FUNDAMENTAL THEOREMS IN A BAYES CONTROLLED PROCESS

Furukawa, Nagata
Department of Mathematics, Kyushu University

https://doi.org/10.5109/13048

出版情報：統計数理研究. 14 (1/2), pp.103-110, 1970-03. Research Association of Statistical Sciences
バージョン：
権利関係：

FUNDAMENTAL THEOREMS IN A BAYES CONTROLLED PROCESS

By

Nagata FURUKAWA*

(Received February 5, 1970)

§ 1. Introduction.

The purpose of this paper is to discuss some results concerning basic parts of a Bayes controlled process with a finite horizon.

A Bayes controlled process which we shall treat is as follows:

We assume that the probability distributions of random variables $X_1, X_2, \cdots, X_k$ are of known type, but include unknown parameters, respectively, the prior probability distributions of which are given. Then we are to maximize the total expected reward in n trials at each of which we are free to choose among $X_1, X_2, \cdots, X_k$.

Our main results are the following: There exists an optimal policy among the class of policies depending outcomes only through the posterior probability distributions of unknown parameters (Theorem 1). Let L_i denote the reward when X_i is chosen, where the value of L_i depends on the true parameter value but not the observations x_i. Let $\rho^*_{N+1}(\xi)$ for a given prior distribution ξ be the total expected reward incurred from an optimal policy truncated at $N+1$, and let $T^{X_i}\xi$ be the posterior probability distribution given observation x_i. Then $\rho^*_{N+1}(\xi)=\underset{1\leqq i\leqq k}{\mathrm{Max}}\,[E(L_1)+E(\rho^*_N(T^{X_i}\xi))]$ (Theorem 2).

The statement of Theorem 1 corresponds to the well-known fact in the Bayes sequential analysis [3], where main interest lies in the stopping rule and the terminal decision, whereas in our case we choose either variables at each of n stages.

The statement of Theorem 2 is the recurrence formula known as the dynamic programming formulation. Bradt, Johnson and Karlin [4], and Feldman [6] give some results concerning an optimal policy in the truncated "two-armed bandit" problem which is a special type of our controlled process. Their arguments are based on our recurrence formula without any proof of it.

§ 2. A Bayes controlled process with a finite horizon.

The elements of a Bayes controlled process with a finite horizon are the following.

(i) A sample space $\mathfrak{X}=(z, \Omega, p)$. Let $Z=(K, U_1)\times(K, U_2)\times\cdots\times(K, U_N)$ be a state space, where K is a set $\{1, 2, \cdots, k\}$ and each U_i is an abstract space. We call $(K, U_1)\times(K, U_2)\times\cdots\times(K, U_n)$ a n-state space. Hence a N-state space is the state space defined above. The outcome $z\in Z$ is written in various forms for convenience:

* Department of Mathematics, Kyushu University, Fukuoka.

$$z=((j_1, u_1), (j_2, u_2), \cdots, (j_N, u_N))$$
$$=(z_{1j_1}, z_{2j_2}, \cdots, z_{Nj_N})$$
$$=(z_1, z_2, \cdots, z_N)$$

where $j_h=1, 2, \cdots, k$ for $h=1, 2, \cdots, N$. Let $\Theta=(\theta_1, \theta_2, \cdots, \theta_k)$, let $P_\theta(z)$ be the probability density of z with respect to a measure λ on z for a fixed parameter Θ, and let $\Omega=\{\Theta\}$ the parameter space.

(ii) The apriori distribution of the parameter ξ. Let $\xi(\Theta)$ be the apriori probability density of Θ with respect to a measure φ on Ω.

(iii) A policy $\mathfrak{S}_N$. For each n let D_{nj} $(j=1, 2, \cdots, k)$ be subsets of a n-state space such that $D_{ni}\cap D_{nj}=\Phi$ for $i\neq j$ and $\sum_{j=1}^{k} D_{nj}$ coincide with the n-state space, i.e., $\{D_{nj};\ 1\leqq j\leqq k\}$ is a partition of a n-state space. Let one of S_{0j} $(1\leqq j\leqq k)$ be the state space Z, and let all the rest empty sets. And let S_{nj} be a cylinder set over D_{nj} for $n=1, 2, \cdots, N$ and $j=1, 2, \cdots, k$, i.e., $S_{nj}=\{(z_1, z_2, \cdots, z_N)\,|\,(z_1, z_2, \cdots, z_n)\in D_{nj}\}$. Here $\{D_{nj};\ 1\leqq n\leqq N,\ 1\leqq j\leqq k\}$ has the meaning of a policy such that according as the outcome $(z_1, z_2, \cdots, z_n)$ up to the n-th stage belongs to D_{nj} we choose X_j at the $(n+1)$-th stage. Since there is one to one correspondence between $\{D_{nj}\}$ and $\{S_{nj}\}$, a policy $\mathfrak{S}_N=\{S_{nj};\ 0\leqq n\leqq N,\ 1\leqq j\leqq k\}$ is a class of cylinder sets over partitions of n-state spaces for $n=0, 1, \cdots, N$.

(iv) A total expected reward ρ_N. Let $L_j(\Theta)=L_j(\theta_j)$ represent the reward when X_j is chosen and θ_j is the true parameter value of the distribution of X_j. We assume that all L_j are ξ-integrable real-valued function on Ω. Let $T_{i_0i_1\cdots i_r}=S_{0i_0}\cap S_{1i_1}\cap\cdots\cap S_{ri_r}$. Then, for given ξ, the total expected reward for a policy S_{N+1} truncated at $N+1$ is given by

$$\rho_{N+1}(\xi, \mathfrak{S}_N)=\sum_{r=0}^{N}\sum_{i_0=1}^{k}\sum_{i=1}^{k}\cdots\sum_{i_r=1}^{k}\int_{T_{i_0i_1\cdots i_r}}\int_{\Omega} L_{i_r}(\theta_{i_r})\xi(\Theta)p_\theta(z)d\varphi(\Theta)d\lambda(z)\,. \tag{2.1}$$

The triple $(\Omega, \mathfrak{S}_N, \rho_N)$ is an N-stages Bayes controlled process.

For the sake of simplicity, in the subsequent sections we shall develope arguments and give the proofs of a lemma and theorems in the case $k=2$ only. Since, in the case $k>2$, we can proceed arguments in the same way as in the case $k=2$ with only formal complication, it suffices to prove the results in the case $k=2$.

§ 3. Bayes solutions.

In this section we shall prove a lemma useful for main results. This lemma implies there always exists an optimal policy in the truncated Bayes controlled process.

Let $X_1=X$, $X_2=Y$, let $z_{n1}=(1, u_n)=x_n$, $z_{n2}=(2, u_n)y_n$, and let $\theta_1=\omega$, $\theta_2=\theta$.

Let $F_j(z)$ denote the set $\{(t_1, t_2, \cdots, t_N)\,|\,t_1=z_1, t_2=z_2, \cdots, t_j=z_j\}$, and for (ξ, p)-integrable function h on $\Omega\times Z$ let $E_j(h)$ be the conditional expectation of h given $z_1, z_2, \cdots, z_j$, when (ω, θ) has a distribution ξ, and z has distribution $p_{(\omega,\theta)}$ for fixed (ω, θ).

Then

$$E_j(h)=\frac{\int_{F_{j(z)}}\int_{\Omega}\xi(\omega,\theta)p_{\omega,\theta}(t)h(\omega,\theta,t)d\varphi d\lambda}{\int_{F_{j(z)}}\int_{\Omega}\xi(\omega,\theta)p_{\omega,\theta}(t)d\varphi d\lambda}\,. \tag{3.1}$$

Let $P_{\xi}(z)=\int_{\Omega}\xi(\omega,\theta)p_{\omega,\theta}(z)d\varphi(\omega,\theta)$, then we have along the same line as in [3]

$$\int_{T_{i_0i_1\cdots i_r}}\int_{\Omega}L_{i_r}(\omega,\theta)\xi(\omega,\theta)p_{\omega,\theta}(z)d\varphi d\lambda=\int_{T_{i_0i_1\cdots i_r}}E_r(L_{i_r}(\omega,\theta))P_{\xi}(z)d\lambda\,. \tag{3.2}$$

Applying this result to (2.1) we obtain

$$\rho_{N+1}(\xi,\mathfrak{S}_N)=\sum_{r=0}^{N}\sum_{i_0=1}^{2}\cdots\sum_{i_r=1}^{2}\int_{T_{i_0i_1\cdots i_r}}E_r(L_{i_r}(\omega,\theta))P_{\xi}(z)d\lambda(z)\,. \tag{3.3}$$

We define

$$\begin{aligned}
&a_N(z_1,z_2,\cdots,z_N)=\mathrm{Max}\,[E_N(L_1(\omega,\theta)),\,E_N(L_2(\omega,\theta))]\,,\\
&a_j(z_1,z_2,\cdots,z_j)=\mathrm{Max}\,[E_j(L_1(\omega,\theta))+E_j[a_{j+1}(z_1,\cdots,z_j,x_{j+1})],\\
&\qquad\qquad E_j(L_2(\omega,\theta))+E_j[a_{j+1}(z_1,\cdots,z_j,y_{j+1})]\,,\\
&\qquad\qquad (0\leqq j\leqq N-1)\,,
\end{aligned} \tag{3.4}$$

by induction backward on j. Hence a_0 is a constant. And let

$$\begin{aligned}
&S^*_{N1}=\{z\,|\,E_N(L_1)>E_N(L_2)\}\,,\\
&S^*_{N2}=\{z\,|\,E_N(L_1)\leqq E_N(L_2)\}\,,\\
&S^*_{j1}=\{z\,|\,E_j(L_1)+E_j[a_{j+1}(z_1,\cdots,z_j,x_{j+1})]>E_j(L_2)+E_j[a_{j+1}(z_1,\cdots,z_j,y_{j+1})]\}\,,\\
&S^*_{j2}=\{z\,|\,E_j(L_1)+E_j[a_{j+1}(z_1,\cdots,z_j,x_{j+1})]\leqq E_j(L_2)+E_j[a_{j+1}(z_1,\cdots,z_j,y_{j+1})]\}\,,\\
&\qquad\qquad (0\leqq j\leqq N-1)\,.
\end{aligned} \tag{3.5}$$

Then $\mathfrak{S}^*_N=\{S^*_{n1},S^*_{n2};\,0\leqq n\leqq N\}$ forms a class of cylinder sets over partitions of n-state spaces for $n=0,1,\cdots,N$.

Thus $\mathfrak{S}^*_N$ is a possible policy and is characterized as follows: After the n-th stage of observations we compare the total expected reward from a policy with choice of X at $(n+1)$-th stage followed by an optimal policy for remaining stages and the one from a policy with choice of Y at $(n+1)$-th stage followed by an optimal policy for the remains. We choose X if the former is larger than the latter, and Y if otherwise. We shall show that $\mathfrak{S}^*_N$ is in fact a Bayes solution for our process.

LEMMA. *The policy* $\mathfrak{S}^*_N$ *defined by* (3.5) *is a Bayes solution against* ξ. *Furthermore* $\rho_{N+1}(\xi,\mathfrak{S}^*_N)=a_0$.

PROOF. Let $\mathfrak{S}_N=\{S_{n1},S_{n2};\,0\leqq n\leqq N\}$ be an arbitrary policy, and define

$$\begin{aligned}
R_n(\xi,\mathfrak{S}_N)=&\sum_{r=0}^{n-1}\sum_{i_0=1}^{2}\sum_{i_1=1}^{2}\cdots\sum_{i_r=1}^{2}\int_{T_{i_0i_1\cdots i_r}}E_r(L_{i_r})P_{\xi}(z)d\lambda\\
&+\sum_{i_0=1}^{2}\sum_{i_1=1}^{2}\cdots\sum_{i_n=1}^{2}\int_{T_{i_0i_1\cdots i_r}}a_n(z_{1i_0},z_{2i_1},\cdots,z_{ni_{n-1}})P_{\xi}(z)d\lambda\,.
\end{aligned} \tag{3.6}$$

Then

(3.7) $R_1(\xi, \mathfrak{S}_N) \leqq a_0$,

for

(3.8)
$$\begin{aligned} R_1(\xi, \mathfrak{S}_N) &= \int_{S_{01}} E(L_1)P_\xi(z)d\lambda + \int_{S_{02}} E(L_2)P_\xi(z)d\lambda \\ &\quad + \sum_{i_1=1}^{2} \int_{S_{01}\cap S_{1i_1}} a_1(z_{11})P_\xi(z)d\lambda + \sum_{i_1=1}^{2} \int_{S_{02}\cap S_{1i_1}} a_1(z_{12})P_\xi(z)d\lambda \\ &= \int_{S_{01}} E(L_1)P_\xi(z)d\lambda + \int_{S_{02}} E(L_2)P_\xi(z)d\lambda \\ &\quad + \int_{S_{01}} a_1(z_{11})P_\xi(z)d\lambda + \int_{S_{02}} a_1(z_{12})P_\xi(z)d\lambda \\ &= \int_{S_{01}} [E(L_1)+E(a_1(z_{11}))]P_\xi(z)d\lambda + \int_{S_{12}} [E(L_2)+E(a_1(z_{12}))]P_\xi(z)d\lambda \\ &\leqq \int_{S_{01}} a_0 P_\xi d\lambda + \int_{S_{02}} a_0 P_\xi d\lambda \\ &= a_0 , \end{aligned}$$

(3.9) $R_{n+1}(\xi, \mathfrak{S}_N) \leqq R_n(\xi, \mathfrak{S}_N)$,

for

(3.10)
$$\begin{aligned} R_{n+1}(\xi, \mathfrak{S}_N) &= \sum_{r=0}^{2} \sum_{i_0=1}^{2} \cdots \sum_{i_r=1}^{2} \int_{T_{i_0 i_1 \cdots i_r}} E_r(L_{i_r})P_\xi(z)d\lambda \\ &\quad + \sum_{i_0=1}^{2} \cdots \sum_{i_{n+1}=1}^{2} \int_{T_{i_0 i_1 \cdots i_{n+1}}} a_{n+1}(z_{1i_0}, z_{2i_1}, \cdots, z_{n+1 i_n})P_\xi(z)d\lambda \\ &= \sum_{r=0}^{n} \sum_{i_0=1}^{2} \cdots \sum_{i_r=1}^{2} \int_{T_{i_0 i_1 \cdots i_r}} E_r(L_{i_r})P_\xi(z)d\lambda \\ &\quad + \sum_{i_0=1}^{2} \sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} \int_{T_{i_0 i_1 \cdots i_n}} a_{n+1}(z_{1i_0}, z_{2i_1}, \cdots, z_{n+1,i_n})P_\xi(z)d\lambda \\ &= \sum_{r=0}^{n-1} \sum_{i_0=1}^{2} \cdots \sum_{i_r=1}^{2} \int_{T_{i_0 i_1 \cdots i_r}} E_r(L_{i_r})P_\xi(z)d\lambda \\ &\quad + \sum_{i_0=1}^{2} \sum_{i_n=1}^{2} \int_{T_{i_0 i_1 \cdots i_n}} [E_n(L_{i_n})+E_n(a_{n+1}(z_{1i_0}, z_{2i_1}, \cdots, z_{n+1,i_n}))]P_\xi(z)d\lambda \\ &\leqq \sum_{r=0}^{n-1} \sum_{i_0=1}^{2} \cdots \sum_{i_r=1}^{2} \int_{T_{i_0 i_1 \cdots i_r}} E_r(L_{i_r})P_\xi(z)d\lambda \\ &\quad + \sum_{i_0=1}^{2} \cdots \sum_{i_n=1}^{2} \int_{T_{i_0 i_1 \cdots i_n}} a_n(z_{1i_0}, z_{2i_1}, \cdots, z_{ni_{n-1}})P_\xi(z)d\lambda \\ &= R_n(\xi, \mathfrak{S}_N) . \end{aligned}$$

(3.11) $\rho_{N+1}(\xi, \mathfrak{S}_N) \leqq R_N(\xi, \mathfrak{S}_N)$,

for

(3.12)
$$\begin{aligned} R_N(\xi, \mathfrak{S}_N) &= \sum_{r=0}^{N-1} \sum_{i_0=1}^{2} \cdots \sum_{i_r=1}^{2} \int_{T_{i_0 i_1 \cdots i_r}} E_r(L_{i_r})P_\xi(z)d\lambda \\ &\quad + \sum_{i_0=1}^{2} \cdots \sum_{i_N=1}^{2} \int_{T_{i_0 i_1 \cdots i_N}} a_N(z_{1i_0}, z_{2i_1}, \cdots, z_{Ni_{N-1}})P_\xi(z)d\lambda \end{aligned}$$

$$\geqq \sum_{r=0}^{N-1} \sum_{i_0=1}^{2} \cdots \sum_{i_r=0}^{2} \int_{T_{i_0 i_1 \cdots i_r}} E_r(L_{i_r}) P_{\xi}(z) d\lambda$$

$$+ \sum_{i_0=1}^{2} \cdots \sum_{i_N=1}^{2} \int_{T_{i_0 i_1 \cdots i_N}} E_N(L_{i_N}) P_{\xi}(z) d\lambda$$

$$= \rho_{N+1}(\xi, \mathfrak{S}_N) .$$

Hence we have

$$\rho_{N+1}(\xi, \mathfrak{S}_N) \leqq R_N(\xi, \mathfrak{S}_N) , \tag{3.11}$$

for

$$R_N(\xi, \mathfrak{S}_N) = \sum_{r=0}^{N-1} \sum_{i_0=1}^{2} \cdots \sum_{i_r=1}^{2} \int_{T_{i_0 i_1 \cdots i_r}} E_r(L_{i_r}) P_{\xi}(z) d\lambda \tag{3.12}$$

$$+ \sum_{i_0=1}^{2} \cdots \sum_{i_N=1}^{2} \int_{T_{i_0 i_1 \cdots i_N}} a_N(z_{1 i_0}, z_{2 i_1}, \cdots, z_{N i_{N-1}}) P_{\xi}(z) d\lambda$$

$$\geqq \sum_{r=0}^{N-1} \sum_{i_0=1}^{2} \cdots \sum_{i_r=0}^{2} \int_{T_{i_0 i_1 \cdots i_r}} E_r(L_{i_r}) P_{\xi}(z) d\lambda$$

$$+ \sum_{i_0=1}^{2} \cdots \sum_{i_N=1}^{2} \int_{T_{i_0 i_1 \cdots i_N}} E_N(L_{i_N}) P_{\xi}(z) d\lambda$$

$$= \rho_{N+1}(\xi, \mathfrak{S}_N) .$$

Hence we have

$$\rho_{N+1}(\xi, \mathfrak{S}_N) \leqq R_N(\xi, \mathfrak{S}_N) \leqq R_{N-1}(\xi, \mathfrak{S}_N) \leqq \cdots \leqq R_1(\xi, \mathfrak{S}_N) \leqq a_0 \quad \text{for all } \mathfrak{S}_N . \tag{3.13}$$

On the other hand we get inductively

$$\rho_{N+1}(\xi, \mathfrak{S}_N^*) = R_N(\xi, \mathfrak{S}_N^*) = \cdots = R_1(\xi, \mathfrak{S}_N^*) = a_0 \tag{3.14}$$

by (3.4) and (3.5). Finally, from these results, we get

$$a_0 = \rho_{N+1}(\xi, \mathfrak{S}_N^*) \geqq \rho_{N+1}(\xi, \mathfrak{S}_N) \quad \text{for all } \mathfrak{S}_N , \tag{3.15}$$

which completes the proof.

§ 4. Main theorems.

In this section we shall consider the case in which the spaces U_i $(i=1, 2, \cdots, N)$ are the same (say U), the random variables X_i $(i=1, 2, \cdots, N)$ are independent with the same probability density $f_{\omega}(x)$ with respect to a measure μ on $(1, U)$, and the random variables Y_i $(i=1, 2, \cdots, N)$ independent with the same probability density $g_{\theta}(y)$ with respect to a measure ν on $(2, U)$.

Let $\xi_j(\omega, \theta)$ be the posterior probability density of (ω, θ) given the observations $z_1, z_2, \cdots, z_j$, then the posterior probability density given $z_1, z_2, \cdots, z_j, x_{j+1}$ becomes

$$\xi_{j+1}(\omega, \theta) = \frac{\xi_j(\omega, \theta) f_{\omega}(x_{j+1})}{\int_{\Omega} \xi_j(\omega, \theta) f_{\omega}(x_{j+1}) d\varphi} . \tag{4.1}$$

Similarly we have the posterior probability density given $z_1, z_2, \cdots, z_j, y_{j+1}$

$$\xi_{j+1}(\omega, \theta) = \frac{\xi_j(\omega, \theta) g_\theta(y_{j+1})}{\int_\Omega \xi_j(\omega, \theta) g_\theta(y_{j+1}) d\varphi} . \tag{4.2}$$

We let Ξ be the space of all distributions on Ω. Then a point ξ_j in Ξ is transformed respectively into a different point ξ_{j+1} in Ξ by the transformations given in (4.1) and (4.2). We shall designate the transformation (4.1) be T^X and the transformation (4.2) be T^Y, hence equations (4.1) and (4.2) can be expressed respectively as $\xi_{j+1} = T^X \xi_j$ and $\xi_{j+1} = T^Y \xi_j$.

THEOREM 1. *There exists an optimal policy among the class of policies depending outcomes only through a posteriori probability distributions of unknown parameters.*

PROOF. It suffices to prove that for any N and ξ there exists a class of partitions of Ξ, $\{\Xi_j^1, \Xi_j^2;\ 0 \leqq j \leqq N\}$, such that an optimal policy $\mathfrak{S}_N^*$ is expressed by

$$\begin{aligned} S_{j1}^* &= \{(z_1, z_2, \cdots, z_N) \mid \xi_j \in \Xi_j^1\} , \\ S_{j2}^* &= \{(z_1, z_2, \cdots, z_N) \mid \xi_j \in \Xi_j^2\} , \qquad \text{for} \quad j = 0, 1, \cdots, N . \end{aligned} \tag{4.3}$$

Define

$$\begin{aligned} H_0^1(\xi_j) &= E_j[L_1(\omega, \theta)] , \\ H_0^2(\xi_j) &= E_j[L_2(\omega, \theta)] , \end{aligned} \tag{4.4}$$

and

$$\begin{aligned} H_0(\xi_1) &= \operatorname{Max}[H_0^1(\xi_1), H_0^2(\xi_1)] , \\ H_j(\xi_l) &= \operatorname{Max}[H_0^1(\xi_l) + E_1(H_{j-1}(T^X \xi_l)), H_0^2(\xi_l) + E_1(H_{j-1}(T^Y \xi_l))] , \\ &\qquad (1 \leqq j \leqq N,\ 0 \leqq l \leqq N) , \end{aligned} \tag{4.5}$$

by induction on j. Consequently we have

$$\begin{aligned} a_N(z_1, z_2, \cdots, z_{N-1}, x_N) &= \operatorname{Max}[E_{N, z_1 z_2 \cdots z_{N-1} x_N}(L_1), E_{N, z_1 z_2 \cdots z_{N-1} x_N}(L_2)] \\ &= \operatorname{Max}[H_0^1(\xi_N(x_1 \cdots z_{N-1} x_N)), H_0^2(\xi_N(z_1 \cdots z_{N-1} x_N))] \\ &= \operatorname{Max}[H_0^1(T^X \xi_{N-1}), H_0^2(T^Y \xi_{N-1})] \\ &= H_0(T^X \xi_{N-1}) , \end{aligned} \tag{4.6}$$

usiug (3.4), (4.4) and (4.5). And similarly we obtain

$$a_N(z_1, z_2, \cdots, z_{N-1} y_N) = H_0(T^Y \xi_{N-1}) . \tag{4.7}$$

From (4.6) and (4.7) it follows that

$$\begin{aligned} &a_{N-1}(z_1, z_2, \cdots, z_{N-2}, x_{N-1}) \\ &\quad = \operatorname{Max}[E_{N-1}(L_1) + E_{N-1}\{a_N(z_1 \cdots z_{N-2} x_{N-1} x_N)\}, \\ &\qquad\qquad E_{N-1}(L_2) + E_{N-1}\{a_N(z_1 \cdots z_{N-2} x_{N-1} y_N)\}] \\ &\quad = \operatorname{Max}[H_0^1(\xi_{N-1}) + E_{N-1}\{H_0(T^X \xi_{N-1})\}, H_0^2(\xi_{N-1}) + E_{N-1}\{H_0(T^Y \xi_{N-1})\}] \\ &\quad = H_1(\xi_{N-1}(z_1 \cdots z_{N-2} x_{N-1})) = H_1(T^X \xi_{N-2}) , \end{aligned} \tag{4.8}$$

and similarly

$$a_{N-1}(z_1, z_2, \cdots, z_{N-2}, y_{N-1}) = H_1(T^Y \xi_{N-2}) . \tag{4.9}$$

In general we have by induction

$$
\begin{aligned}
&a_{N-j}(z_1, \cdots, z_{N-j-1}, x_{N-j}) = H_j(T^X\xi_{N-j-1}) \qquad (0 \leqq j \leqq N-1)\,,\\
&a_{N-j}(z_1, \cdots, z_{N-j-1}, y_{N-j}) = H_j(T^Y\xi_{N-j-1}) \qquad (0 \leqq j \leqq N-1)\,,\\
&a_0 = H_N(\xi)\,.
\end{aligned} \tag{4.10}
$$

Since an optimal policy is expressed by (3.5) in virtue of Lemma, if we let $\Xi_N^1 = \{\xi \mid H_0^1(\xi) > H_0^2(\xi)\}$ then from (4.4) we get

$$
\{z \mid \xi_N \in \Xi_N^1\} = \{z \mid H_0^1(\xi_N) > H_0^2(\xi_N)\} = \{z \mid E_N(L_1) > E_N(L_2)\} = S_{N1}^*\,. \tag{4.11}
$$

And similarly if we let $\Xi_N^2 = \{\xi \mid H_0^1(\xi) \leqq H_0^2(\xi)\}$, then

$$
\{z \mid \xi_N \in \Xi_N^2\} = S_{N2}^*\,. \tag{4.12}
$$

Let

$$
\Xi_j^1 = \{\xi \mid H_0^1(\xi) + E(H_{N-j-1}(T^X\xi)) > H_0^2(\xi) + E(H_{N-j-1}(T^Y\xi))\}
$$

for $0 \leqq j \leqq N-1$. Then we have from (4.4) and (4.10)

$$
\begin{aligned}
\{z \mid \xi_j \in \Xi_j^1\} &= \{z \mid H_0^1(\xi_j) + E_j(H_{N-j-1}(T^X\xi_j)) > H_0^2(\xi_j) + E_j(H_{N-j-1}(T^Y\xi_j))\}\\
&= S_{j1}^*\,,
\end{aligned} \tag{4.13}
$$

and similarly

$$
\{z \mid \xi_j \in \Xi_j^2\} = S_{2j}^* \qquad \text{for} \quad j = 0, 1, \cdots, N-1\,. \tag{4.14}
$$

It is easily seen from the definition that the one of Ξ_0^1 and Ξ_0^2 is an empty set and the other whole space. Hence the theorem is proved.

Now we shall prove the recurrence formula which an optimal policy satisfies.

THEOREM 2. *For any* N

$$
\rho_{N+1}^*(\xi) = \mathrm{Max}\,[E(L_1) + E(\rho_N^*(T^X\xi)),\ E(L_2) + E(\rho_N^*(T^Y\xi))]\,. \tag{4.15}
$$

PROOF. By Lemma and (4.10) we get

$$
\rho_{N+1}^*(\xi) = a_0 = H_N(\xi) \tag{4.16}
$$

for any N. Furthermore from (4.4), (4.5) and (4.16) it follows that

$$
\begin{aligned}
H_N(\xi) &= \mathrm{Max}\,[H_0^1(\xi) + E(H_{N-1}(T^X\xi)),\ H_0^2(\xi) + E(H_{N-1}(T^Y\xi))]\\
&= \mathrm{Max}\,[E(L_1) + E(\rho_N^*(T^X\xi)),\ E(L_2) + E(\rho_N^*(T^Y\xi))]\,,
\end{aligned} \tag{4.17}
$$

which completes the proof.

§ 5. Remarks.

In this section we shall refer to the case of an infinite horizon.

Since we can now redefine the space of all ξ's, Ξ, as a state space in virtue of Theorem 1, the Bayes controlled process is a Markovian decision process in the sence of Blackwell [1], [2] and Strauch [7]. Consequently from their results, in a Bayes controlled process with an infinite horizon, the existence of an optimal policy implies the existence of an optimal stationary policy.

From Theorem 2 we have a functional equations, which an optimal policy satisfy, as follows;

$$\rho^*(\xi) = \mathrm{Max}\,[E(L_1)+E(\rho^*(T^X\xi)),\ E(L_2)+E(\rho^*(T^Y\xi))]\,, \tag{5.1}$$

where ρ^* is a total expected reward from an optimal policy over infinite future.

Now above considerations lead us to the fact that under assuming the existence of an optimal policy the solution of the functional equation (5.1) becomes an optimal stationary policy. The author wishes to discuss the solving the equation (5.1) under some restriction for the reward function in another occasion.

The author would like to mention the work of Dynkin [5]. Dynkin gives a sufficient partition of policies in the infinite stage controlled process, which corresponds to the statement of Theorem 1 in our case.

References

[1] Blackwell, D. (1962). Discrete dynamic programming. Ann. Math. Statist. **33** 719–726.
[2] Blackwell, D. (1965). Discounted dynamic programming. Ann. Math. Statist. **36** 226–235.
[3] Blackwell, D. and Girshick, M. A. (1954). *Theory of games and statistical decisions*. John Wiely.
[4] Bradt, R. N., Johnson, S. M. and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. Ann. Math. Statist. **27** 1060–1074.
[5] Dynkin, E. B. (1965). Controlled stochastic processes. Theory of prob. and its appl. **10** 3–8 (In Russian)
[6] Feldman, P. (1962). Contribution to the "Two-armed Bandit" problem. Ann. Math. Statist. **33** 847–856.
[7] Strauch, R. E. (1966). Negadive dynamic programming. Ann. Math. Statist. **37** 871–890.