Journal of the
Operations Research
Society of Japan

VOLUME 6 *August 1964* *NUMBER 4*

RESEARCH IN SEMI-MARKOVIAN DECISION STRUCTURES

RONALD A. HOWARD

Operations Research Center
Massachusetts Institute of Technology

Introduction

The complexity of modern management problems requires correspondingly sophisticated models for management systems. We shall here discuss a statistical model useful in analyzing a wide variety of common problems, problems in the areas of maintenance, replacement, marketing, finance, and inventory control, for example. This model has been under continuing development at M. I. T. for some years. Its present form is the result of encountering practical situations that required successive generalizations of the model.

The basis for the present model is a decision model developed for strictly Markov processes[1]—processes that satisfy the Chapman-Kolmogorov equations. These processes can be characterized as processes in which the main interest is on state transitions rather than on the time required for a transition. The time between transitions can be considered very easily if transitions occur at regularly spaced points of time or if the times of transition are exponentially distributed. The case of regularly spaced transitions occurs frequently in practice because many decisions are made on a weekly, monthly, or annual basis. The case of exponential transition times, however, does not seem to arise very often.

The desire to allow other types of transition behavior was encouraged by the development of the semi-Markov process[2,3,4]. This process allowed the time between transitions to be a random variable conditional on the transition made. The decision model for the strictly Markov process had interesting properties primarily because its structure involved only the limiting state probabilities of the Markov process. When it was noted that a semi-Markov process had the same limiting state probabilities as an exponential Markov process with the same mean times for transitions, it was natural to expect that the decision model could be extended to the semi-Markov process. In this paper we shall point out some of the results of this development and show how this increase in generality does not exact a computational penalty.*

Semi-Markov Processes

Let us begin by defining those parts of semi-Markov process theory that we shall need. Consider a process with N states. Let p_{ij} be the conditional probability that the next transition is to state j given that the last transition was to state i. We shall call these probabilities the transition probabilities; they must satisfy the conditions

$$\sum_{j=1}^{N} p_{ij} = 1 \qquad i = 1, 2, \ldots, N \tag{1}$$

$$p_{ij} \geq 0 \qquad 1 \leq i,\ j \leq N \tag{2}$$

Whenever a process enters a state, we imagine that it selects from these probabilities to determine the next state to which it will move. However, before making this transition, it "holds" for a time τ_{ij} in state i where i is the index of the present state and j that of the state selected as its next state. The quantities τ_{ij} are random variables governed by a corresponding set of density functions, $h_{ij}(.)$, $1 \leq i, j \leq N$. These density functions are called the "holding time" density functions. In general we must specify N^2 of them to determine a semi-Markov process. After holding in state i for the time τ_{ij} the process makes the transition to j and then repeates the whole procedure.

We see that, if we ignore the random nature of transition times and focus on

* Some of the results in this paper were also discovered independently by J. de Cani (University of Pennsylvania), W. S. Jewell (University of California, Berkeley) and P. Schweitzer (Massachusetts Institute of Technology).

the transition instants, the process is an ordinary Markov process. However, when the hold behavior is included the process will not satisfy the Chapman-Kolmogorov equations unless the holding times are all exponentially distributed. Therefore the process is Markovian only at the transition instants; hence the name "semi-Markov" process.

We shall find it useful to develop additional notation for the holding time behavior. We shall use ${}^{c}h_{ij}(.)$ for the cumulative probability distribution of τ_{ij},

$$ {}^{c}h_{ij}(t) = \int_0^t h_{ij}(\tau)d\tau = p\left\{\tau_{ij} \leq t\right\}. \tag{3} $$

and ${}^{cc}h_{ij}(.)$ for the complementary cumulative probability distribution of τ_{ij},

$$ {}^{cc}h_{ij}(t) = \int_t^{\infty} h_{ij}(\tau)d\tau = 1 - {}^{c}h_{ij}(t) = p\left\{\tau_{ij} > t\right\}. \tag{4} $$

We shall assume that the means $\bar{\tau}_{ij}$ of all the holding time distribution are finite and that all the holding time density functions have no impulse component at the origin.

Suppose now that the process enters state i and chooses its successors state j, but that we as observers do not know the successor chosen. The density function that we would assign to its holding time in i, τ_i, would then be $w_i(.)$, where

$$ w_i(t) = \sum_{j=1}^{N} p_{ij} h_{ij}(t). \tag{5} $$

We shall call this function the unconditional waiting time density function for state i. The mean unconditional wait in state i, $\bar{\tau}_i$, is related to the means of the holding times for state i by

$$ \bar{\tau}_i = \sum_{j=1}^{N} p_{ij} \bar{\tau}_{ij}. \tag{6} $$

The cumulative and complementary cumulative probability distributions for the unconditional wait are given by

$$ {}^{c}w_i(t) = \sum_{j=1}^{N} p_{ij} {}^{c}h_{ij}(t) = \int_0^t w_i(\tau)d\tau = p\left\{\tau_i \leq t\right\}, \tag{7} $$

$$ {}^{cc}w_i(t) = \sum_{j=1}^{N} p_{ij} {}^{cc}h_{ij}(t) = \int_t^{\infty} w_i(\tau)d\tau = p\left\{\tau_i > t\right\}. \tag{8} $$

When modeling real systems with semi-Markov processes we must decide whether we want to call a movement from a state to itself a transition of the process. We find it convenient to make the decision different ways in different circumstances. To distinguish we shall say that the system makes a "real" transition only when its state number actually changes, but that it makes a "virtual" transition when it re-enters the same state. Some models require that only real transitions be allowed, in others the virtual transitions are the most important. For example, if we are studying a machine maintenance problem in which the state of the system is the number of machines working, then only real transitions need be considered because a virtual transition would require a simultaneous breakdown and repair. On the other hand in a marketing problem the state variable might be the last brand purchased by the customer. In this case virtual transitions would correspond to repeat purchases of the same brand, a very important phenomenon. In our development we shall allow the possibility of virtual transitions; the diagonal transition probabilities p_{ii} may or may not be zero.

Now let us proceed to the question of state probabilities for a semi-Markov process. Let $\phi_{ij}(t)$ be the probability that the system is in state j at time t given that it entered state i at time zero. We shall call these probabilities the interval transition probabilities of the process. We can write a recursion equation for $\phi_{ij}(t)$ as follows. A system starting in state i can be in state j at time t either because $i = j$ and it never left i during that interval or because it left i at least once and finally managed to reach j by time t. The probabilities of these two mutually exclusive possibilities are added in this equation,

$$\phi_{ij}(t) = \delta_{ij}{}^{cc}w_i(t) + \sum_{k=1}^{N} p_{ik} \int_0^t d\tau h_{ik}(\tau)\ \phi_{kj}(t-\tau)$$
$$1 \leq i,\quad j \leq N;\quad t \geq 0 \qquad (9)$$
$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

The quantity δ_{ij} assures that the term in which it appears occurs only when $i = j$. The complementary cumulative probability ${}^{cc}w_i(t)$ is the probability that the system will not leave its starting state i until after time t. The second term in the equation represents the probability of the sequence of events where the system

makes a transition from i to some state k (maybe itself) at time τ and then proceeds to state j from k in the remaining time $t-\tau$. This probability is summed over all states k to which the initial transition could have been made and integrated over all times of transition τ between 0 and t.

Solving this recurrence equation in its present form is not particularly easy. However, exponential transformation simplifies it considerably. Let

$$f^T(s)=\int_0^\infty f(t)e^{-st}dt \tag{10}$$

be the exponential transform of the function $f(t)$. Then we find the exponential transform of Equation 9 to be

$$\phi_{ij}{}^T(s)=\delta_{ij}{}^{cc}w_i{}^T(s)+\sum_{k=1}^{N} p_{ik}h_{ik}{}^T(s)\ \phi_{kj}{}^T(s), \qquad 1\leq i,\ j\leq N. \tag{11}$$

Since from Equation 8,

$${}^{cc}w_i(t)=\int_t^\infty w_i(\tau)\,d\tau, \tag{12}$$

the exponential transform of the complementary cumulative waiting time distribution in state i is

$${}^{cc}w_i{}^T(s)=\frac{1}{s}\left[1-w_i{}^T(s)\right]. \tag{13}$$

Equation 11 now relates the transform of the interval transition probabilities to the basic process quantities.

A matrix formulation of Equation 11 will prove worthwhile. We shall define matrix of any double-indexed quantity to be the corresponding upper case letter. Thus $\Phi(t)=\{\phi_{ij}(t)\}$, $P=\{p_{ij}\}$, etc. We shall also define diagonal matrices $W(t)=\{\delta_{ij}w_i(t)\}$, ${}^cW(t)=\{\delta_{ij}{}^cw_i(t)\}$, and ${}^{cc}W(t)=\{\delta_{ij}{}^{cc}w_i(t)\}$. The special form of the summation in Equation 11 makes it worthwhile to define a special kind of matrix multiplication designated by $\square$. If A, B, and C are N by N square matrices, then $C=A\square B$ implies $c_{ij}=a_{ij}b_{ij}$, $1\leq i,\ j\leq N$. In other words the "box" operation is an element by element multiplication.

Now we can write Equation 11 in the form

$$\Phi^T(s) = {}^{cc}W^T(s) + (P \square H^T(s))\, \Phi^T(s). \tag{14}$$

The transform of the interval transition probability matrix $\Phi^T(s)$ is thus

$$\Phi^T(s) = [I - P \square H^T(s)]^{-1}\, {}^{cc}W^T(s). \tag{15}$$

The inverse matrix will always exist for the type of process we are considering. We see that the interval transition probabilities depend only on the products of p_{ij} and $h_{ij}(t)$, not on the individual quantities.

Equation 15 provides the means for relating the interval transition probabilities of a semi-Markov process to the process parameters. However, for the purposes of this paper we shall not need to explore the transient behavior of interval transition probabilities, but only their limiting form. Let us define a limiting interval transition probability matrix for the process by

$$\Phi = \lim_{t \to \infty} \Phi(t). \tag{16}$$

By the final value theorem of exponential transforms Φ is also given by

$$\Phi = \lim_{s \to 0} s\, \Phi^T(s). \tag{17}$$

We use Equation 15 to write this limit in the form

$$\Phi = \lim_{s \to 0} s\, \Phi^T(s) = \lim_{s \to 0} s\, [I - P \square H^T(s)]^{-1} \lim_{s \to 0} {}^{cc}W^T(s). \tag{18}$$

We now consider each of the limits on the right of Equation 18 individually. First, by Equation 13,

$$\lim_{s \to 0} {}^{cc}W^T(s) = \lim_{s \to 0} \frac{1}{s}\left[I - W^T(s)\right] . \tag{19}$$

where I is the identity matrix. However, because $W^T(0) = I$, the limit is indeterminate and we must resort to L'Hospital's rule. We find

$$\lim_{s \to 0} {}^{cc}W^T(s) = -\frac{d}{ds} W^T(s) \Big|_{s=0} = \int_0^\infty t\, W(t)\, dt = M. \tag{20}$$

The matrix M is therefore a diagonal matrix of the mean unconditional waiting times in each of the states of the process.

Now we find the other limit in Equation 18,

$$\lim_{s\to 0} s\,[I-P\,\square\, H^T(s)]^{-1}=\lim_{s\to 0} T(s). \tag{21}$$

where we have written $T(s)$ for s times the inverse matrix. Then

$$T(s)=s\,[I-P\,\square\, H^T(s)]^{-1}$$

or

$$T(s)-T(s)\,P\,[\square\, H^T(s)=sI. \tag{22}$$

If we take the limit of Equation 22 as s approaches zero and note that $H^T(0)$ is a matrix with all elements equal to one, we obtain

$$T(0)=T(0)P. \tag{23}$$

We shall now show that the rows of the $T(0)$ matrix must be proportional to the limiting state probabilities of the imbedded Markov process that describes the transitions of the process without respect to their duration. Let us suppose in the remainder of this paper that this imbedded Markov process has only one recurrent chain and therefore a unique set of limiting state probabilities independent of the starting state. What we shall say about this case can be extended fairly easily to the multiple chain problem. Let π_i be the limiting probability of state i for the imbedded process. Then these probabilities must satisfy the equation,

$$\pi_j=\sum_{i=1}^{N}\pi_i\,p_{ij}, \qquad \sum_{i=1}^{N}\pi_i=1. \tag{24}$$

Since the solution is unique for a single chain process, the rows of the $T(0)$ matrix must each be proportional to the row vector Π with elements π_i.

We have now reduced Equation 18 to the form

$$\Phi=T(0)\,M. \tag{25}$$

The elements of Φ must therefore satisfy

$$\phi_{ij}=t_{ij}(0)\bar{\tau}_j=k_i\pi_j\bar{\tau}_j \tag{26}$$

where k_i is an undetermined constant of proportionality between the i^{th} row of $T(0)$ and Π, the limiting state probability vector for the imbedded process. We can now

use the condition that the limiting interval transition probabilities ϕ_{ij} must sum to one over all the states of the process. Thus

$$\sum_{j=1}^{N} \phi_{ij} = 1 = k_i \sum_{j=1}^{N} \pi_j \bar{\tau}_j \tag{27}$$

and therefore

$$k_i = \frac{1}{\sum_{j=1}^{N} \pi_j \bar{\tau}_j} . \tag{28}$$

The constant of proportionality is then the same for all states and we can write from Equation 26 and 28,

$$\phi_{ij} = \frac{\pi_j \bar{\tau}_j}{\sum_{j=1}^{N} \pi_j \bar{\tau}_j} = \phi_j . \tag{29}$$

As we would expect, the limiting interval transition probabilities ϕ_{ij} do not depend on the starting state i. We shall, therefore use only the second subscript to indicate these quantities. Furthermore, these probabilities are equal to the limiting state probabilities of the imbedded Markov process weighted by the mean unconditional waiting times in each state. This result is intuitive and important: the only statistic of each holding time distribution that affects the limiting behavior of the process is its mean.

A Reward Structure

We shall now superimpose on the semi-Markov process a system of rewards. The system will earn both by making transitions and by staying in a state. Let r_{ij} be the transition reward for a transition from state i to state j. It is a lump sum quantity paid when the transition occurs. Let y_i be the yield rate for state i, the amount the system will earn per unit time for all the time it stays in state i. A large class of physical processes can be modeled by this type of reward structure or by simple variations of it. For example, a machine maintenance system whose state was the number of machines in operation might earn profit at a rate corresponding to the number of operation and incur a fixed charge whenever a breakdown occurs.

An important quantity in a system with a reward structure is the total reward

we expect the system to earn in a time t if it enters some state, say state i, at time zero; let this quantity be represented by $v_i(t)$. The system will either stay in i throughout the time t or else make a transition out of state i (but possibly to enter state i again in a virtual transition). If the system stays in state i for the entire interval, then it will earn the yield rate y_i multiplied by the time t. It will also earn whatever terminal reward may be associated with being in state i at the end of the time t, a quantity we shall denote by $v_i(0)$. Such a terminal reward may be extremely important in certain systems where scrap value or another such terminal cost must be included in the formulation. The probability that the system will not leave state i during the interval is just ${}^{cc}w_i(t)$. Therefore the contribution to the expected total reward $v_i(t)$ due to the possibility of the system's not leaving state i is ${}^{cc}w_i(t)[v_i(0)+y_it]$.

The system may leave state i for some state j at some time τ between zero and t. If it does so it will earn transition reward r_{ij}, and the yield rate y_i for a time τ. However, it will also be in a position to earn whatever expected reward is associated with entering state j when a time $t-\tau$ remains in the process, $v_j(t-\tau)$. The sum of these quantities must be multiplied by the probability that the system will make its necessary transition to state j, p_{ij}, and by the probability density function of the hold time for this transition evaluated at the point τ, $h_{ij}(\tau)$. The result must then be summed over all states j in the system and integrated over all values of τ between 0 and t. When these operations have been performed we obtain as the final expression for the expected total reward,

$$v_i(t)={}^{cc}w_i(t)\Big[v_i(0)+y_it\Big]+\sum_{j=1}^{N}p_{ij}\int_0^t d\tau h_{ij}(\tau)\ \Big[r_{ij}+y_i\tau+v_j(t-\tau)\Big],$$
$$i=1,2,\ldots,N;\quad t\geq 0. \tag{30}$$

Equation 30 provides a method for calculating total expected rewards for any starting state and any time period. However, the calculation is involved for all but the simplest cases. The numerical analysis required to solve the equation essentially reduces the problem to that of the fixed transition time Markov process because of the necessity of considering discrete time increments. Consequently, we shall focus our interest here on the behavior of this equation in the situation where t is very large; that is, where the process is going to be allowed to run for a relatively long

time. Fortunately, the behavior we shall study is of practical importance in most processes after only a few transitions are made.

When t is very large ${}^{cc}w_i(t)$ approaches zero because we are concerned only with systems having finite mean holding times. Furthermore, it is easy to show that $t\,{}^{cc}w_i(t)$ approaches zero as well. Therefore when t is very large in Equation 30

$$v_i(t)=\sum_{j=1}^{N} p_{ij}\int_0^\infty d\tau h_{ij}(\tau)\Big[r_{ij}+y_i\tau+v_j(t-\tau)\Big]$$

$$i=1,2,\ldots,N;\quad \text{for } t \text{ large} \qquad (31)$$

or

$$v_i(t)=\sum_{j=1}^{N} p_{ij}r_{ij}+y_i\bar{\tau}_i+\sum_{j=1}^{N} p_{ij}\int_0^\infty d\tau h_{ij}(\tau)v_j(t-\tau)$$

$$i=1,2,\ldots,N;\quad \text{for } t \text{ large} \qquad (32)$$

Let us use a quantity q_i defined by

$$q_i=\frac{1}{\bar{\tau}_i}\sum_{j=1}^{N} p_{ij}r_{ij}+y_i \qquad i=1,2,\ldots,N \qquad (33)$$

and call it "earning" rate in state i to simplify Equation 32. We can then write

$$v_i(t)=q_i\bar{\tau}_i+\sum_{j=1}^{N} p_{ij}\int_0^\infty d\tau h_{ij}(\tau)v_j(t-\tau)$$

$$i=1,2,\ldots,N;\ \text{for } t \text{ large} \qquad (34)$$

By the basic Equation 30 it is possible to show that when t is large $v_i(t)$ has the form

$$v_i(t)=v_i+gt \qquad i=1,2,\ldots,N;\quad \text{for } t \text{ large.} \qquad (35)$$

Indeed, all of our present results can be obtained rigorously, if somewhat laboriously, by transform methods. However, the present development will illustrate the properties we need. Equation 35 says that the expected reward will grow linearly with t when t is large. The growth rate is g, a quantity that we call the gain of the process. The gain is thus the average reward per unit time that the system will earn in the steady state. As we would expect it is independent of the state in which the system is started.

Let us now substitute the result given by Equation 35 into Equation 34. We find

$$v_i+gt=q_i\bar{\tau}_i+\sum_{j=1}^{N}p_{ij}\int_0^\infty d\tau h_{ij}(\tau)\Big[v_j+(t-\tau)g\Big]$$
$$=q_i\bar{\tau}_i+\sum_{j=1}^{N}p_{ij}v_j+gt-g\bar{\tau}_i\,. \tag{36}$$

The quantity gt appears on both sides of this equation and therefore vanishes. We have thus found that for large t,

$$v_i+g\bar{\tau}_i=q_i\bar{\tau}_i+\sum_{j=1}^{N}p_{ij}v_j \qquad i=1,2,\ldots,N \tag{37}$$

Equation 37 allows us to solve for the gain of the process in terms of the transition probabilities and the mean unconditional waits in each state. The equation is homogeneous in the v_i's; we can set one of the quantities v_i arbitrarily. We shall use this freedom to set $v_N=0$. Then for the type of process we are considering, the remaining values will be determined by these equations to within a constant as a direct result of solving the equations implied by expression 37. We can therefore find the gain of the process and also a set of quantities that differ from the quantities v_i by the same constant. We shall find that the solution of these equations plays a central role in the decision structure we shall soon add.

There is another result obtained quite readily from Equation 37. First we multiply this equation by the limiting state probability for state i in the imbedded Markov process and then sum the result for all values of i. The result is

$$\sum_{i=1}^{N}\pi_i v_i+g\sum_{i=1}^{N}\pi_i\bar{\tau}_i=\sum_{i=1}^{N}\pi_i q_i\bar{\tau}_i+\sum_{i=1}^{N}\pi_i\sum_{j=1}^{N}p_{ij}v_j$$
$$=\sum_{i=1}^{N}\pi_i q_i\bar{\tau}_i+\sum_{j=1}^{N}v_j\sum_{i=1}^{N}\pi_i p_{ij}\,. \tag{38}$$

From Equation 24 we see that the first and last terms of this equation are equal. It reduces to

$$g\sum_{i=1}^{N}\pi_i\bar{\tau}_i=\sum_{i=1}^{N}\pi_i q_i\bar{\tau}_i \tag{39}$$

or

$$g=\frac{\sum_{i=1}^{N}\pi_i\bar{\tau}_i q_i}{\sum_{i=1}^{N}\pi_i\bar{\tau}_i}\,. \tag{40}$$

Now if we apply Equation 29 we can write

$$g = \sum_{i=1}^{N} \phi_i q_i \,. \tag{41}$$

The gain of the process is just the sum of the earning rates for each state weighted by the limiting interval transition probabilities for the process. This result is consistent with our intuition about the process.

The semi-Markov process with a reward structure is an excellent model for several stochastic processes. It offers a method by which general results can often be obtained with little difficulty. For example, consider the problem of establishing the order quantity Q for an inventory system with the following characteristics. A single item is subject to random demand with the time between succeeding demands selected independently from the same probability density function; mean time between demands is $1/\lambda$. No stockouts are permitted, leadtime is zero, there is a fixed cost A of placing an order, and a carrying charge of r per unit time to be assessed against each item in inventory. We see that we can model this inventory system as a semi-Markov process whose state index i represents the number of items in inventory at any time. The state index can range from 1 to Q,the order quantity. It cannot be zero because when the last item in inventory is sold the amount of stock in inventory jumps instantaneously to Q. The transition probability structure is therefore very simple: the imbedded Markov process is periodic with period Q. The state index falls by one unit with each transition (demand) unless there is only one unit in inventory in which case it become Q. Therefore the limiting state probability for each state in the imbedded Markov process is just $1/Q$,

$$\pi_i = 1/Q \qquad i = 1, 2, \ldots, Q \,.$$

The holding time in every state is the density function of time between demands. The mean holding time in any state is therefore

$$\bar{r}_i = \frac{1}{\lambda} \qquad i = 1, 2, \ldots, Q \,.$$

Since the expected holding time in all states is the same, we see from Equation 29 that the limiting interval transition probabilities ϕ_i will be equal to the limiting

state probabilities π_i,

$$\phi_i = \pi_i = \frac{1}{Q} \qquad i = 1, 2, \ldots, Q.$$

The reward structure for this example is also simple. Since this system can only lose money we shall deal in costs rather than rewards. There is no transition cost associated with any transition except the transition from state 1 to state Q. The cost associated with this transition is the ordering cost,

$$r_{1Q} = A$$
$$r_{ij} = 0 \qquad \text{otherwise}$$

The yield rate in any state is the inventory carry charge rate of that state. For state i this will be ir,

$$y_i = ir \qquad i = 1, 2, \ldots, Q.$$

We can now calculate the earning rate in each state using Equation 33 and these results. We find

$$q_i = \frac{1}{r_i} \sum_{j=1}^{Q} p_{ij} r_{ij} + y_i \qquad i = 1, 2, \ldots, Q,$$
$$q_1 = \lambda A + r$$
$$q_i = ir \qquad i = 2, 3, \ldots, Q.$$

Now that we have obtained both the limiting interval transition probabilities and the earning rates for the system we can use Equation 41 to write the gain, which in this case is the average cost per unit time of operating the system in the steady state. We find

$$g = \sum_{i=1}^{Q} \phi_i q_i = \frac{1}{Q} \sum_{i=1}^{Q} q_i$$
$$= \frac{1}{Q}\left[\lambda A + r \sum_{i=1}^{Q} i\right]$$
$$= \frac{1}{Q}\left[\lambda A + r \frac{Q(Q+1)}{2}\right]$$
$$= \frac{\lambda A}{Q} + \frac{r(Q+1)}{2}$$

Now all that remains is to find the value of Q that minimizes this average cost rate. If we ignore the discrete nature of Q, differentiate g with respect to Q and set the derivative equal to zero, we find that the minimizing Q is given by

$$Q = \sqrt{\frac{2A\lambda}{r}} ,$$

which is, of course, the famous Wilson lot-size formula. The best integral value of Q can be found by difference methods if necessary.

Note that the generality of the semi-Markov process has allowed us to establish this result for demand processes that range from Poisson to constant time between purchases. Such modifications as quantity-dependent ordering costs and inter-demand distributions that depend on the inventory level would complicate the model only slightly. Even the problem of multiple-demands occurring at the same time can be treated. The freedom to expand the problem under consideration in several interesting ways at small increase in difficulty is a particularly appealing feature of semi-Markov models.

A Decision Structure

Suppose now that when the system is in state i there are various alternatives for its operation. Associated with the k^{th} alternative in state i are all the process parameters: p_{ij}^k, $h_{ij}^k(.)$, y_i^k, and r_{ij}^k. They are given a superscript k to indicate this association. Thus if a system enters state i and follows alternative k in that state, the next transition that will be made, the time of that transition, the yield rate for occupancy of state i, and the reward for the transition to the next state will all be governed by the alternative. We shall assume that the alternatives available are finite but that they may be different in number from one state to another. The problem we pose is this. Suppose we are free to select one alternative in each state for the operation of the system. We shall call such a selection in any one state a decision and in all states a policy. The policy once established will be fixed for all time—every time the system enters a state it will follow the alternative dictated by the policy. The problem is now simply stated: Find the policy that maximizes the gain (average reward per unit time in the steady state) for the system.

As in the case of strictly Markovian processes, the solution can be found by an iteration scheme. We shall first establish its plausibility and then prove that it works. If we solve Equation 37 for g, we obtain

$$g = q_i + \frac{1}{\bar{\tau}_i}\left[\sum_{j=1}^{N} p_{ij} v_j - v_i\right] \qquad i = 1, 2, \ldots, N. \tag{42}$$

It seems reasonable that to maximize g we want to make the right hand side of this equation as large as possible. That is exactly what we do in the iteration scheme. We begin with an arbitrary policy for the system and then solve Equation 37 to determine the gain of this policy and the quantities v_i which we shall call the relative values. Then we use these v_i's to improve our policy by finding the alternative in each state i that maximizes the test quantity

$$q_i^k + \frac{1}{\bar{\tau}_i^k}\left[\sum_{j=1}^{N} p_{ij}^k v_j - v_i\right] \tag{43}$$

We shall not change the old decision unless another alternative has a strictly greater value of the test quantity. When a new decision (possibly the same) has been made in each state, a new policy has been found. Note that the relative nature of the v_i's does not affect the policy improvement process because the test quantity, expression 43, is unchanged if any constant is added to all the v_i's. The test quantity involves the transition probabilities of each alternative, p_{ij}^k, the mean unconditional holding time in state i under that alternative, $\bar{\tau}_i^k$ and the earning rate it prescribes for state i, q_i^k. The earning rate for the alternative is in turn defined by Equation 33 as

$$q_i^k = \frac{1}{\bar{\tau}_i^k} \sum_{j=1}^{N} p_{ij}^k r_{ij}^k + y_i^k \qquad i = 1, 2, \ldots, N. \tag{44}$$

Consequently we see that the holding time density functions enter into the calculations only through their means. We do not find this result surprising because we already know that the limiting interval transition probabilities for the semi-Markov process depend only on the same characteristics of the holding time density functions and because the gain depends only on the earning rates and the limiting interval transition probabilities.

The iteration cycle is shown in Fig. 1. We select an arbitrary policy and evaluate it in the upper box in the figure. Then we enter the policy improvement box with

the relative values for this policy and find the alternative in each state that maximizes the test quantity. When we have performed this operation for all states in the system, a new policy has been found. We shall soon show that this policy can only increase in gain as a result of the policy improvement operation. The iteration terminates when the policy on two successive cycles is identical.

Let us illustrate the type of calculation involved with an example. Suppose that a machine facility can be in either of two states. It is in state 1 when the machine is operating and in state 2 when the machine is out of order. When the machine is in state 1 it can be maintained with either normal or expensive maintenance; these two options are the two alternatives in state 1. Under either maintenance policy the machine will sooner or later break down so that $p_{11}^k = 0, p_{12}^k = 1$ for either alternative. However, the density function of time until the next breakdown will be different in each case. Under alternative 1, normal maintenance, the time to breakdown will have the density function $h_{12}^1(t) = 5e^{-5t}$. Under alternative 2, expensive maintenance, it will be $h_{12}^2(t) = 16te^{-4t}$. The yield rate for normal maintenance will be larger than

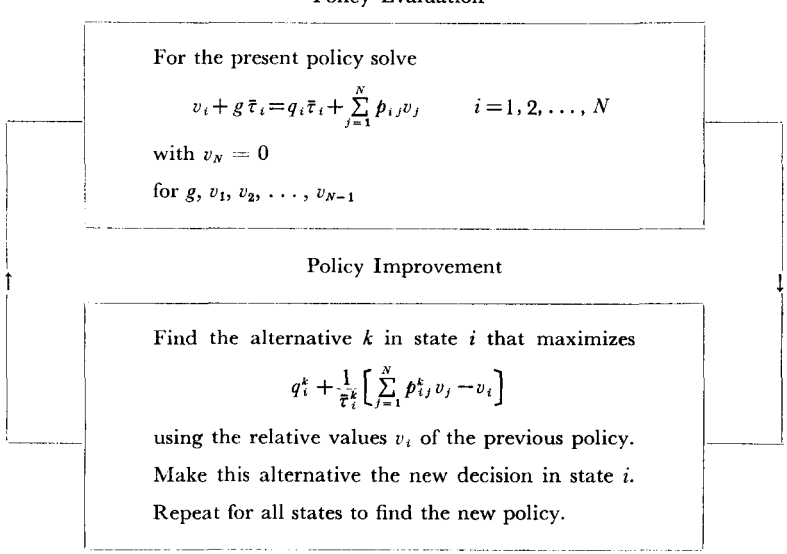

Fig. 1 The Iteration Cycle

under expensive maintenance because of the higher maintenance cost; in fact, $y_1^1 = 6$, $y_1^2 = 4$. There is no transition reward associated with the transition from state 1 to state 2.

When the system is in state 2 and the machine is out of order, there are again two alternatives, inside-plant and outside-plant repair crews. Under either alternative, the machine will sooner or later be fixed: $p_{21}^k = 1$, $p_{22}^k = 0$. The time for the repair, $h_{21}^1(t) = 64te^{-8t}$; for outside repair it is $h_{21}^2(t) = 7e^{-7t}$. There is a fixed charge of 0.5 for restarting the machine after either repair, and a charge per unit time of 1 for the inside crew and 1.5 for the outside crew: $r_{21}^1 = r_{21}^2 = -0.5$, $y_2^1 = -1$, $y_2^2 = -1.5$.

The data for the example are summarized in Table I. The mean time to the next transition under each alternative is of particular interest. We see that the effect of the more expensive maintenance is to increase the expected time until the next breakdown, and that the effect of using the outside repair crew is to shorten the expected time for completion of the repair. The earning rates for each state and alternative calculated using Equation 44 also appear in the table. One way of finding the policy that maximizes the gain of the process would be first to find the limiting interval transition probabilities for each of the 4 possible policies in this problem by

State	Alternative	Transition Probabilities		Holding Time Densities		Mean Unconditional Waits
i	k	p_{i1}^k	p_{i2}^k	$h_{i1}^k(t)$	$h_{i2}^k(t)$	$\bar{\tau}_i^k$
1	1	0	1	—	$5e^{-5t}$	1/5
	2	0	1	—	$16te^{-4t}$	1/2
2	1	1	0	$64te^{-8t}$	0	1/4
	2	1	0	$7e^{-7t}$	0	1/7

Transition Rewards		Yield Rates	Earning Rates
r_{i1}^k	r_{i2}^k	y_i^k	q_i^k
—	0	6	6
—	0	4	4
−0.5	—	−1	−3
−0.5	—	−1.5	−5

Table I. A Machine Repair Example

using Equation 29 with $\pi_1 = \pi_2 = 1/2$. Then we could use Equation 41 to find the gain of each policy and finally choose the one with the maximum gain. This procedure is feasible for this problem, but not for even a slightly larger one because the number of possible policies grows very quickly. The iteration method is valuable because it can treat problems that would be difficult to solve by exhaustion.

We begin the iteration procedure by choosing a policy arbitrarily. A convenient first choice is the policy that maximizes the earning rate in each state. In this example it is the policy formed by the first alternative in each state. We shall denote a policy by a column vector $\underline{d}$ whose i_{th} element is the alternative selected in the i_{th} state. Therefore the initial policy is described by

$$\underline{d} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{45}$$

We now write the policy evaluation equations 37 for this policy,

$$\begin{aligned} v_1 + 1/5g &= 6/5 + v_2 \\ v_2 + 1/4g &= -3/4 + v_1 \end{aligned} \tag{46}$$

When we set $v_2 = 0$, we find immediately, $g = 1$, $v_1 = 1$. The gain of the system under the initial policy is 1 per unit time.

Now we try to improve the policy by maximizing the test quantity 43 in each state using the values $v_1 = 1$, $v_2 = 0$. The calculations involved appear in Table II.

State	Alternative	Test Quantity
i	k	$q_i^k + \frac{1}{\bar{\tau}_i^k}\left[\sum_{j=1}^{N} p_{ij}^k v_j - v_j\right]$
1	1	$6+5(-1) = 1$
	2	$4+2(-1) = 2$
2	1	$-3+4(1) = 1$
	2	$-5+7(1) = 2$

Table II. Policy Improvement for Machine Repair Example

We see that the second alternative in each state now has the higher value of the test quantity. Therefore the policy is changed to

$$\underline{d} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}. \tag{47}$$

We now solve the policy evaluation equation for this policy both to assure ourselves that the gain has, in fact, increased and to provide the relative values v_i required for possible further improvement. The equations are

$$\begin{aligned} v_1 + 1/2\, g &= 2 + v_2 \\ v_2 + 1/7\, g &= -5/7 + v_1 . \end{aligned} \tag{48}$$

With $v_2 = 0$ we obtain the solution $g = 2$, $v_1 = 1$. The gain has doubled to 2 per unit time. If we now attempt policy improvement we see that we are going to repeat the calculations in Table II because v_1 and v_2 have the same values as before. Therefore the policy found by the second alternative in each state has the highest gain of any policy in the system; the iteration has converged.

The management of the machine facility should therefore find it most profitable in the long run to use expensive maintenance and the outside repair crew. Furthermore it has learned something from the observation that $v_1 - v_2 = 1$. The difference in total expected reward over an infinite time period due to starting in one state rather than the other.

The proof that the iteration cycle works in the way we have described is very similar to the proof for the strictly Markov case[1]. Suppose that we have evaluated a policy A and then used the policy improvement procedure to obtain a new policy B. How are the gains of policies A and B related? Let us use a superscript A or B to indicate the quantities corresponding to each policy. The gains g^B and g^A are then found by solving the policy evaluation equations for each policy,

$$v_i^B + g^B \bar{\tau}_i^B = q_i^B \bar{\tau}_i^B + \sum_{j=1}^{N} p_{ij}^B v_i^B \qquad i = 1, 2, \ldots, N \tag{49}$$

and

$$v_i^B + g^A \bar{\tau}_i^A = q_i^A \bar{\tau}_i^A + \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad i = 1, 2, \ldots, N \tag{50}$$

Now we divide Equation 49 by $\bar{\tau}_i^B$ and Equation 50 by $\bar{\tau}_i^A$ to obtain

$$\frac{v_i^B}{\bar{\tau}_i^B} + g^B = q_i^B + \frac{1}{\bar{\tau}_i^B} \sum_{j=1}^{N} p_{ij}^B v_j^B \qquad i = 1, 2, \ldots, N \tag{51}$$

$$\frac{v_i^A}{\bar{\tau}_i^A} + g^A = q_i^A + \frac{1}{\bar{\tau}_i^A} \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad i = 1, 2, \ldots, N \tag{52}$$

By substracting Equation 52 from Equation 51 we form an expression involving the difference in gains for the two policies, $g^B - g^A$:

$$\frac{v_i^B}{\bar{\tau}_i^B} - \frac{v_i^A}{\bar{\tau}_i^A} + g^B - g^A = q_i^B - q_i^A + \frac{1}{\bar{\tau}_i^B}\sum_{j=1}^{N} p_{ij}^B v_j^B - \frac{1}{\bar{\tau}_i^A}\sum_{j=1}^{N} p_{ij}^A v_j^A$$
$$i = 1, 2, \ldots, N \qquad (53)$$

Because policy B was generated by using the values from policy A, we know from the policy improvement procedure that

$$q_i^B + \frac{1}{\bar{\tau}_i^B}\left[\sum_{j=1}^{N} p_{ij}^B v_j^A - v_i^A\right] \geq q_i^A + \frac{1}{\bar{\tau}_i^A}\left[\sum_{j=1}^{N} p_{ij}^A v_j^A - v_i^A\right]$$
$$i = 1, 2, \ldots, N \qquad (54)$$

Let γ_i be the result of subtracting the right side of equation 54 from the left side,

$$\gamma_i = q_i^B - q_i^A + \frac{1}{\bar{\tau}_i^B}\left[\sum_{j=1}^{N} p_{ij}^B v_j^A - v_i^A\right] - \frac{1}{\bar{\tau}_i^A}\left[\sum_{j=1}^{N} p_{ij}^A v_j^A - v_i^A\right] \geq 0\,. \qquad (55)$$

If we subtract Equation 55 from Equation 53, we obtain

$$\frac{v_i^B}{\bar{\tau}_i^B} - \frac{v_i^A}{\bar{\tau}_i^A} + g^B - g^A = \gamma_i - \frac{1}{\bar{\tau}_i^B}\left[\sum_{j=1}^{N} p_{ij}^B v_j^A - v_i^A\right] + \frac{1}{\bar{\tau}_i^A}\left[\sum_{j=1}^{N} p_{ij}^A v_j^A - v_i^A\right]$$
$$+ \frac{1}{\bar{\tau}_i^B}\sum_{j=1}^{N} p_{ij}^B v_j^B - \frac{1}{\bar{\tau}_i^A}\sum_{j=1}^{N} p_{ij}^A v_j^A$$
$$i = 1, 2, \ldots, N \qquad (56)$$

or

$$\frac{1}{\bar{\tau}_i^B}\left(v_i^B - v_i^A\right) + g^B - g^A = \gamma_i + \frac{1}{\bar{\tau}_i^B}\sum_{j=1}^{N} p_{ij}^B\left(v_j^B - v_j^A\right)$$
$$i = 1, 2, \ldots, N \qquad (57)$$

Now if we multiply through by $\bar{\tau}_i^B$ and use the notation that $x^A = x^B - x^A$ for any quantity x, then

$$v_i^A + g^A \bar{\tau}_i^B = \gamma_i \bar{\tau}_i + \sum_{j=1}^{N} p_{ij}^B v_j^A \qquad i = 1, 2, \ldots, N \qquad (58)$$

Now we realize that Equation 58 is of exactly the same form as the policy evaluation equations. The only difference is that the earning rates have been replaced by the increases in the test quantity, the γ_i's. Therefore by using the results of Equation 41 we

know that

$$g^A = \sum_{i=1}^{N} \phi_i^B \gamma_i \,. \tag{60}$$

The increase in gain from policy A to policy B is the sum of the increases in the test quantity that have been made in the course of policy improvement weighted by the limiting interval transition probability of the process under policy B. We have immediately that $g^A \geq 0$ and that it is strictly greater than zero if an increase in the test quantity has been made in any recurrent state under policy B.

We can easily show that it is impossible for a policy B with a higher gain than policy A to exist and be undiscovered by the policy improvement procedure. Suppose that such a policy exists so that $g^A > 0$. Suppose further that the policy iteration procedure has converged on policy A; then in all states $\gamma_i \leq 0$. However, since $\phi_i^B \geq 0$ for all i, Equation 60 shows that $g^A > 0$. We have therefore reached a contradiction of our assumption that $g^A > 0$ and have consequently shown that such a situation cannot arise. The policy iteration procedure must converge on a policy having the highest gain possible. Sometimes we may encounter situations where other policies have an equally high gain, but the tied policies are easily found from the tied alternatives in the policy improvement procedure in the final iteration.

Discounting

In some applications whose span covers large periods of time we must account for the fact that a dollar received in the future has a lower value than a dollar received today. We accomplish this by discounting any payment in the future to some present value. For our purposes we shall treat the case where discounting is exponential at a rate $\alpha > 0$; that is, payment received at a time t in the future is considered to be worth $e^{-\alpha t}$ as much now. It follows that the present value of any stream, $f(t)$, of payments stretching out into the future has a present value equal to the exponential transform of $f(t)$ evaluated at the point $\alpha: f^T(\alpha)$. Calculating this quantity is seldom trivial, but it can usually be adequately approximated.

The most important effect of discounting on our problem is avoidance of the linearly increasing nature of the expected total reward under a given policy, $v_i(t)$. When discounting is used the present value of the stream of payments expected from

the process in an infinite time is finite for any starting state i. We shall now proceed to calculate the present values for a process operating under a given policy.

The equations that determine the present values of the discounted process have the same probabilistic structure and rewards that appear in Equation 30. The only difference is that each of the payments the process generates must be discounted to its beginning. We shall use $v_i(t)$ now to mean the present value of the expected rewards the system will generate in time t if it started in state i at time zero. We can then write

$$v_i(t) = {}^{cc}w_i(t)\left\{e^{-at}v_i(0) + y_i\frac{1}{a}\left[1-e^{-at}\right]\right\} + \sum_{j=1}^{N} p_{ij}\int_0^t d\tau h_{ij}(\tau)\left\{r_{ij}e^{-a\tau} + y_i\frac{1}{a}\left[1-e^{-a\tau}\right] + e^{-a\tau}v_j(t-\tau)\right\}$$
$$i = 1, 2, \ldots, N;$$
$$t \geq 0. \qquad (61)$$

All lump payments or expectations are simply multiplied by e^{-a} raised to the time at which they occur. The only reward not receiving this treatment is that generated by the yield rate, y_i. For example, this reward is paid continuously at the rate y_i throughout the interval $(0, t)$ if the system does not change state. The present value of such a stream we easily find to be $y_i\frac{1}{a}\left[1-e^{-at}\right]$ from its exponential transform. The term within the integral arises from the rate y_i's being paid over the interval $(0, \tau)$.

Equation 61 could now be used to find the present value of allowing the system to operate for any time period starting in any state. However, we are once more interested in the behavior of the system if it is allowed to operate for a very long time. We shall use v_i for the present value of infinite time operation starting in state i and find the equation for this quantity by observing the form of Equation 61 when t is very large. We see that the entire first term vanishes in this case and write

$$v_i = v_i(\infty) = \lim_{t\to\infty}\sum_{j=1}^{N} p_{ij}\int_0^t d\tau h_{ij}(\tau)\left\{r_{ij}e^{-a\tau} + y_i\frac{1}{a}\left[1-e^{-a\tau}\right] + e^{-a\tau}v_j(t-\tau)\right\}$$
$$i = 1, 2, \ldots, N \qquad (62)$$

or

$$v_i = \sum_{j=1}^{N} p_{ij} r_{ij} h_{ij}{}^T(a) + y_i \frac{1}{a}\left[1 - w_i^T(\alpha)\right] + \sum_{j=1}^{N} p_{ij} h_{ij}{}^T(a) v_j \qquad i = 1, 2, \ldots, N \tag{63}$$

We can simplify Equation 63 by defining a quantity $\rho_i(a)$ as the present value of the expected reward that will be obtained while the system occupies state i and upon its departure from state i; thus,

$$\rho_i(a) = \sum_{j=1}^{N} p_{ij} r_{ij} h_{ij}{}^T(a) + y_i \frac{1}{a}\left[1 - w_i^T(a)\right] \qquad i = 1, 2, \ldots, N; \quad a > 0 \tag{64}$$

Now Equation 63 becomes

$$v_i = \rho_i(a) + \sum_{j=1}^{N} p_{ij} h_{ij}{}^T(a) v_j \qquad i = 1, 2, \ldots, N \tag{65}$$

We can solve Equation 65 uniquely for the present value of starting in each state of the system.

We gain insight into Equation 65 by writing it in matrix form. Let $\underline{v}$ and $\underline{\rho}(a)$ be column vectors of v_i and $\rho_i(a)$. The Equation 65 can be written

$$\underline{v} = \underline{\rho}(a) + \left[P \square H^T(a)\right] \underline{v} \tag{66}$$

or

$$\underline{v} = \left[I - P \square H^T(a)\right]^{-1} \underline{\rho}(a) . \tag{67}$$

Since $P \square H^T(a)$ has all its eigenvalues strictly less than one in magnitude, we can write $[I - P \square H^T(a)]^{-1}$ in the form

$$\left[I - P \square H^T(a)\right]^{-1} = \sum_{m=0}^{\infty} (P \square H^T(a))^m = C \tag{68}$$

Because all the elements of $P \square H^T(a)$ are positive, Equation 68 shows that all elements of $C = [I - P \square H^T(a)]^{-1}$ are also positive. We shall use this property later.

The decision problem in this process is to find the policy that maximizes the present value of all states. Fortunately no conflict between the states arises in fulfilling this goal; that is, it is not possible to increase the present value of one state at the

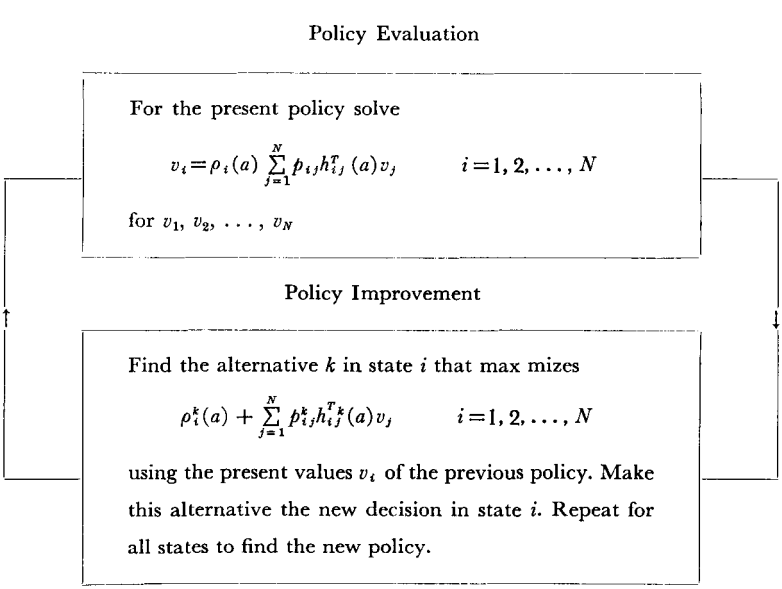

Fig. 2 The Iteration Cycle with Discounting

expense of another. An iteration scheme of the same type we described previously accomplishes the job. It is shown in Fig. 2. We select an arbitrary policy and evaluate its present values by using Equation 65. Then we improve the policy by finding the alternative k in each state that maximizes the test quantity

$$\rho_i^k(a)+\sum_{j=1}^{N} p_{ij}^k h_{ij}^T(a)v_j \tag{69}$$

sing the present values of the original policy. We can see immediately that the test quantity is plausible. When the procedure has been performed for all the states, we have a new policy that we proceed to evaluate, etc. Once more the process terminates when the same policy has been found on two successive iterations.

We shall now prove that the iteration cycle works by following an argument very similar to the one used in the case without discounting. Suppose that policy A has been evaluated and has produced a successor policy B in the policy Improvement procedure. The present values for the policies A and B must each satisfy Equation 65,

$$v_i^B = \rho_i^B(a) + \sum_{j=1}^{N} p_{ij}^B h_{ij}^{TB}(a) v_j^B \qquad i = 1, 2, \ldots, N \tag{70}$$

$$v_j^A = \rho_i^A(a) + \sum_{j=1}^{N} p_{ij}^A h_{ij}^{TA}(a) v_j^A \qquad i = 1, 2, \ldots, N \tag{71}$$

If we subtract Equation 71 from Equation 70 we obtain

$$v_i^B - v_i^A = \rho_i^B(a) - \rho_i^A(a) + \sum_{j=1}^{N} p_{ij}^{B} h_{ij}^{TB}(a) v_j^B - \sum_{j=1}^{N} p_{ij}^A h_{ij}^{TA}(a) v_j^A$$
$$i = 1, 2, \ldots, N \tag{72}$$

Since policy B was produced from policy A as the result of policy improvement,

$$\rho_i^B(a) + \sum_{j=1}^{N} p_{ij}^B h_{ij}^{TB}(a) v_j^A \geq \rho_i^A(a) + \sum_{j=1}^{N} p_{ij}^A h_{ij}^{TA}(a) v_j^A$$
$$i = 1, 2, \ldots, N \tag{73}$$

We define γ_i equal to the difference between the sides of Inequation 73,

$$\gamma_i = \rho_i^B(a) - \rho_i^A(a) + \sum_{j=1}^{N} p_{ij}^B h_{ij}^{TB}(a) v_j^A - \sum_{j=1}^{N} p_{ij}^A h_{ij}^{TA}(a) v_j^A$$
$$i = 1, 2, \ldots, N \tag{74}$$

If we now subtract Equation 74 from Equation 72 and write $v_i{}^A = v_i^B - v_i^A$, we find

$$v_i{}^A = \gamma_i + \sum_{j=1}^{N} p_{ij}^B h_{ij}^{TB}(a) v_j{}^A \qquad i = 1, 2, \ldots, N \tag{75}$$

Equation 75 has the same form as the policy evaluation equations except that the improvements in the test quantity γ_i have replaced the $\rho_i(a)$. Finally if we define column vectors $\underline{v}^A$ and $\underline{\gamma}$ with components $v_i{}^A$ and γ_i we can use Equations 67 and 68 to write Equation 75 as

$$\underline{v}^A = C\underline{\gamma} \tag{76}$$

Since we have already shown that the elements of C and $\underline{\gamma}$ are non-negative, the elements of $\underline{v}^A$ must be non-negative. The quantity $v_i{}^A$ will be greater than zero so that the present value of starting in state i will be greater under policy B than it was under policy A if it is possible to increase the test quantity in any state j that can be reached from state i under policy B. The argument that the iteration process cannot

stop until it has reached the highest possible present values for each state now follows by analogy with the non-discounting case.

Transient Behavior

In a large number of processes of practical importance we are more interested in the reward the process earns while passing through transient states than we are in the gain of the process. For example, we can often model terminal control systems with different possible trajectories by the kind of model we have been discussing with all states but the terminal state transient. In this type of situation the total reward earned by the process before it enters the recurrent state is the quantity of greatest interest. We can learn about the behavior of such a transient process by writing the policy evaluation equation 37 in the form

$$v_i = (q_i - g)\bar{\tau}_i + \sum_{j=1}^{N} p_{ij} v_j \qquad i = 1, 2, \ldots, N \tag{77}$$

We interpret this equation by saying that the value of being in state i is equal to the difference between the earning rate and the gain multiplied by the mean waiting time in the state plus the sum of the relative values of all states weighted by the probability of reaching each of them on the next transition. We can clarify the issues involved by assuming that all states but state N are transient and that state N is a trapping state. Since $P_{NN} = 1$, Equation 77 for $i = N$ produces immediately that

$$g = q_N \, . \tag{78}$$

If the system were started in state N then the total reward in time t would be simply $q_N t$. Therefore from Equation 35, v_N must be zero, and the other values are uniquely determined by Equation 77,

$$v_i = (q_i - g)\bar{\tau}_i + \sum_{j=1}^{N-1} p_{ij} v_j \qquad i = 1, 2, \ldots, N-1 \tag{79}$$

The first term on the left side of this equation represents the contribution to the expected value of state i due to the present occupancy of this state; the second term represents the expectation of future profit on later transitions. Let $\underline{e}$ be an $N-1$ element column vector with components $\{(q_i - g)\bar{\tau}_i)\}$ and let P^* be the $N-1 \times N-1$ matrix with elements $\{p_{ij}: i = 1, 2, \ldots, N-1; j = 1, 2, \ldots, N-1\}$.

Then Equation 79 becomes

$$\underset{\sim}{v} = \underset{\sim}{e} + P^* \underset{\sim}{v} \tag{80}$$

or

$$\underset{\sim}{v} = [I - P^*]^{-1} \underset{\sim}{e} \tag{81}$$

The inverse matrix $[I - P^*]^{-1}$ always exists for a transient process of the form we have defined. Let

$$\bar{N} = \{\bar{\nu}_{ij}\} = [I - P^*]^{-1} \tag{82}$$

It is a well-known property of Markov processes that the quantity $\bar{\nu}_{ij}$ is the mean number of times the system will occupy state j in an infinite number of transitions. Then

$$\underset{\sim}{v} = \bar{N} \underset{\sim}{e} \tag{83}$$

or

$$v_i = \sum_{j=1}^{N-1} \bar{\nu}_{ij}(q_j - g)\bar{\tau}_j \tag{84}$$

$$= \sum_{j=1}^{N-1} \bar{\nu}_{ij} q_j \bar{\tau}_j - g \sum_{j=1}^{N-1} \bar{\nu}_{ij} \bar{\tau}_j \tag{85}$$

Equation 85 says that the value of state i is equal to the sum of the expected earnings from a single occupancy of state j, $q_j \bar{\tau}_j$, multiplied by the expected number of times state j will be occupied if the system is started in state i less the gain of the system multiplied by the expected total time the system will spend in the transient process. The problem of maximizing the expected total reward from starting in a certain transient state in this process is therefore one of selecting the most favorable set of mean numbers of occupancies and earning rates rather than one of selecting the most favorable set of limiting interval transition probabilities and earning rates. We might therefore expect that a slightly different iteration procedure would be advisable.

To establish such a procedure we might examine Equation 77. It suggests that v_i could be maximized by a policy improvement procedure that maximized the test quantity

$$(q_i^k - g)\bar{\tau}_i^k + \sum_{j=1}^{N} p_{ij}^k v_j \qquad i = 1, 2, \ldots, N \tag{86}$$

with respect to all alternatives k in state i. Note that this test quantity involves not only the relative values of state i under the previous policy, but also the gain of that policy. Let us use the same type of proof we used before to find the properties of this test criterion.

If policy B has been produced from policy A by using this test criterion in a policy improvement, then we can write Equations 49 and 50 for each policy individually and subtract them. We obtain

$$v_i^B - v_i^A + g^B \bar{\tau}_i^B - g^A \bar{\tau}_i^A = q_i^B \bar{\tau}_i^B - q_i^A \bar{\tau}_i^A + \sum_{j=1}^{N} p_{ij}^B v_j^B - \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad i = 1, 2, \ldots, N \tag{87}$$

From the properties of the policy improvement,

$$(q_i^B - g^A)\,\bar{\tau}_i^B + \sum_{j=1}^{N} p_{ij}^B v_j^A \geq (q_i^A - g^A)\,\bar{\tau}_i^A + \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad i = 1, 2, \ldots, N \tag{88}$$

Let ζ_i equal the difference between the left and right sides of this equation,

$$\zeta_i = (q_i^B - g^A)\,\bar{\tau}_i^B + \sum_{j=1}^{N} p_{ij}^B v_j^A - (q_i^A - g^A)\,\bar{\tau}_i^A - \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad i = 1, 2, \ldots, N \tag{89}$$

If we subtract Equation 89 from Equation 87, we find

$$v_i^B - v_i^A + (g^B - g^A)\bar{\tau}_i^B = \zeta_i + \sum_{j=1}^{N} p_{ij}^B (v_j^B - v_j^A) \qquad i = 1, 2, \ldots, N \tag{90}$$

or

$$v_i^A + g^A \bar{\tau}_i^B = \zeta_i + \sum_{j=1}^{N} p_{ij}^B v_j^A \qquad i = 1, 2, \ldots, N \tag{91}$$

If we are dealing with a system in which only state N is recurrent and if the gain of that state is fixed, the $g^A = 0$ and Equation 91 has the same form as Equation 79 with $g = 0$. Therefore by Equation 85.

$$v_i^A = \sum_{j=1}^{N-1} \bar{\nu}_{ij}^B \zeta_j \qquad i = 1, 2, \ldots, N-1 \tag{92}$$

The change in values for each state is given by the sum of the mean number of transitions made to each state under the policy multiplied by the increases in the test quantity. Since both of the factors are non-negative, all the v_i^A are non-negative. This iteration procedure will therefore converge on the policy with the largest values for the same reasons that were discussed in the case with discounting.

We have therefore found a test procedure that will maximize the value of the transient states. We might ask about the effect that using this procedure in all problems would have upon the gain of a process that did not have this transient structure. Equations 91 still apply. We see that they are identical to Equation 58 if

$$\zeta_i = \gamma_i \bar{\tau}_i^B, \tag{93}$$

an identity that is easily established. We can therefore write

$$g^A = \sum_{j=1}^{N} \phi_i^B \frac{\zeta_i}{\bar{\tau}_i^B} \tag{94}$$

Equation 94 shows that the increase in gain is not equal to the increases in the test quantities multiplied by the interval transition probabilities of policy B, but rather a division by $\bar{\tau}_i^B$ is involved. Nevertheless all the arguments about the necessity for the iteration procedure to increase the gain and ultimately to converge on the policy of highest gain apply equally well to this criterion. We do not yet know whether this criterion will increase the gain as quickly as the original criterion, but we do know that it will work and that it will maximize the values of a transient process.

Yet there is a disturbing thought—how do we know that the original test quantity, which we shall call the "gain-oriented" criterion, does not miximize the values of a transient process just like the new test quantity, which we shall call the "value-oriented" criterion. Let us check. Equation 58 shows that if $g^A = 0$, then the v_i^A must satisfy Equation 92 after the substitution of ζ_i from Equation 93; that is,

$$v_i^A = \sum_{j=1}^{N} \bar{\nu}_{ij}^B \gamma_j \bar{\tau}_j^B \tag{95}$$

We see that the gain-oriented criterion involves not only the mean number of transitions to each state and the increases in test quantity but also the mean waiting times in each state. Yet here again all the arguments that show how Equation 92 implies

a valid iteration procedure still apply. In other words, the gain-oriented criterion will also maximize the values of transient states.

We have now see that both the gain-oriented and value-oriented test quantities will accomplish both task we require. The matter of their relative efficiency is still open to question. However, it is easy to show that their efficiencies do differ by considering two examples. The first example is a one state system with three alternatives in the state. The system can make only virtual transitions; the alternatives specify the earning rates and mean waiting times. The data are

$$\begin{array}{lll} q_1^1 = 3 & q_1^2 = -2 & q_1^3 = 1 \\ \bar{\tau}_1^1 = 1 & \bar{\tau}_1^2 = 3 & \bar{\tau}_1^3 = 8 \end{array} \tag{96}$$

This trivial system has a trivial solution. Only the gain is involved and from Equation 78 we know that it is just equal to the earning rate of the state. Since alternative 1 produces the highest earning rate, 3, it is clear that this is the alternative on which we want any iteration scheme to converge. Suppose we start with a policy that specifies alternative 3, $\underline{d} = [3]$. How will solving the problem using the gain and value-oriented criteria affect the number of iterations required for convergence? We shall begin with the gain-oriented criterion. We already know that for the initial policy $g = 1$, $v_i = 0$. Therefore, the policy improvement procedure of expression 43 becomes simply $\underset{k}{\text{Max}}\, q_1^k$. We see that this leads immediately to the policy $\underline{d} = [1]$, and that the procedure has converged to give a gain $g = 3$ in one iteration.

Now we shall solve the same problem using the value-oriented criterion of expression 86. We see that it becomes $\underset{k}{\text{Max}}\, (q_1^k - g)\bar{\tau}_1^k$. The test quantities for the three alternatives using the gain of the original policy are then 2, 3, and 0. Therefore, the second alternative is best and we have a policy $\underline{d} = [2]$ with gain 2. Now we repeat the policy improvement again and obtain test quantities 1, 0, and -8. Now the first alternative is best and we have found the policy $\underline{d} = 1$ with gain 3. Further attempts at improvement lead to the same policy.

We have just seen how in a particular example involving only gains the gain-oriented test quantity was able to find the optimum policy in one less iteration. But is it always better for any problem? No, as we shall now see in a second example with two states. Let state 2 be a trapping state with gain zero and let state 1 be a transient

state that must enter state 2 on its first transition. State 1 has three alternative ways of going to state 2 described by the three alternatives in display 96. Since the process can earn money only while it occupies state 1 and since the expected amount it will earn doing this is $q_i\bar{\tau}_i$, we see that it will earn the most, 8 by following alternative 3. In this problem we find quickly that $g = 0$, $v_2 = 0$, $v_1 = q_1\bar{\tau}_1$ so that solving the policy evaluation equation is no problem. Let us choose as our intial policy the first alternative in state 1, and indicate this choice simply by $\underline{d} = [1]$.

Now we have to choose a test quantity. Let us begin this time with the value-oriented criterion of expression 86. Therefore we find immediately that the third alternative is the best and we have converged on the optimum policy $\underline{d} = [3]$ with a value of 8 in only one iteration.

The gain-oriented criterion of Equation 43 reduces in this problem to the form

$$\max_k \; q_1^k - \frac{v_1}{\bar{\tau}_1^k}$$

For the original policy, $v_1 = 3$ and so we have for the three test quantities 0, 1, and 5/8. Therefore alternative 2 is the best and we change to policy $\underline{d} = [2]$ for which $v_1 = 6$. Then we compute the test quantities again and find -3, 0, and 1/4. Now alternative 3 is the best and we have found the optimum policy $\underline{d} = [3]$ with value $v_1 = 8$. Of course, further attempts at improvement lead to the same policy.

Now we have seen by example that the value-oriented criterion can save iterations in primarily value maximization problems while the gain-oriented criterion can save iterations in primarily gain maximization problems. We still have not resolved the question of when each criterion should be used, but we know that the answer has significance. It would seem advisable to consider plans like using the gain-oriented criterion until the gain converged and then switching over to the value-oriented criterion. However, more research on this computational question is needed.

Linear Programming

Decision problems in semi-Markov processes can be solved by using linear programming just as in the strictly Markovian case; it may or may not be advantageous computationally to use linear programming algorithms. We shall illustrate how the problems we have discussed can be formulated as linear programs.

The basic equations for the semi-Markovian reward process without discounting are Equations 37,

$$v_i + g\bar{\tau}_i = q_i\bar{\tau}_i + \sum_{j=1}^{N} p_{ij}v_j \qquad i = 1, 2, \ldots, N \tag{97}$$

By solving these equations for g we can write

$$g = q_i + \frac{1}{\bar{\tau}_i}\left(\sum_{j=1}^{N} p_{ij}v_j - v_i\right) \qquad i = 1, 2, \ldots, N \tag{98}$$

If the gain g and the values v_i are those pertinent to the optimal policy, then for any possibly non-optimal alternative k in state i,

$$g \geq q_i^k + \frac{1}{\bar{\tau}_i^k}\left(\sum_{j=1}^{N} p_{ij}^k v_j - v_i\right) \qquad i = 1, 2, \ldots, N \tag{99}$$

Let K_i be the number of alternatives in state i, and let $K = \sum_{i=1}^{N} K_i$ be the number of alternatives in all states. Then the gain g must satisfy all the inequalities

$$g \geq q_i^k + \frac{1}{\bar{\tau}_i^k}\left(\sum_{j=1}^{N} p_{ij}^k v_j - v_i\right) \qquad \begin{matrix} i = 1, 2, \ldots, N; \\ k = 1, 2, \ldots, K_i \end{matrix} \tag{100}$$

The gain we seek is the smallest number g that meets these requirements.

We can place this minimization problem in the form of a standard linear program by introducing matrix notation. Since $v_N = 0$ we can first write Equation 100 as

$$\frac{1}{\bar{\tau}_i^k}\sum_{j=1}^{N-1}(\delta_{ij} - p_{ij}^k)v_j + g \geq q_j^k \qquad \begin{matrix} i = 1, 2, \ldots, N; \\ k = 1, 2, \ldots, K_i \end{matrix} \tag{101}$$

Now we construct a $K \times N$ matrix B with elements b_{ij} defined by

$$\begin{aligned} &b_m(i,k),\ j = \frac{1}{\bar{\tau}_i^k}(\delta_{ij} - p_{ij}^k) && j = 1, 2, \ldots, N-1 \\ &b_m(i,k),\ N = 1 \end{aligned} \tag{102}$$

where

$$m(i,k) = k + K_0 + K_1 + \ldots + K_{i-1}\ ((k = 1, 2, \ldots, K_i)\ i = 1, 2, \ldots, N);\ K_0 = 0.$$

The matrix B is shown in Fig. 3. Then we define two N–element column vectors $\underline{v}$ and $\underline{t}$, and one K–element column vector $\underline{q}$. These vectors are defined by

$$\begin{aligned} &v_i = v_i, \qquad i = 1, 2, \ldots, N-1 \qquad && t_i = 0, \qquad i = 1, 2, \ldots, N-1 \\ &v_N = g && t_N = 1 \\ &q_m(i, k) = q_i^k \end{aligned} \tag{103}$$

and also shown in Fig. 3.

Now we can write the linear program as

$$\begin{aligned} &\text{Min } \underline{t}^{tr}\ \underline{v} \\ &\text{subject to } B\ \underline{v} \geq \underline{q} \\ &\qquad \underline{v} \text{ unconstrained in sign} \end{aligned} \tag{104}$$

Here the superscript *tr* means transpose. This linear program could now be solved by conventional methods.

The dual of this program is interesting in itself. We find immediately from linear programming duality theory that the dual is

$$\begin{aligned} &\text{Max } \underline{q}^{tr}\underline{\phi} \\ &\text{subject to } B^{tr}\underline{\phi} = \underline{t} \\ &\qquad \underline{\phi} \geq 0 \end{aligned} \tag{105}$$

The vector $\underline{\phi}$ is a K–element column vector. The constraint involving the matrix B is an equality constraint because the primal variables $\underline{v}$ are unconstrained in sign. The dual variables $\underline{\phi}$ are constrained in sign because the constraints involving B in the primal are inequalities. If we write the detailed equations implied by the matrix formulation 105, we find that the dual variables are the different possible limiting interval transition probabilities for the semi-Markov process. Formulation 105 is therefore a method for maximizing the right-hand side of Equation 41.

The decision process with discounting is also susceptible to linear programming. In this case the basic equations are Equations 65

$$v_i = \rho_i(a) + \sum_{j=1}^{N} p_{ij} h_{ij}^{T}(a) v_j \qquad i = 1, 2, \ldots, N \tag{106}$$

$$
B = \begin{pmatrix}
(1-p_{11}^1)\frac{1}{\bar{\tau}_1^1} & -p_{12}^1\frac{1}{\bar{\tau}_1^1} & -p_{13}^1\frac{1}{\bar{\tau}_1^1} & \cdots & -p_{1,N-1}^1\frac{1}{\bar{\tau}_1^1} & 1 \\
(1-p_{11}^2)\frac{1}{\bar{\tau}_1^2} & -p_{12}^2\frac{1}{\bar{\tau}_1^2} & -p_{13}^2\frac{1}{\bar{\tau}_1^2} & \cdots & -p_{1,N-1}^2\frac{1}{\bar{\tau}_1^2} & 1 \\
\vdots & & & & & \\
(1-p_{11}^{K_1})\frac{1}{\bar{\tau}_1^{K_1}} & -p_{12}^{K_1}\frac{1}{\bar{\tau}_1^{K_1}} & -p_{13}^{K_1}\frac{1}{\bar{\tau}_1^{K_1}} & \cdots & -p_{1,N-1}^{K_1}\frac{1}{\bar{\tau}_1^{K_1}} & 1 \\
-p_{21}^1\frac{1}{\bar{\tau}_2^1} & (1-p_{22}^1)\frac{1}{\bar{\tau}_2^1} & -p_{23}^1\frac{1}{\bar{\tau}_2^1} & \cdots & -p_{2,N-1}^1\frac{1}{\bar{\tau}_2^1} & 1 \\
-p_{21}^2\frac{1}{\bar{\tau}_2^2} & (1-p_{22}^2)\frac{1}{\bar{\tau}_2^2} & -p_{23}^2\frac{1}{\bar{\tau}_2^2} & \cdots & -p_{2,N-1}^2\frac{1}{\bar{\tau}_2^2} & 1 \\
\vdots & & & & & \\
-p_{21}^{K_2}\frac{1}{\bar{\tau}_2^{K_2}} & (-p_{22}^{K_2}\frac{1}{\bar{\tau}_2^{K_2}} & & \cdots & -p_{2,N-1}^{K_2}\frac{1}{\bar{\tau}_2^{K_2}} & 1 \\
-p_{N1}^1\frac{1}{\bar{\tau}_N^1} & -p_{N2}^1\frac{1}{\bar{\tau}_N^1} & & \cdots & -p_{N,N-1}^1\frac{1}{\bar{\tau}_N^1} & 1 \\
-p_{N1}^2\frac{1}{\bar{\tau}_N^2} & -p_{N2}^2\frac{1}{\bar{\tau}_N^2} & & \cdots & -p_{N,N-1}^2\frac{1}{\bar{\tau}_N^2} & 1 \\
\vdots & & & & & \\
-p_{N1}^{K_N}\frac{1}{\bar{\tau}_N^{K_N}} & -p_{N2}^{K_N}\frac{1}{\bar{\tau}_N^{K_N}} & & \cdots & -p_{N,N-1}^{K_N}\frac{1}{\bar{\tau}_N^{K_N}} & 1
\end{pmatrix}
\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix}
\begin{matrix} \uparrow \\ \\ \\ \downarrow \\ \uparrow \\ \\ \\ \downarrow \\ \uparrow \\ \\ \\ \downarrow \end{matrix}
\begin{matrix} \uparrow \\ \\ K_1 \\ \downarrow \\ \uparrow \\ \\ K_2 \\ \downarrow \\ \uparrow \\ \\ K_N \\ \downarrow \end{matrix}
\quad K
$$

$$
\longleftarrow \qquad N \qquad \longrightarrow
$$

Fig. 3 Matrix Definitions for Linear Programming Formulation

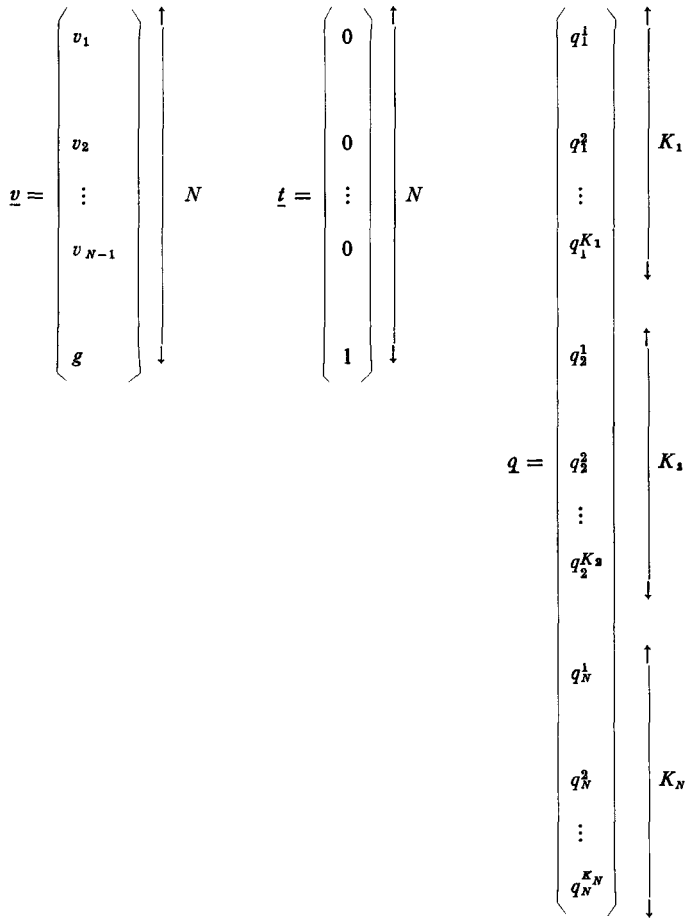

Fig. 3 Matrix Definitions for Linear Programming Formulation

We now want to maximize the present values v_i. By the same argument that we followed before we know that the v_i's we seek are the smallest v_i's satisfying.

$$v_i \geq \rho_i^k(a) + \sum_{j=1}^{N} p_{ij}^k h_{ij}^{T^k}(a) v_j \qquad i = 1, 2, \ldots, N \tag{106}$$

or

$$\sum_{j=1}^{N} \left[\delta_{ij} - p_{ij}^k h_{ij}^{T^k}(a) \right] v_j \geq \rho_i^k(a) \tag{108}$$

We construct the matrix formulation by defining a matrix D whose elements are

$$d_m(i, k), j = \delta_{ij} - p_{ij}^k h_{ij}^{T^k}(a) \qquad j = 1, 2, \ldots, N \tag{109}$$

We also define two N-element vectors $\underline{v}$ and $\underline{s}$ and one K-element vector $\underline{\rho}(a)$ with elements

$$\begin{aligned} &v_i = v_i, \quad i = 1, 2, \ldots, N; \quad s_i = 1, \quad i = 1, 2, \ldots, N \\ &\rho_{m(i,k)}(a) = \rho_i^k(a) \end{aligned} \tag{110}$$

These new definitions may be easily visualized in the format of Fig. 3. Since maximizing the sum of the present values is equivalent to maximizing each of them individually in our problem, the linear program is

$$\begin{aligned} &\text{Min } \underline{s}^{tr}\ \underline{v} \\ &\text{subject to } D\ \underline{v} \geq \underline{\rho}(a) \\ &\qquad \underline{v} \text{ unconstrained in sign} \end{aligned} \tag{111}$$

Once more the dual of this problem is significant. Let $\underline{c}$ be a K-element column vector. Then the dual of this linear program is

$$\begin{aligned} &\text{Max } \underline{\rho}^{tr}(\alpha)\ \underline{c} \\ &\text{subject to } D^{tr}\ \underline{c} = \underline{s} \\ &\qquad \underline{c} \geq 0 \end{aligned} \tag{112}$$

The detailed equations implied by formulation 112 show that the components of $\underline{c}$ are just the discounted number of times we expected each state to be occupied in an infinite number of transitions. Therefore this formulation is equivalent to maximizing

the set of present values as expressed by Equation 67. Once more we see that the dual formulation has an important interpretation.

Now that we have seen a linear programming formulation of the problems we have been discussing it is worth mentioning that the dual is computationally much more convenient than the primal. The primal has $2N+K$ variables because the constraints are inequalities and the variables are unconstrained in sign; it has K equations. The dual has K variables and N equations. Since K is usually much larger than N, we find it more convenient to solve the dual formulation in almost all cases.

Summary

Decision models of the type we have discussed are interesting mathematical developments, but they are useful in practical problems only if we can supply them with necessary data. Consequently, we are investigating schemes for deriving this data from both experience and experiment. We are also investigating how to solve the decision process when the data are uncertain. Although the practicality of these methods will be considerably enhanced by the completion of this research, the model as it stands continues to have important application in several areas of management systems.

REFERENCES

1. Howard, R. A., Dynamic Programming and Markov Processes, Technology Press-Wiley, Cambridge, 1960.
2. Levy, Paul, "Systems Semi-Markovian a au plus une infinite denombrable d'etats possibles, "Proc. Int. Congr. Math., Amsterdam, Vol. 2 (1954), p. 294. "Processes Semi-Markovian," ibid., Vol. 3 (1954), pp. 416–426.
3. Smith, W. L., "Regenerative stochastic processes, "Proc. Roy. Soc. (London), Ser. A, Vol. 232 (1955), pp. 6–31 (cf. the abstract of this paper in Proc. Int. Congr. Math., Amsterdam, Vol. 2 (1954), pp. 304–305.
4. Smith, W. L., "Renewal theory and its ramifications," J. Roy. Stat. Soc., Ser. B, Vol. 20 (1958), pp. 243–302.
5. Pyke, R., "Markov Renewal Processes with Finitely Many States," Annals, Math. Stati., 32: 1243–59.
6. Howard, R. A., "Semi-Markovian Decision Processes," Proceedings of the 34th Session of the International Statistical Institute, Ottawa, Canada August 21–29, 1963.