

by

Yutaka Umezawa
The University of Tokyo

First Draft, April 1980
Revised, February 1989
Third Draft, April 1989

PREPARED UNDER
THE PROJECT ON MANAGEMENT SCIENCE
RESEARCH INSTITUTE FOR THE JAPANESE ECONOMY

# ON DISCRETE-TIME CONTROL PROCESSES UNDER UNCERTAINTY: A DECISION THEORETIC APPROACH*

Yutaka Umezawa
*The University of Tokyo*

## Introduction

The purpose of this paper is to present a general model of discrete-time control processes under uncertainty, to identify, through a decision theoretic analysis of the model, elements of the processes and fundamental structures of the corresponding multi-stage decision problems, and to clarify the effect, in a multiperiod setting, of information collected and used to obtain the optimal control.

Two models are presented with respect to multiperiod control processes under uncertainty. One is a rather specific model, Model I, of typical control processes (Section 1), and the other is a quite abstract one, Model II, of multi-stage decision processes (Section 2). Through the analysis of Model II, a fundamental structure of multiperiod control problems is made clear, and an algorithm for the derivation of optimal control strategies is presented (Section 3). Finally model I is shown to be reducible to Model II (Section 4).

Various types of adaptive control process models, with procedures to derive optimal control strategies for the corresponding multi-stage decision problems, have been presented in many books and articles. Those of Bellman [1961, ch. 16], Fel'dbaum [1965, ch. 6], and Aoki [1967, ch. 4] are typical ones among them. In each of the three cases, the logic underlying the derivation of the optimal control strategies is only that of dynamic programming; *i.e.*, the principle of optimality: "An optimal strategy has the property that whatever the initial state and initial action are, the remaining actions must constitute an optimal strategy with regard to the state resulting from the first action."

The control objective is to take a sequence of actions during the course of a process so as to maximize the expected value of a criterion function. In adaptive control situations, however, any

* The models discussed in this paper are almost the same with those presented in Umezawa [1970] written in Japanese, except that in this paper we could remove the critical restriction that the action space at any time dose not depend on the previous actions.

concepts related to expectation is not well-defined. With respect to which probability distribution should we take the expectation? The prior probability distribution or the posterior? The principle of optimality does not give any answer to this question at all.

Bellman maintained that the dynamic programming recurrence equation technique can be used even in the adaptive case to establish the existence of optimal strategies and various structural characteristics of the solution. But he succeeded in neither of them. Fel'dbaum and Aoki respectively presented specific adaptive control models, the former being rather complex while the latter simple, and showed how the optimal control strategies are derived. The derivation methods they used are so primitive and complicated that the structural characteristics of the optimal control have not been made clear.

In this paper we derive the algorithm to obtain the optimal strategy for the corresponding multi-stage decision problems through a formal decision-theoretic analysis of Model II, which is a more general version of the excellent model of the control problem defined by Miyasawa [1970]. It is easily shown that Model I is a special case of Model II and models of Bellman, Fel'dbaum, and Aoki are also special cases of Model II. Accordingly any findings obtained through the analysis of Model II are applicable also to our Model I and to their models. And our main findings are (1) the essential elements of the adaptive control process are (*a*) uncertain factor which is out of the control (*b*) information concerning the uncertain factor (*c*) action to achieve the immediate goal as well as to learn about unknown aspects of the process, and (*d*) reward function which measures the degree of the goal achievement, (2) state of the system, the controlled object, is included in the uncertain factor, (3) the information consists of not only the whole series of the previous observations of the system but also the whole series of the previous actions, and (4) for the algorithm to derive the optimal control strategy to function, only two conditions are needed: (*a*) a conditional probability on the set of uncertain factors given any information is known for every period and (*b*) a conditional probability on the next-period-observation set, given any information and actiom, is known for every period.

One might be able to say that our algorithm is exactly the adaptive-control version of the dynamic recurrence equation based upon the principle of optimality. However, without making formal analysis, no one can argue that this version is really relevant to the adaptive control case.

## 1. A Model of Discrete-time Managerial Control Processes under Uncertainty: Model I

We consider discrete-time control processes under uncertainty. Suppose there is a system under the control of a controller as shown by the figure. Let $x_t$ be the state of the system at time point $t$, $t=0,1,\cdots,n$. We also call the time interval from $t=k$ to $t=k+1$ period $k$. A control action is taken at each period. Given the previous state $x_{t-1}$ and an implemented action $c_t$ which is an element of the action space $A_t$, $x_t$ is determined by

$$x_t=F_t(x_{t-1}, c_t, \xi_t) \qquad (1)$$

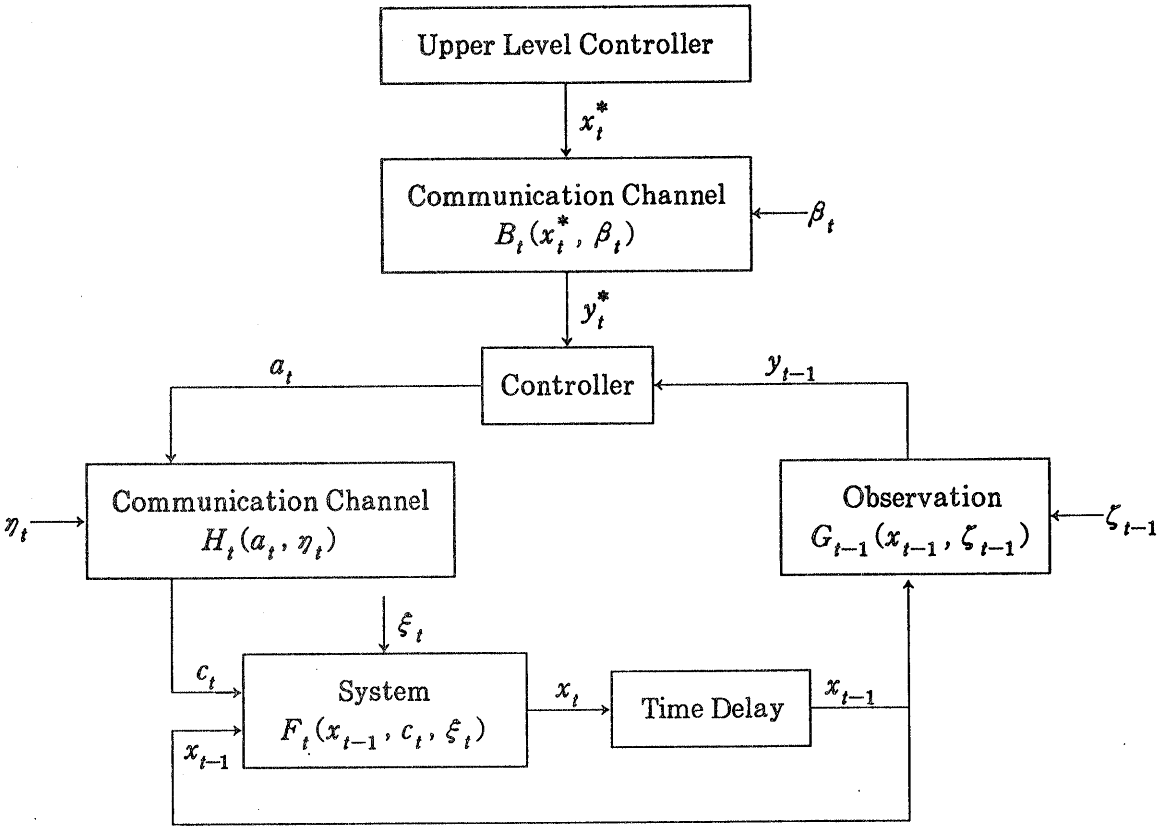

**Fig. 1 Schematic Diagram of the Model I**

where $\xi_t$ is a random disturbance which is out of the controller's control. To decide which action to choose the controller observes $x_{t-1}$ at time $t$. The result of the obserbvation is $y_{t-1}$, given by

$$y_{t-1}=G_{t-1}(x_{t-1}, \zeta_{t-1}) \tag{2}$$

where $\zeta_{t-1}$ is a measurement error. The upper level controller gives the controller a goal $x_t^*$ for the period $t$ which is transformed into $y_t^*$ when accepted by the controller;

$$y_t^*=B_t(x_t^*, \beta_t) \tag{3}$$

where $\beta_t$ is a disturbance through the communication channel between the two hierarchy levels. Obtaining the observation $y_{t-1}$ and goal $y_t^*$, the controller chooses an action $a_t \in A_t$ to control the system, but because of the communication disturbance $\eta_t$, the actually implemented action $c_t$ is determined by

$$c_t=H_t(a_t, \eta_t) \tag{4}$$

The previous actions $\{a_1, a_2, \cdots, a_{t-1}\}$, which is designated by $a^{t-1}$, restrict the action space $A_t$ at the period $t$. To make this explicit we denote the action space by $A_t(a^{t-1})$, It must hold that $c_t \in A_t(a^{t-1})$.

For convenience of notation, as in the above, we introduce a simplification. Let $f_t$ be any function or variable related to time $t$. By $f^k$ we express the series $\{f_1, f_2, \cdots, f_k\}$ or $\{f_0, f_1, \cdots, f_k\}$.

The controller can choose an action $a_t \in A_t(a^{t-1})$ on the basis of the whole series of previous goals, observations, and actions which we call information at $t$ and denote by $I_t$; *i. e.*,

$$I_t=\{y^{*t}, y^{t-1}, a^{t-1}\} \tag{5}$$

A desision rule of the controller at $t$ is a function $d_t$ which, for each possible information $I_t$, specifies an action $a_t \in A_t(a^{t-1})$; *i. e.*,

$$a_t = d_t(I_t) \tag{6}$$

a strategy $d$ is a totality of the decision rules for each period $t$, $t=1,2,\cdots,n$; *i. e.*,

$$d=\{d_1, d_2, \cdots, d_n\}=\{d^n\} \tag{7}$$

We denote by $D$ the set of all possible strategies $d$. Sometimes $n$ is referred to as the time horizon.

As a function of the goal $x_t^*$, the state of the system $x_t$, and the implemented action $c_t$, a reward $v_t$ of period $t$ is given by

$$v_t=u_t(x_t^*, x_t, c_t) \tag{8}$$

The controller wants to maximize the expected value of the sum of the rewards through the time horizon computed with respect to a set of given prior distributions. Reward function $u_t$ is a bounded numerical function for all $t$, $t=1,2,\cdots,n$.

Given a set $X$ and a probability distribution on $X$, we denote either probability of $x$ (discrete case) or density of $x$ (continuous case) by $p(x)$ for any $x \in X$, and refer to $p$ as probability distribution in both cases (Henceforth, instead of probability distribution we simply write p.d.). Also we denote any conditional probability distribution (abbreviated c.p.d.) given $x_1$ by $p(x|x_1)$.

With respect to the goal given by the upper controller $x_t^*$, we assume that c.p.d.'s $p(x_t^*|x^{*t-1})$, $t=2,3,\cdots,n$, and a prior p.d. $p(x_1^*)$ are known.

These constitute the description of Model I. Now we can complete the model formulation as follows:

### Model I

| | | |
|---|---|---|
| Goal | $x_t^*$ | |
| State of the System | $x_t$ | |
| Observation | $o_t=\{y_t^*, y_{t-1}\}$ | (9) |
| Control Action | $a_t \in A_t(a^{t-1})$ | |
| Implemented Action | $c_t=H_t(a_t, \eta_t) \in A_t(a^{t-1})$ | (4) |
| Uncertain Factors | $z_t=\{\xi_t, \eta_t, \zeta_{t-1}, \beta_t\}$ | (10) |
| Information | $I_t=\{y^{*t}, y^{t-1}, a^{t-1}\}=\{o^t, a^{t-1}\}$ | (5) |
| System Equation | $x_t=F_t(x_{t-1}, c_t, \xi_t)$ | (1) |
| Observation Equation | $y_t=G_t(x_t, \zeta_t)$ | (2) |
| | $y_t^*=B_t(x_t^*, \beta_t)$ | (3) |
| Reward Function | $v_t=u_t(x_t^*, x_t, c_t)$ | (8) |
| Time Horizon | $n$ | |
| Decision Rule | $d_t : I_t \to A_t(a^{t-1})$ | (6) |
| Strategy | $d=\{d^n\} \in D$ | (7) |
| Criterion Function | $E[\sum_{t=1}^{n} u_t(x_t^*, x_t, c_t)]$ | |

Concerning probability properties of uncertain factors, Model I is classified into two versions.

### Model I−1. Stochastic Control Model

A. Uncertain factors $\xi_t$, $\eta_t$, $\zeta_{t-1}$ and $\beta_t$ are mutually independently distributed and p.d.'s of each one of them are known for $t=1,2,\cdots,n$.

B. $\xi_t$, $\eta_t$, $\zeta_{t-1}$ and $\beta_t$ are also serially independently distributed.

C. A prior p.d. $p(x_0)$ is known.

### Model I−2. Adaptive Control Model

A. For all $t$, $\xi_t$ has a known c.p.d. $p(\xi_t|\xi^{t-1}, \theta_\xi)$ given $\xi^{t-1}$ and an unknown parameter $\theta_\xi \in \Theta_\xi$. A prior p.d. $p(\theta_\xi)$ over $\Theta_\xi$ is known.

B. Given $\eta^{t-1}$, $\zeta^{t-2}$, $\beta^{t-1}$ and unknown parameters $\theta_\eta \in \Theta_\eta$, $\theta_\zeta \in \Theta_\zeta$, $\theta_\beta \in \Theta_\beta$, uncertain factors $\eta_t$, $\zeta_{t-1}$ and $\beta_t$ are mutually independent and have known c.p.d.'s $p(\eta_t|\eta^{t-1}, \theta_\eta)$, $p(\zeta_{t-1}|\zeta^{t-2}, \theta_\zeta)$, $p(\beta_t|\beta^{t-1}, \theta_\beta)$ respectively. Under the same condition they are also independent of the factor $\xi_t$ given $\xi_{t-1}$ and $\theta_\xi \in \Theta_\xi$. These hold for all $t$, $t=1,2,\cdots,n$. For unknown parameters $\theta_\eta$, $\theta_\zeta$ and $\theta_\beta$, prior p.d.'s $p(\theta_\eta)$, $p(\theta_\zeta)$, and $p(\theta_\beta)$ are known respectively.

C. The initial state variable $x_0$ has a known prior p.d. $p_0(x_0|\theta_x)$ given $\theta_x \in \Theta_x$. Also a prior p.d. $p(\theta_x)$ over $\Theta_x$ is known.

## 2. A Model of Multiperiod Decision Processes under Uncertainty: Model II

We will define another abstract model which has only essential elements, from the view point of multistage decision theory, of the discrete-time control processes under uncertainty. The correspondence between Model I and Model II will be discussed later.

Model II is defined as follows. The elements whose totality constitutes the control processes with a time horizon $n$ are an uncertain factor $\theta_t \in \Theta_t$, an observation $o_t \in O_t$, an action $a_t \in A_t$, and a bounded numerical reward function $r_t$ defined on $\Theta_t \times A_t$, for all $t$, $t=1,2,\cdots,n$. The series of actions taken up to the previous period $a^{t-1}$ restricts the space $A_t$ of possible actions at time $t$. We denote this by $A_t(a^{t-1})$, which is assumed to be a bounded closed set. The choice of an action $a_t$ is based on an information $I_t$ defined by

$$I_t = \{o^t, a^{t-1}\} \tag{11}$$

A decision rule $d_t$ of period $t$ specifies an action $a_t \in A_t(a^{t-1})$ for every possible $I_t$. The set of possible information $I_t$ at time $t$ depends on previous decision rules $d^{t-1}$. We denote this information set by $\boldsymbol{I}_t(d^{t-1})$. Accordingly the decision rule $d_t$ is a function such that $d_t(I_t) \in A_t(a^{t-1})$ for all $I_t \in \boldsymbol{I}_t(d^{t-1})$. We call a set of decision rules $d_t$ for each period $t$, $t=1,2,\cdots,n$, a strategy and denote it by $d$; *i.e.*,

$$d = \{d_1, d_2, \cdots, d_n\} = \{d^n\} \tag{12}$$

Also we denote by $D$ the set of all strategies. A strategy $d^*$ is optimal if it maximizes within $D$ the expected value of the reward sum $\sum_{t=1}^{n} r_t$ with respect to given probability distributions of the uncertain

factor $\theta_t$. A prior p.d. $p(\theta_1)$ on $\Theta_1$ is known and a set of specifications of probability distributions are given with respect to the other uncertain factors $\theta_t$, $t=2,3,\cdots,n$. Further, under the condition that $\theta^{t-1}$ are already given, $\theta_t$ is absolutely independent of the action $a_t$, $t=1,2,\cdots,t$.
Now we can summarize the model as follows:

**Model II**

| | | |
|---|---|---|
| Obervation | $o_t \in O_t$ | |
| Action | $a_t \in A_t(a^{t-1})$ | |
| Uncertain Factor | $\theta_t \in \Theta_t$ | |
| Information | $I_t=\{o^t, a^{t-1}\}\in \boldsymbol{I}_t(d^{t-1})$ | (11) |
| Reward Function | $v_t=r_t(\theta_t, a_t)$ | (13) |
| Time Horizon | $n$ | |
| Decision Rule | $d_t : \boldsymbol{I}_t(d^{t-1}) \to A_t(a^{t-1})$ | (14) |
| Strategy | $d=\{d^n\}$ | (12) |
| Criterion Function | $E[\sum_{t=1}^{n} r_t(\theta_t, a_t)]$ | |

In the next section we will investigate the procedures to derive optimal strategies for the control processes.

## 3. Derivation of Optimal Strategies on Model II

First we will introduce some simplified ways of expressing the expected value of a function. Let $Z_1$ and $Z_2$ be sets of real numbers. Let $M(Z_1 \times Z_2)$ be a set of functions defined on the product space $Z_1 \times Z_2$, $M(Z_1)$ and $M(Z_2)$ be sets of functions defined on $Z_1$ and $Z_2$ respectively, $P(Z_1)$ and $P(Z_2)$ be sets of p.d. functions defined on $Z_1$ and $Z_2$ respectively. Let $p_1$ be a function in $M(Z_1 \times Z_2)$ and also in $P(Z_1)$ for any $z_2 \in Z_2$, $p_2$ be a function in $M(Z_1 \times Z_2)$ and also in $P(Z_2)$ for any $z_1 \in Z_1$. For any $u \in M(Z_1 \times Z_2)$, we define $p_1 u \in M(Z_2)$ and $p_2 u \in M(Z_1)$ by

$$p_1 u(z_2)=\int_{Z_1} u(z_1, z_2)\, p_1(z_1, z_2)\, dz_1 \tag{15}$$

$$p_2 u(z_1)=\int_{Z_2} u(z_1, z_2)\, p_2(z_1, z_2)\, dz_2 \tag{16}$$

Further, let $p_3$ and $p_4$ be in $P(Z_1)$ and in $P(Z_2)$ respectively. We define $p_3 u \in M(Z_2)$ and $p_4 u \in M(Z_1)$ by

$$p_3 u(z_2)=\int_{Z_1} u(z_1, z_2)\, p_3(z_1)\, dz_1 \tag{17}$$

$$p_4 u(z_1)=\int_{Z_2} u(z_1, z_2)\, p_4(z_2)\, dz_2 \tag{18}$$

Similary, for any $v \in M(Z_1)$, we define a value $p_3 v$ by

$$p_3 v=\int_{Z_1} v(z_1)\, p_3(z_1)\, dz_1 \tag{19}$$

We assume the following two conditions.

Fundamental Conditions for Deriving Optimal Strategies

**Condition $C_1$.** For every possible information $I_t$, a conditional probability distribution $\pi_t$ on the set of uncertain factors $\Theta_t$, given $I_t$,

$$\pi_t(\theta_t | I_t),\ \theta_t \in \Theta_t,\ t=1,2,\cdots,n,$$

is known.

**Condition $C_2$.** For every possible combination of information $I_t$ and action $a_t$, a conditional probability distribution $\omega_{t+1}$ on the observation set $O_{t+1}$, given $I_t$ and $a_t$,

$$\omega_{t+1}(o_{t+1} | I_t, a_t),\ o_{t+1} \in O_{t+1},\ t=0,1,\cdots,n-1,$$

is known.

Next we state an algorithm for deriving optimal strategies for the control process with a time horizon $n$.

**Algorithm**

Step 1. For every information $I_{n+1}$, define $\phi_{n+1}$ by

$$\phi_{n+1}(I_{n+1})=0 \tag{20}$$

Step 2. Let $t$ equal $n$.

Step 3. For every possible pair of information $I_t$ and action $a_t$, obtain $\psi_t$ by

(A1) $\psi_t(I_t, a_t)=\pi_t r_t(I_t, a_t)+\omega_{t+1}\phi_{t+1}(I_t, a_t)$,

then for every $I_t$, define $\phi_t$, $a_t^*$, and $d_t^*$ by

(A2) $\phi_t(I_t)=\max\limits_{a_t \in A_t(a^{t-1})} \psi_t(I_t, a_t)=\psi_t(I_t, a_t^*)$

(A3) $d_t^*(I_t)=a_t^*$

Step 4. Let $t$ equal $t-1$. If $t$ equals 1, then go to Step 5. Otherwise go back to Step 3.

Step 5. Define an optimal strategy $d^*$ by

$$d^*=\{d_1^*, d_2^*, \cdots, d_n^*\} \tag{21}$$

then stop.

Since $I_{t+1}=\{o^{t+1}, a^t\}=\{I_t, a_t, o_{t+1}\}$, $\phi_{t+1}(I_{t+1})$ can be regarded as a function of $I_t$, $a_t$, and $o_{t+1}$. Hence we can express $\psi_t(I_t, a_t)$ obtained by Equation (A1) as follows:

$$\begin{aligned}\psi_t(I_t, a_t)=&\int_{\Theta_t} \mathrm{r}_t(\theta_t, a_t)\pi_t(\theta_t | I_t)\, d\theta_t\\ &+\int_{O_{t+1}} \phi_{t+1}(I_t, a_t, o_{t+1})\, \omega_{t+1}(o_{t+1} | I_t, a_t) do_{t+1}\end{aligned} \tag{22}$$

The first term of the right hand side of this equation is the expected value of the reward of period $t$, given $I_t$. The second term is, roughly speaking, the expected value of the sum of the rewards of all the later periods resulting from optimal decision rules for these periods, with the expectation being at the stage when information $I_t$ is gained.

We should note that, for the value $\phi_t(I_t, a_t)$ to be obtained for a pair of $I_t$ and $a_t$, both of the two conditions $C_1$ and $C_2$ must be satisfied.

Now we will show that a strategy $d^*$ obtained through the algorithm is optimal. For that, we need the next two lemmas.

**Lemma 1** *Let $d^t=\{d_1, d_2, \cdots, d_t\}$ be a series of decision rules for period 1 through $t$, $t \leqq n$, $\mathbf{I}(d^{t-1})$ be the set of all information $I_t$ corresponding to the first $t-1$ decision rules $d^{t-1}$ included in $d^t$. Let $q_{t,d^{t-1}}$ denote a probability distribution on $\mathbf{I}_t(d^{t-1})$. Then, for all $t$, $t=1,2,\cdots,n$*

$$q_{t,d^{t-1}}(I^t) = \prod_{k=0}^{t-1} \omega_{k+1}(o_{k+1} | I_k, d_k(I_k)) \tag{23}$$

*where*

$$q_{1,d^0}(I_1) = \omega_1(o_1) \tag{24}$$

*Proof* By the definition of $I_t$ and the multiplication law of the probability

$$\begin{aligned} q_{t,d^{t-1}}(I_t) &= q_{t,d^{t-1}}(I_{t-1}, a_{t-1}, o_t) \\ &= q_{t,d^{t-1}}(I_{t-1}, d_{t-1}(I_{t-1}), o_t) \\ &= q_{t-1,d^{t-2}}(I_{t-1})\ \omega_t(o_t | I_{t-1}, d_{t-1}(I_{t-1})) \end{aligned} \tag{25}$$

since the conditional probability, given $I_{t-1}$, that the action $a_{t-1}=d_{t-1}(I_{t-1})$ is taken is unity. This recursive formula yields (23). Q.E.D.

Examining the right hand sides of (25), we recognize that for each decision rule $d_t$ for period $t$ the function $\omega_{t+1}$ gives the probability of $o_{t+1}$, given every information $I_t \in \mathbf{I}(d^{t-1})$. Therefore we will denote the function by $\omega_{t+1,d_t}$, when it is necessary to express explicitly the function's dependancy on the specific decision rule $d_t$.

By this lemma, especially by a direct use of (25), We can derive a corollary.

**Corollary** *Let $u_t$ be any bounded numerical function of information $I_t$. Then*

$$q_{t,d^{t-1}}\, u_t = q_{t-1,d^{t-2}}\ \omega_{t,d_{t-1}} u_t \tag{26}$$

**Lemma 2** *Let $R_t(d^t)$ be the expected value of the reward $r_t$ of period $t$ which corresponds to a series of decision rules $d^t$, with the expectation being taken at the beginning of the initial period. Then, for $t=1,2,\cdots,n$*

$$R_t(d^t) = q_{t,d^{t-1}}\, \pi_t r_t(d_t)$$

*Proof* Let $p_{d^{t-1}}$ be the joint p.d. of $\theta_t$ and $I_t$ for $d^{t-1}$. Then

$$R_t(d^t) = \int r_t(\theta_t, d_t(I_t))\, p_{d^{t-1}}(\theta_t, I_t)\, d\theta_t\, dI_t$$

Since

$$\begin{aligned} p_{d^{t-1}}(\theta_t, I_t) &= p(\theta_t | I_t) p(I_t) \\ &= \pi_t(\theta_t | I_t) q_{t,d^{t-1}}(I_t), \end{aligned}$$

using the simplified notation for integration we may write

$$R_t(d^t)=\int_{I_t(d^{t-1})}\{\int_{\Theta_t} r_t(\theta_t, d_t(I_t))\pi_t(\theta_t|I_t)d\theta_t\}q_{t,d^{t-1}}(I_t)dI_t$$

$$=\int_{I_t(d^{t-1})}\pi_t r_t(I_t, d_t)q_{t,d^{t-1}}(I_t)dI_t$$

$$=q_{t,d^{t-1}}\pi_t r_t(d_t) \qquad \text{Q.E.D.}$$

Let $R(d)$ be the expected value of the total reward for a strategy $d\in D$, with the expectation being taken at the beginning of period 1. It is clear that $R(d)$ is equal to the sum of $R_t(d^t)$ throughout the whole processes; *i.e.*,

$$R(d)=E[\sum_{t=1}^{n} r_t(\theta_t, d_t(I_t))]$$

$$=\sum_{t=1}^{n} E[r_t(\theta_t, d_t(I_t))]$$

$$=\sum_{t=1}^{n} R_t(d^t) \tag{27}$$

For $k=1,2,\cdots,t$, let $S_k(d)$ be the sum of the expected values of the rewarde of period $k$ through period $n$ corresponding a strategy $d\in D$; *i.e.*,

$$S_k(d)=\sum_{t=k}^{n} R_t(d^t) \tag{28}$$

where $d^t$ is the series of the first $t$ decision rules of the strategy $d$. Then by (27)

$$S_1(d)=\sum_{t=1}^{n} R_t(d^t)=R(d) \tag{29}$$

Therefore $S_1(d)$ is also a criterion function for the optimal strategy.

**Theorem** *Any strategy $d^*$ obtained through the algorithm is optimal*; *i.e.*, *for all strategy $d\in D$*

$$R(d^*)=S_1(d^*)\geqq R(d)=S_1(d) \tag{30}$$

*Proof* Let $d_t^*$ be a function determined by Equation (A3) at step 3. We prove the theorem by showing that the next two equations hold for any strategy $d\in D$ and for all $t$, $t=1,2,\cdots,n$;

$$S_t(d^{t-1}, d_t^*, d_{t+1}^*, \cdots, d_n^*)=q_{t,d^{t-1}}\phi_t \tag{31}$$

$$S_t(d^{t-1}, d_t^*, d_{t+1}^*, \cdots, d_n^*)\geqq S_t(d) \tag{32}$$

We show these by an inductive way of proving.

By (20) and Equation (A1)

$$\phi_n(I_n, d_n(I_n))=\pi_n r_n(I_n, d_n(I_n))$$

Multiplying both sides of this equation by $q_{n,d^{n-1}}(I_n)$, and integrating each with respect to $I_n$, we get

$$q_{n,d^{n-1}}\phi_n(d_n)=q_{n,d^{n-1}}\pi_n r_n(d_n)$$

so that, by Lemma 2

$$S_n(d)=R_n(d)=q_{n,d^{n-1}}\pi_n r_n(d_n)$$

$$=q_{n,d^{n-1}}\phi_n(d_n) \tag{33}$$

On the other hand, from Equation (A2), we get for all $a_n\in A_n(a^{n-1})$ and $I_n\in \mathbf{I}_n(d^{n-1})$

$$\phi_n(I_n)=\phi_n(I_n, a_n^*)\geqq \phi_n(I_n, a_n)$$

which, by Equation (A3), is equivalent to

$$\phi_n(I_n)=\phi_n(I_n, d_n^*(I_n))\geqq \phi_n(I_n, d_n(I_n)).$$

Again, multiplying by $q_{n,d^{n-1}}(I_n)$ and integrating with respet to $I_n$ yield

$$q_{n,d^{n-1}}\phi_n = q_{n,d^{n-1}}\varphi_n(d_n^*) \geqq q_{n,d^{n-1}}\varphi_n(d_n) \qquad (34)$$

Therefore by (33) and (34)

$$S_n(d^{n-1}, d_n^*) = q_{n,d^{n-1}}\varphi_n(d_n^*) = q_{n,d^{n-1}}\phi_n$$

and

$$S_n(d^{n-1}, d_n^*) = q_{n,d^{n-1}}\varphi_n(d_n^*) \geqq q_{n,d^{n-1}}\varphi_n(d_n) = S_n(d)$$

These prove that (31) and (32) hold for $t=n$.

Next we shall assume that (31) and (32) hold for $t=k$. By (28), for all $d \in D$

$$S_{k-1}(d) = R_{k-1}(d^{k-1}) + S_k(d)$$

so that

$$\begin{aligned} S_{k-1}(d^{k-1}, d_k^*, d_{k+1}^*, \cdots, d_n^*) &= R_{k-1}(d^{k-1}) + S_k(d^{k-1}, d_k^*, d_{k+1}^*, \cdots, d_n^*) \\ &= q_{k-1,d^{k-2}}\pi_{k-1}r_{k-1}(d_{k-1}) + q_{k,d^{k-1}}\phi_k \\ &= q_{k-1,d^{k-2}}\pi_{k-1}r_{k-1}(d_{k-1}) + q_{k-1,d^{k-2}}\omega_{k,d_{k-1}}\phi_k \\ &= q_{k-1,d^{k-2}}(\pi_{k-1}r_{k-1}(d_{k-1}) + \omega_{k,d_{k-1}}\phi_k) \end{aligned}$$

by Lemma 2, the induction hypothesis (31) with $t=k$, and (26). Since Equation (A1) can also be written as

$$\varphi_t(I_t, d_t(I_t)) = \pi_t r_t(I_t, d_t(I_t)) + \omega_{t+1}\phi_{t+1}(I_t, d_t, (I_t)),$$

we obatain

$$S_{k-1}(d^{k-1}, d_k^*, \cdots, d_n^*) = q_{k-1,d^{k-2}}\varphi_{k-1}(d_{k-1}) \qquad (35)$$

On the other hand, from Equation (A2), we get for all $a_{k-1} \in A_{k-1}(a^{k-2})$ and $I_{k-1} \in \boldsymbol{I}_{k-1}(d^{k-2})$

$$\phi_{k-1}(I_{k-1}) = \varphi_{k-1}(I_{k-1}, a_{k-1}^*) \geqq \varphi_{k-1}(I_{k-1}, a_{k-1})$$

which, by Equation (A3), is equivalent to

$$\phi_{k-1}(I_{k-1}) = \varphi_{k-1}(I_{k-1}, d_{k-1}^*(I_{k-1})) \geqq \varphi_{k-1}(I_{k-1}, d_{k-1}(I_{k-1}))$$

Multiplied by $q_{k-1,d^{k-2}}(I_{k-1})$ and integrated with respect to $I_{k-1}$, this yields

$$q_{k-1,d^{k-2}}\phi_{k-1} = q_{k-1,d^{k-2}}\varphi_{k-1}(d_{k-1}^*) \geqq q_{k-1,d^{k-2}}\varphi_{k-1}(d_{k-1}) \qquad (36)$$

Thus we have, by (35), (36), and the induction hypothesis (32) with $t=k$,

$$S_{k-1}(d^{k-2}, d_{k-1}^*, \cdots, d_n^*) = q_{k-1,d^{k-2}}\varphi_{k-1}(d_{k-1}^*) = q_{k-1,d^{k-2}}\phi_{k-1}$$

and

$$\begin{aligned} S_{k-1}(d^{k-2}, d_{k-1}^*, \cdots, d_n^*) &\geqq q_{k-1,d^{k-2}}\varphi_{k-1}(d_{k-1}) \\ &= S_{k-1}(d^{k-1}, d_k^*, \cdots, d_n^*) \\ &= R_{k-1}(d^{k-1}) + S_k(d^{k-1}, d_k^*, \cdots, d_n^*) \\ &\geqq R_{k-1}(d^{k-1}) + S_k(d) \\ &= S_{k-1}(d) \end{aligned}$$

These prove that (31) and (32) hold for $t=k-1$.

Consequently, (32) with $t=1$ shows that for any $d \in D$

$$S_1(d^*) \geqq S_1(d)$$

and, by (31) with $t=1$ and (24)

$$R(d^*) = \max_{d \in D} S_1(d) = q_{1,d^0}\phi_1 = \omega_1\phi_1 \qquad (37)$$

Q.E.D.

The analysis of the case where the action space at time $t$, $A_t$, does not depend on the previous action $a^{t-1}$ was made by Miyasawa [1970] and Umezawa [1970].

## 4. Relations between the Two Models

By defining each of elements of Model II in terms of those of Model I, we can relate Model I to Model II as follows:

| Model II | Model I | |
|---|---|---|
| $a_t$ | $=a_t$ | (38) |
| $\theta_t$ | $=\{x_t^*, \beta_t, x_{t-1}, \zeta_{t-1}, \eta_t, \xi_t\}$ | (39) |
| $o_t$ | $=\{y_t^*, y_{t-1}\}$ | (40) |
| $r_t(\theta_t, a_t)$ | $=u_t(x_t^*, F_t(x_{t-1}, H_t(a_t, \eta_t), \xi_t), H_t(a_t, \eta_t))$ | (41) |

The uncertain elements which appear in the right-hand side of (41) are $x_t^*$, $x_{t-1}$, $\eta_t$, and $\xi_t$. These elements are all included in the set $\theta_t$. Accordingly we can obtain the expected value of the reward $u_t$, given $I_t$, if the p.d. $p(\theta_t|I_t)$ is known, as condition $C_1$ requires.

Since conditions $C_1$ and $C_2$ are necessary for the algorithm to function (see step 3), we now examine whether they are satisfied in Model I, using relations (38) through (40).

Since by (40)

$$I_t=\{o^t, a^{t-1}\}=\{y^{*t}, y^{t-1}, a^{t-1}\} \tag{42}$$

we have

$$\begin{aligned}\pi_t(\theta_t|I_t)&=p(x_t^*, \beta_t, x_{t-1}, \zeta_{t-1}, \eta_t, \xi_t|y^{*t}, y^{t-1}, a^{t-1})\\&=\int\cdots\int p(x^{*t}, \beta^t, x^{t-1}, \zeta^{t-1}, \eta^t, \xi^t, \theta_\beta, \theta_\zeta, \theta_\eta, \theta_\xi|y^{*t}, y^{t-1}, a^{t-1})\\&\quad\times d(x^{*t-1}, \beta^{t-1}, x^{t-2}, \zeta^{t-2}, \eta^{t-1}, \xi^{t-1}, \theta_\beta, \theta_\zeta, \theta_\eta, \theta_\xi)\\&=\int\cdots\int p(\Omega(t)|y^{*t}, y^{t-1}, a^{t-1})d\Omega(t-1)\end{aligned} \tag{43}$$

where

$$\Omega(t)=\{x^{*t}, \beta^t, x^{t-1}, \zeta^{t-1}, \eta^t, \xi^t, \theta_\beta, \theta_\zeta, \theta_\eta, \theta_\xi\} \tag{44}$$

Therefore we can obtain $\pi_t(\theta_t|I_t)$ if $p(\Omega(t)|y^{*t}, y^{t-1}, a^{t-1})$ is known. We will show inductively that this c.p.d. of $\Omega(t)$, given $\{y^{*t}, y^{t-1}, a^{t-1}\}$ is known for all $t$, $t=1,2,\cdots,n$.

Using the multiplication law of the probability repetitively, we have

$$\begin{aligned}&p(\Omega(t+1), y_{t+1}^*, y_t|y^{*t}, y^{t-1}, a^t)\\&\quad=p(\Omega(t), x_{t+1}^*, \beta_{t+1}, x_t, \zeta_t, \eta_{t+1}, \xi_{t+1}, y_{t+1}^*, y_t|m_{t+1})\\&\quad=p(\Omega(t)|m_{t+1})p(x_{t+1}^*|m_{t+1}, \Omega(t))p(\beta_{t+1}|m_{t+1}, \Omega(t), x_{t+1}^*)\\&\qquad\times p(x_t|m_{t+1}, \Omega(t), x_{t+1}^*, \beta_{t+1})p(\zeta_t|m_{t+1}, \Omega(t), x_{t+1}^*, \beta_{t+1}, x_t)\\&\qquad\times p(\eta_{t+1}|m_{t+1}, \Omega(t), x_{t+1}^*, \beta_{t+1}, x_t, \zeta_t)\\&\qquad\times p(\xi_{t+1}|m_{t+1}, \Omega(t), x_{t+1}^*, \beta_{t+1}, x_t, \zeta_t, \eta_{t+1})\\&\qquad\times p(y_{t+1}^*|m_{t+1}, \Omega(t), x_{t+1}^*, \beta_{t+1}, x_t, \zeta_t, \eta_{t+1}, \xi_{t+1})\end{aligned}$$

$$\times p(y_t \,|\, m_{t+1}, \Omega(t), x^*_{t+1}, \beta_{t+1}, x_t, \zeta_t, \eta_{t+1}, \xi_{t+1}, y^*_{t+1}) \tag{45}$$

where

$$m_{t+1} = \{y^{*t}, y^{t-1}, a^t\} \tag{46}$$

According to the specifications of Model Ⅰ, (45) can be written simply

$$\begin{aligned} &p(\Omega(t+1), y^*_{t+1}, y_t \,|\, y^{*t}, y^{t-1}, a^t) \\ &\quad = p(\Omega(t) \,|\, y^{*t}, y^{t-1}, a^{t-1}) p(x^*_{t+1} \,|\, x^{*t}) p(\beta_{t+1} \,|\, \beta^t, \theta_\beta) \\ &\qquad \times \delta(x_t - F_t(x_{t-1}, H_t(a_t, \eta_t), \xi_t)) p(\zeta_t \,|\, \zeta^{t-1}, \theta_\zeta) \\ &\qquad \times p(\eta_{t+1} \,|\, \eta^t, \theta_\eta) p(\xi_{t+1} \,|\, \xi^t, \theta_\xi) \delta(y^*_{t+1} - B_{t+1}(x^*_{t+1}, \beta_{t+1})) \\ &\qquad \times \delta(y_t - G_t(x_t, \zeta_t)) \end{aligned} \tag{47}$$

Now we assume the p.d. $p(\Omega(t) \,|\, y^{*t}, y^{t-1}, a^{t-1})$ be known. Then, since the other factors in the right-hand side of (47) are all known, the left-hand side becomes known. By integrating this with respect to $\Omega(t+1)$ we have

$$\begin{aligned} &p(y^*_{t+1}, y_t \,|\, y^{*t}, y^{t-1}, a^t) \\ &\quad = \int \cdots \int p(\Omega(t+1), y^*_{t+1}, y_t \,|\, y^{*t}, y^{t-1}, a^t) d\Omega(t+1) \end{aligned} \tag{48}$$

Also by the multiplication law, we have

$$\begin{aligned} &p(\Omega(t+1) \,|\, y^{*t+1}, y^t, a^t) \\ &\quad = p(\Omega(t+1), y^*_{t+1}, y_t \,|\, y^{*t}, y^{t-1}, a^t) / \{p(y^*_{t+1}, y_t \,|\, y^{*t}, y^{t-1}, a^t)\} \end{aligned} \tag{49}$$

The numerator and the denominator in the right-hand side of this equation are given by (47) and (48) respectively, so that $p(\Omega(t+1) \,|\, y^{*t+1}, y^t, a^t)$ is also known.

The only thing left is to show that $p(\Omega(t) \,|\, y^{*t}, y^{t-1}, a^{t-1})$ is known for $t=1$; *i.e.*, $p(\Omega(1) \,|\, y^*_1, y_0)$ is known. Similarly with (49)

$$p(\Omega(1) \,|\, y^*_1, y_0) = \{p(\Omega(1), y^*_1, y_0)\} / \{\int \cdots \int p(\Omega(1), y^*_1, y_0) d\Omega(1)\} \tag{50}$$

By the specifications of the model, we have

$$\begin{aligned} p(\Omega(1), y^*_1, y_0) &= p(x^*_1, \beta_1, x_0, \zeta_0, \eta_1, \xi_1, \theta_\beta, \theta_\zeta, \theta_\eta, \theta_\xi, y^*_1, y_0) \\ &= p(x^*_1) p(\theta_\beta) p(\beta_1 \,|\, \theta_\beta) p(x_0) p(\theta_\zeta) p(\zeta_0 \,|\, \theta_\zeta) \\ &\quad \times p(\theta_\eta) p(\eta_1 \,|\, \theta_\eta) p(\theta_\xi) p(\xi_1 \,|\, \theta_\xi) \\ &\quad \times \delta(y^*_1 - B(x^*_1, \beta_1)) \delta(y_0 - G_0(x_0, \zeta_0)) \end{aligned} \tag{51}$$

and

$$p(x_0) = \int p_0(x_0 \,|\, \theta_x) p(\theta_x) d\theta_x \tag{52}$$

Since the factors in the right-hand side of (51) are all known, $p(\Omega(1) \,|\, y^*_1, y_0)$ is also known. Thus, by (50), $p(\Omega(t) \,|\, y^{*t}, y^{t-1}, a^{t-1})$ with $t=1$ is also known.

On the other hand. We have by (40)

$$\omega_{t+1}(o_{t+1} \,|\, I_t, a_t) = p(y^*_{t+1}, y_t \,|\, y^{*t}, y^{t-1}, a^t)$$

Since this is given by (48), Condition $C_2$ is also satisfied in Model Ⅰ.

Thus, it has become clear that both of conditions $C_1$ and $C_2$ are satisfied in Model Ⅰ, so that an optimal strategy can be derived through the algorithm.

## References

Aoki, M., *Optimization of Stochastic Systems*, Academic Press, 1967.

Bellman, R., *Adaptive Control Processes: A Guided Tour*, Princeton Univ. Press, 1961.

Fel'dbaum, A.A., *Optimal Control Systems*, Translated by A. Kraiman, Academic Press, 1965.

Miyasawa, K., "A General Theory of Control Problems," (in Japanese), *Keizaigaku-Ronshu* (Univ. of Tokyo), Vol.36, No.1 (1970), pp.2-16.

Umezawa, Y., "Planning and Control Processes under Uncertainty," (in Japanese), *Keizaigaku-Ronshu* (Univ. of Tokyo), Vol.36, No.2 (1970), pp.12-44.