The Optimal Control of Partially Observable Markov Processes over a Finite Horizon

Richard D. Smallwood

Stanford University, Stanford, California, and Xerox Palo Alto Research Center, Palo Alto, California

and

Edward J. Sondik

Stanford University, Stanford, California

(Received October 27, 1971)

This paper formulates the optimal control problem for a class of mathematical models in which the system to be controlled is characterized by a finite-state discrete-time Markov process. The states of this internal process are not directly observable by the controller; rather, he has available a set of observable outputs that are only probabilistically related to the internal state of the system. The formulation is illustrated by a simple machine-maintenance example, and other specific application areas are also discussed. The paper demonstrates that, if there are only a finite number of control intervals remaining, then the optimal payoff function is a piecewise-linear, convex function of the current state probabilities of the internal Markov process. In addition, an algorithm for utilizing this property to calculate the optimal control policy and payoff function for any finite horizon is outlined. These results are illustrated by a numerical example for the machine-maintenance problem.

THE TWO CONCEPTS of state and state transition are essential to the modeling of complex dynamic systems. The concept of state allows one to focus on the features of the system that are essential to the problem at hand, while the concept of state transition provides the mechanism for structuring the system's dynamic behavior. In most situations, there is an element of uncertainty in the transitions of the process from one state to another, and this leads naturally to the use of Markov processes as quantitative models of the system.

Unfortunately, in many practical applications we are not permitted exact observation of the state of the process. For example, there are many situations in medicine in which we would like to model the dynamics of the patient's physiological state as a Markov process, but this state is not directly observable. In such cases, we can often model what is observable as probabilistically related to the true state of the system. Figure 1 presents a pictorial representation of such a model, termed a partially observable Markov process. In this paper we shall consider partially observable Markov processes for which the underlying Markov process is a discrete-time finite-state Markov process; in addition, we shall limit the discussion to processes for which the number of possible outputs at each observation is finite.

As an example of such a system, consider a hypothetical manufacturing operation that produces a finished product once an hour at the end of each hour. This machine consists of two identical internal components, each of which must operate once upon the product before it is finished. Unfortunately, each component can fail spontaneously and, if a component has failed, there is some probability that, when operating upon the product, it will cause the product to be defective. For the sake of simplicity, let us assume that the finished product is either not defective or defective with a corresponding profit of one unit or zero units, respectively. If the machine must be disassembled in order to examine the status of the internal components, then its internal state is not directly observable, and so Fig. 1 is a valid representation of the process.

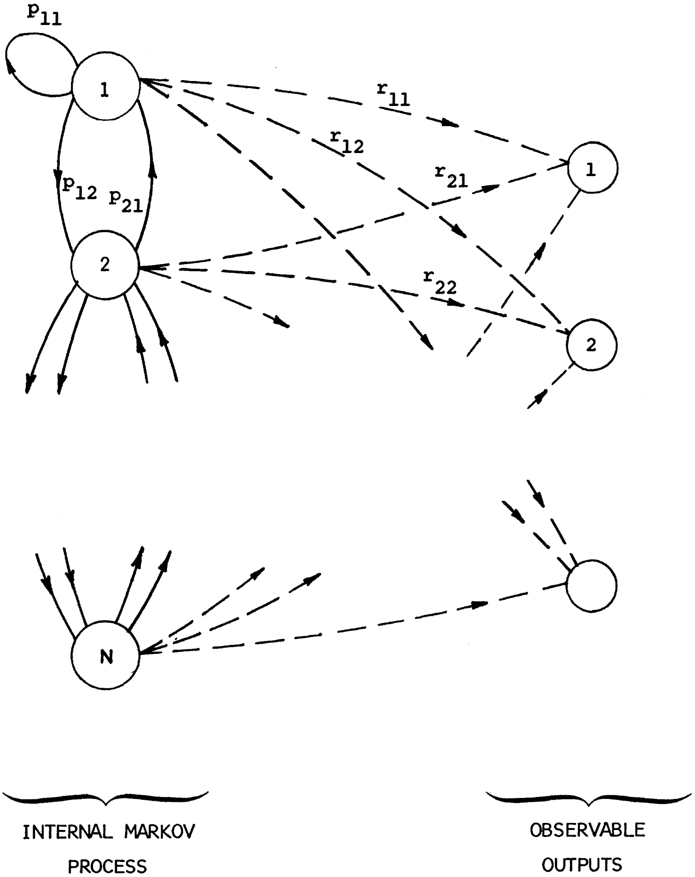


Fig. 1. The partially observable Markov process.

Suppose that there are several control options available to us during each one-hour production interval. For the simplest alternative, we simply continue the manufacturing process with no examination of the finished product. A second alternative is to examine the quality of the product as it rolls off the production line at the end of the hour. For the third alternative, we stop the machine for the one-hour interval, disassemble it, inspect the internal components and replace any that have failed. Finally, the fourth alternative uses the hour to replace both components without inspecting them first. For this manufacturing example, we wish to know which of these control alternatives is optimal for each possible history of the machine's operation.

This example illustrates the characteristics of the general optimal control problem for partially observable Markov processes. This paper formulates and solves this general optimal control problem for a process that is to operate for only a finite number of time periods. A later paper will examine this control problem for a process that is to operate into the indefinite future.

ECKLES[2] has considered the control of partially observable Markov processes as applied to machine replacement problems similar to the one in the preceding example. In a second area, the partially observable Markov process has also been applied to the human learning process (SMALLWOOD[6]); in this application, the unobservable states of the Markov process correspond to the state of knowledge of the student, the observable outputs represent the discrete responses by the student to particular questions about the subject matter, and the alternative controls represent various mechanisms for presenting material to the student. MATHESON[4] and Smallwood[7,8] have considered the optimal control problem for simple examples of such learning models.

A third application of partially observable Markov processes has been in the decoding of Markov sources transmitting over a noisy channel (DRAKE[1]). In this case, the internal state of the Markov process corresponds to the state of the Markov source and the discrete observations represent the outputs of the noisy channel. A fourth application area for this class of model is in medical diagnosis and decision-making. In this case, the unobservable state of the process corresponds to the physiological status of the patient, the discrete observations represent the results of diagnostic tests or the patient's response to particular therapies, and the control alternatives correspond to different tests or therapies available to the patient.

A fifth application area is in the search for a moving object. POLLOCK[5] has formulated a two-state moving-target model that can be converted very simply into a three-state partially observable Markov process, the additional state corresponding to detection of the target. For the general application of partially observable Markov processes to the search for a moving object, the unobservable states of the internal process represent the status of the target object, the discrete observations correspond to the outcome of some expenditure of search resources (e.g., 'found it' or 'did not find it'), and the control alternatives represent particular feasible expenditures of the search resources (e.g., 'look in location i'). Table I summarizes these five application areas.

I. PROPERTIES OF THE MODEL

TO BEGIN THE explicit formulation of the control problem for a partially observable Markov process, we assume that the internal dynamics of the system under control can be modeled by an N-state discrete-time Markov process. If there are n control periods remaining, the problem is to select the alternative from the available set

TABLE I

APPLICATION AREAS

Application area	State	Observation	Control alternative
Machine maintenance and replacement	Status of machine	Quality of product or result of internal inspection	Examination of output or internal examination
Human learning and instruction	Status of knowledge	Response to query	Alternative presentations of material
Decoding of Markov sources	Status of source	Output of noisy channel	(Not considered)
Medical diagnosis and decision-making	Physiological status of patient	Outcome of test or response to therapy	Test or therapy
Search for moving object	Status of target object	Result of search	Expenditure of search resources

$A(n)$ that will optimize the performance of the system during its remaining lifetime. If alternative a is selected, then the conditional probability that the internal process will make its next transition to state j if it is presently in state i will be written as p^a_{ij}. An observation will follow each transition, with $r^a_{j\theta}$ denoting the probability of observing output θ if the new internal state of the process is j and alternative a is controlling the system. Figure 2 illustrates this sequence of events.

With this representation of the process, it is easy to see that, if $r^a_{j\theta}$ is independent of j, then the observation of the output will yield no additional information about the internal state of the process. This is the case of the nonobservable Markov process. The other extreme is the more usual case that has been studied extensively in the literature (HOWARD[3]). If there is one output for each internal state of the process and if for each alternative $r^a_{j\theta}=1$ if and only if $j=\theta$, then the process is said to be completely observable.

The calculation of an optimal control policy requires a reward structure for the process. Thus, we define $w^a_{ij\theta}$ as the immediate award accrued if, while under the control of alternative a during one control interval, the process makes a transition from state i to j and then produces output θ. The analysis to follow assumes that the controller has no direct observation of the accrued rewards; that is, the controller only observes the outputs of the observation part of the process. If this assumption is violated, then it is easy to redefine the observation outputs of the process to include the rewards that are immediately available to the controller.

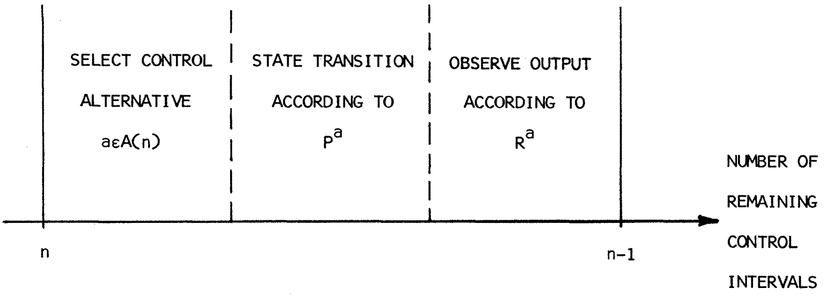


Fig. 2. The sequence of operations.

The uncertainties in the dynamics of the internal process and in the observation process produce uncertainty about the internal state of the system. For this formulation of the control problem, the current state of information about the internal state of the system can be encoded as the information vector $\pi = [\pi_1, \pi_2, \cdots, \pi_N]$, where π_i is the probability that the current internal state of the system is i. In other words, if the controller only has available to him his past observations of the process's outputs, then at any time the vector π is a sufficient statistic for his past sequence of observations. Appendix A presents a proof of this rather intuitive result.

From this result it follows that the dynamic behavior of the information vector π is itself a discrete-time continuous-state Markov process. This dynamic behavior of the state of information is crucial to the calculation of the optimal control. If our prior state of information about the internal state of the system is denoted by π, and if we observe the output θ after using alternative a, then we must be able to calculate our updated state of information. If π_j' is the updated probability that the internal state of the system is j given the new information, then the application of simple probability operations based on the sequence of events shown in Fig. 2 yields the following equation (Appendix A contains the complete derivation):

$$\pi_j' = [\textstyle\sum_i \pi_i p_{ij}^a r_{j\theta}^a] / [\textstyle\sum_{i,j} \pi_i p_{ij}^a r_{j\theta}^a]. \qquad (1)$$

Equation (1) defines a transformation from the vector π to the vector π'. Since this transformation plays an important role in the succeeding development, it is useful to introduce the notation

$$\pi' = T(\pi | a, \theta). \qquad (2)$$

Figure 3 illustrates some of the properties of this transformation for the three-state case. In this portrayal, the space of possible π vectors is represented by an equilateral triangle, with each point in the triangle corresponding to a possible state of the information vector π. For each information vector π, the perpendicular distance from the point to the side opposite the ith vertex is just equal to π_i. Thus, points closer to the ith vertex correspond to states of information in which the process is believed more likely to be in state i. The transformation in (2) then transforms a point in the space of information vectors for one time period into another point in the space of information vectors for the succeeding time period. Furthermore, as illustrated in Fig. 3, there will be one such transformation for each possible output of the observation process.

With this as a background, the remainder of this section will formulate and examine a dynamic-programming approach to calculating the optimal control policy for a partially observable Markov process. To this end, we define $V_n(\pi)$ as the maximum expected reward that the system can accrue during the lifetime of the process if the current information vector is π and there are n control intervals remaining before the process terminates. Then, expanding over all possible next transitions and observations yields the recursive equation

$$V_n(\pi) = \max_{a \epsilon A(n)} [\textstyle\sum_{i=1}^{i=N} \pi_i \sum_{j=1}^{j=N} p_{ij}^a \sum_\theta r_{j\theta}^a \{w_{ij\theta}^a + V_{n-1}[T(\pi | a, \theta)]\}]. \qquad (3)$$

This equation can be simplified somewhat by defining the expected immediate reward for state i if alternative a is used during the next control interval as

$$q_i^a = \textstyle\sum_{j,\theta} p_{ij}^a r_{j\theta}^a w_{ij\theta}^a. \qquad (4)$$

Equation (3) then becomes

$$V_n(\pi) = \max_{a \epsilon A(n)} \left[\sum_i \pi_i q_i^a + \sum_{i,j,\theta} \pi_i p_{ij}^a r_{j\theta}^a V_{n-1}[T(\pi|a,\theta)]\right]. \qquad (5)$$

Equation (5) is valid for $n \geqq 1$. It only remains to define the value of terminating the process in each internal state. If q_i^0 is the expected value of terminating the process in state i, then the expected terminal value for the process with a final information vector π is just

$$V_0(\pi) = \sum_i \pi_i q_i^0 = \pi \cdot q^0. \qquad (6)$$

It is instructive to rewrite (5) in matrix form. To this end, we define $\mathrm{pr}\{\theta|\pi, a\}$ as the probability of next observing output θ if the current information vector is π and the next alternative selected is a. With this definition, (5) becomes

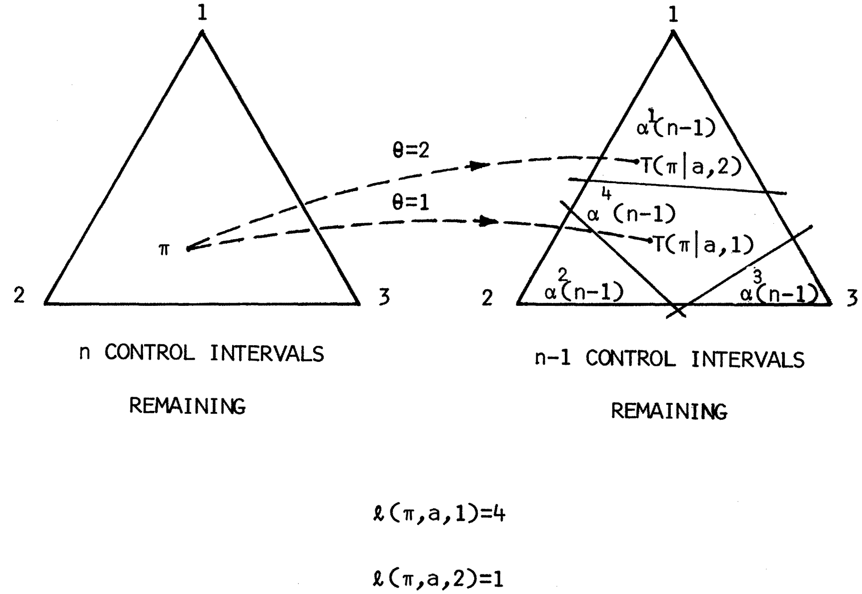


Fig. 3. The information-vector transformation $T(\pi|a,\theta)$ for the three-state case.

$$V_n(\pi) = \max_{a \epsilon A(n)} \left[\pi \cdot q^a + \sum_\theta \mathrm{pr}\{\theta|\pi, a\} V_{n-1}[T(\pi|a,\theta)]\right], \qquad (5')$$

valid for $n \geqq 1$. In this form, the equation represents a dynamic-programming problem over a continuous state space, the space of information vectors. This is consistent with the previous comment that the information vector π is itself the state of a discrete-time, continuous-state Markov process. Appendix A discusses this in more detail.

Although (5) [or equivalently (5′)] appears rather formidable, its solution has a rather simple form. In particular, we shall show that $V_n(\pi)$ is piecewise linear and convex, and can thus be written as

$$V_n(\pi) = \max_k \left[\sum_{i=1}^{i=N} \alpha_i^k(n) \pi_i\right] \qquad (7)$$

for some set of vectors $\alpha^k(n)=[\alpha_1^k(n), \alpha_2^k(n), \cdots, \alpha_N^k(n)]$, $k=1, 2, \cdots$. We shall use the term *α-vector* to refer to one of the vectors in (7).

The proof of this important property proceeds by induction. Equation (6) demonstrates that $V_n(\pi)$ has the desired form for $n=0$. Now, assuming that $V_{n-1}(\pi)$ is of the form in (7), we shall prove that this implies that $V_n(\pi)$ is of the same form. This involves a rather straightforward substitution into (5). The critical part of (5) is $V_{n-1}[T(\pi|a, \theta)]$. Substituting (1) into this quantity yields

$$V_{n-1}[T(\pi|a, \theta)]=[\max_k \textstyle\sum_j \alpha_j^k(n-1) \sum_i \pi_i p_{ij}^a r_{j\theta}^a]/[\sum_{i,j} \pi_i p_{ij}^a r_{j\theta}^a]. \qquad (8)$$

As illustrated in Fig. 3, if $V_{n-1}(\cdot)$ is piecewise linear and convex, the space of information vectors can be divided into a finite set of convex regions separated by linear hyperplanes such that $V_{n-1}(\pi)=\pi\cdot\alpha^k(n-1)$ within a region for a single index k. It will prove convenient for the succeeding development to define a function $l(\pi, a, \theta)$ that is equal to the corresponding α-vector index for the region containing the transformed information vector $T(\pi|a, \theta)$. In other words, if there are n control intervals remaining, if our current information state vector is π, and if we apply control alternative a and observe an output θ, then the total expected accrued rewards $V_{n-1}[T(\pi|a, \theta)]$ from the optimal policy during the remaining $n-1$ control intervals will just be the quantity on the right side of (8) with the index k equal to $l(\pi, a, \theta)$:

$$V_{n-1}[T(\pi|a, \theta)]=[\textstyle\sum_j \alpha_j^{l(\pi,a,\theta)}(n-1) \sum_i \pi_i p_{ij}^a r_{j\theta}^a]/[\sum_{i,j} \pi_i p_{ij}^a r_{j\theta}^a]. \qquad (9)$$

Figure 3 illustrates an example of $l(\pi, a, \theta)$ for the three-state, two-outcome situation.

With this definition, substituting (9) into (5) yields

$$V_n(\pi)=\max_{a\epsilon A(n)} [\textstyle\sum_i \pi_i[q_i^a+\sum_{\theta,j} p_{ij}^a r_{j\theta}^a \alpha_j^{l(\pi,a,\theta)}(n-1)]]. \qquad (10)$$

To show that the expression in (10) is of the same form as (7), let us focus on the outer bracketed quantity in (10) for some control alternative a. If we can demonstrate that this quantity is piecewise linear and convex in π, then, since the maximum of a set of piecewise linear convex functions is itself piecewise linear and convex, this will prove that $V_n(\pi)$ is of the form in (7). To this end, notice first that, for each a and θ, $l(\pi, a, \theta)$ is a finitely valued function of π. This fact, plus the assumed convexity of $V_{n-1}(\cdot)$ and the continuity of $T(\pi|a, \theta)$, imply that $l(\pi, a, \theta)$ partitions the space of information vectors into a finite number of regions such that $l(\pi, a, \theta)$ is single-valued over each region. If we hold a constant, the function $l(\pi, a, \theta)$ defines a different partition of the space of information vectors for each output θ. Now, let us take the common refinement defined by the union of the region boundaries within each partition. This common refinement is a new partition such that the inner summand of (10) is constant over each region of this new partition. The net result is that the bracketed quantity in (10) is piecewise linear over the space of information vectors. The convexity of the bracketed quantity follows easily from the maximization in the definition of $l(\pi, a, \theta)$ [*see* (8) and (9)]. Thus, the outer bracketed quantity in (10) is of the form in (7) for each control alternative a; it then follows that $V_n(\cdot)$ is also of this form. This completes the proof of the piecewise-linear, convex form of $V_n(\cdot)$.

There are two important practical points to keep in mind. First, if the set of α vectors for $V_{n-1}(\cdot)$ has been calculated, then it is possible using (8) and (10) to calculate the optimum control alternative and the corresponding α-vector for any specified information vector π for the n-horizon case. This property will be most useful when we derive an algorithm for calculating the α-vectors in Section III. Secondly, the calculation of a new α-vector using (10) yields an optimal control alternative associated with each new α-vector. Thus, in storing the optimal control policy, it is not necessary to store the complete description of the policy regions as illustrated in Fig. 3; we need only store the set of α-vectors along with the appropriate control alternative for each α-vector. Then, to find the optimal control alternative for some information vector π, we merely carry out the maximization in (7) and then use the control alternative associated with the maximizing α-vector. This represents a considerable practical saving over previous solutions to this problem.

II. THE MACHINE-MAINTENANCE EXAMPLE

To ILLUSTRATE THE ideas of the preceding section, let us consider a more explicit formulation of the machine-maintenance example outlined earlier.

The machine under consideration has two identical internal components, each of which must operate on the product before it is finished. Since the components are identical, we can model the internal dynamics of the machine by a three-state discrete-time Markov process with the three states corresponding to zero, one, or two internal components that have failed. If component breakdowns are independent of one another, and if there is a probability of 0.1 that an operational component will break down during the manufacture of a product, then the matrix of transition probabilities for the normal operation of the manufacturing process is

$$\begin{bmatrix} 0.81 & 0.18 & 0.01 \\ 0 & 0.9 & 0.1 \\ 0 & 0 & 1 \end{bmatrix}.$$

We shall assume that, if a component has failed, then, in its processing of the product, there is a fifty-fifty chance that it will cause the product to be defective. Thus, for the control alternative in which we examine the quality of the finished product, the probabilities of observing a nondefective product are 1.0, 0.5, and 0.25 if there are zero, one, or two faulty internal components, respectively. If there is a profit of one or zero units for producing a nondefective or defective product, respectively, then the expected immediate production profit for a machine that begins the production cycle with zero, one, or two internal components that have failed is 0.9025, 0.427, and 0.25, respectively. For the sake of simplicity, these calculations have assumed that the breakdown of an internal component during a production cycle precedes its operation upon the product, that is, that the transitions governing the internal dynamics of the machine precede the actual manufacture of the product during any production cycle. This ensures correspondence with the assumed sequence of operations illustrated in Fig. 2.

For this maintenance problem there are four control alternatives available during each production cycle (control interval). In the first alternative, we simply manufacture another item, but without examining whether or not the resulting item is defective. For the second alternative, we proceed as in the manufacture alternative, except that we examine the finished product at a cost of 0.25 units. There are two observable outputs for this alternative corresponding to the production of a nondefective or defective product. In the third control alternative, the manufacturing process is interrupted for one production cycle, the machine is dismantled, and the two internal components are inspected and replaced if they have failed.

The replacement cost for each component is one unit and there is a 0.5-unit additional cost for inspecting the status of the internal components. The fourth and final control alternative involves the replacement of both internal components with no prior inspection. This alternative accrues the two-unit cost for the replacements, but does not incur an inspection cost. Table II lists the numerical values of the problem parameters for each of these four control alternatives.

A reasonable choice for the terminal reward is the replacement cost (i.e., salvage value) for the internal components that are still operable. Thus, we have $q_0^0=2$, $q_1^0=1$, $q_2^0=0$, where the subscripts refer to the number of faulty internal components at the termination of the process.

Figures 4a, 4b, 4c, and 4d portray the complete optimum policy regions for the cases in which the are 3, 4, 7, and 11 control intervals remaining. In these figures the solid lines specify the regions over which the optimal control is constant, while the dotted lines subdivide the regions into subregions over which a fixed α-vector $\alpha^k(n)$ maximizes the bracketed quantity in (7).

TABLE II

PARAMETER VALUES FOR THE EXAMPLE

Control alternative	P^a	R^a	q^a
Manufacture	$\begin{bmatrix} 0.81 & 0.18 & 0.01 \\ 0 & 0.9 & 0.1 \\ 0 & 0 & 1.0 \end{bmatrix}$	$\begin{bmatrix} 1.0 & 0.0 \\ 1.0 & 0.0 \\ 1.0 & 0.0 \end{bmatrix}$	$\begin{bmatrix} 0.9025 \\ 0.475 \\ 0.250 \end{bmatrix}$
Examine	$\begin{bmatrix} 0.81 & 0.18 & 0.01 \\ 0 & 0.9 & 0.1 \\ 0 & 0 & 1.0 \end{bmatrix}$	$\begin{bmatrix} 1.0 & 0.0 \\ 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix}$	$\begin{bmatrix} 0.6525 \\ 0.225 \\ 0.00 \end{bmatrix}$
Inspect	$\begin{bmatrix} 1.0 & 0 & 0 \\ 1.0 & 0 & 0 \\ 1.0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1.0 & 0.0 \\ 1.0 & 0.0 \\ 1.0 & 0.0 \end{bmatrix}$	$\begin{bmatrix} -0.50 \\ -1.50 \\ -2.50 \end{bmatrix}$
Replace	$\begin{bmatrix} 1.0 & 0 & 0 \\ 1.0 & 0 & 0 \\ 1.0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1.0 & 0.0 \\ 1.0 & 0.0 \\ 1.0 & 0.0 \end{bmatrix}$	$\begin{bmatrix} -2.0 \\ -2.0 \\ -2.0 \end{bmatrix}$

There are several important characteristics of these results. First of all, there is a great variation in the size and shape of the optimum policy regions. In particular, the optimal region for the 'manufacture' control alternative decreases in size as the number of remaining control intervals increases from 3 to 7; but then it increases in size again in going from a control horizon of 7 to 11. If we were to portray the optimum policy regions for longer horizons, the sizes and shapes of the regions for this problem eventually stabilize. However, not all problems exhibit this property; a later paper on this infinite-horizon problem will examine this property more closely.

Secondly, the location of the optimum policy regions in Figs. 4a–4d is intuitively appealing. That is, the 'manufacture' alternative is optimal if we are reasonably sure that both internal components are working, the 'replace' alternative is optimal when we feel that both components have failed, the 'inspect and replace' alternative is optimal when we believe that exactly one internal component has failed, and the 'examine output' alternative is optimal if we are uncertain whether the number of broken-down components is zero or two.

Finally, it is important to notice that, while the region for a particular α-vector is convex, the complete region for a control alternative is not necessarily convex. In fact, as illustrated in Fig. 4d, the separate regions for a single control alternative can even be disjoint.

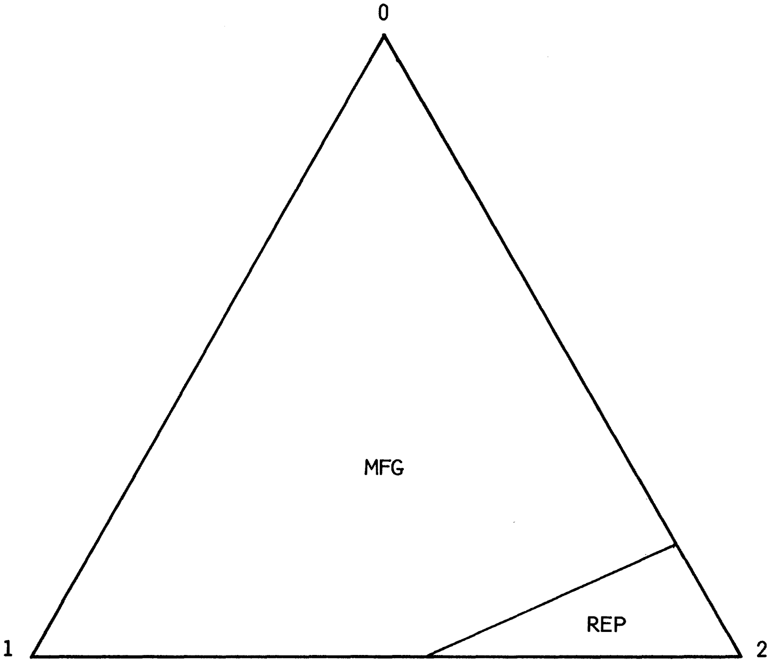


Fig. 4a. The optimal-policy and α-vector regions for the control horizon with $n = 3$. (MFG = manufacture; REP = replace only.)

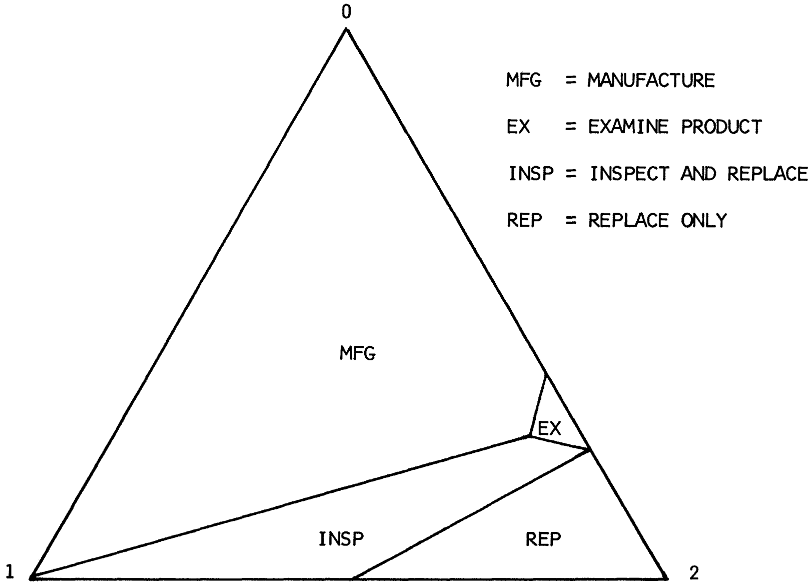


Fig. 4b. The optimal-policy and α-vector regions for the control horizon with $n = 4$.

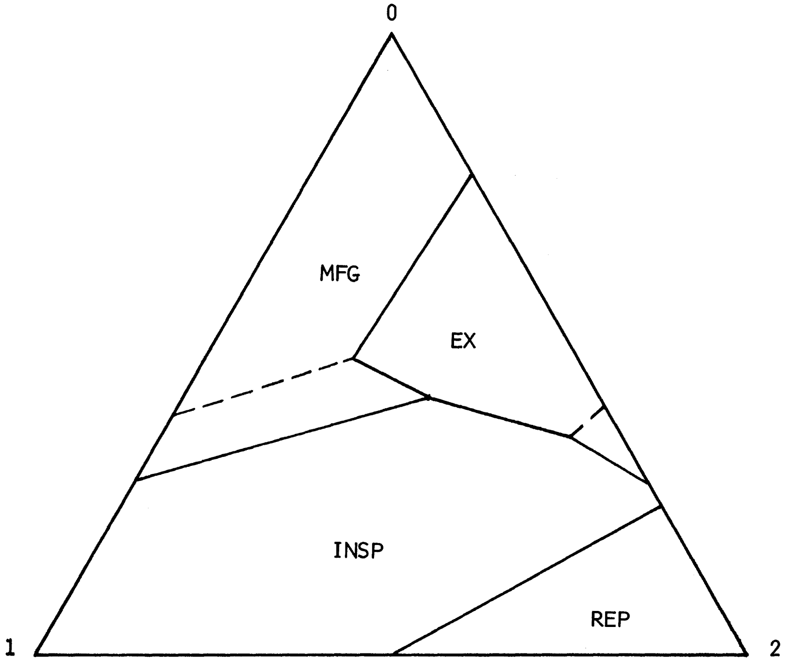


Fig. 4c. The optimal-policy and α-vector regions for the control horizon with $n=7$.

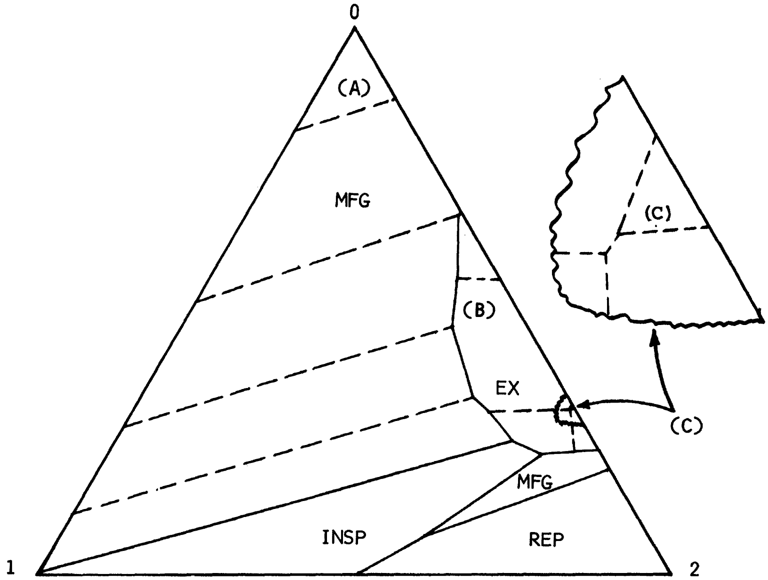


Fig. 4d. The optimal-policy and α-vector regions for the control horizon with $n=11$.

Although diagrams such as the ones in Figs. 4a–4d are convenient for problems that contain only two or three internal states, such diagrams are impractical for more complex problems. In this case, the numerical values of the α-vectors and their associated mapping onto the control alternatives are the most convenient way to portray the complete optimal control policy. However, for the cases in which we desire the optimal control policy for only a single information-state vector and control horizon, a very simple decision-tree format can be used to protray the optimal policy. In fact, a single decision tree will be valid for a complete region of information vectors. Thus, as an example, Fig. 5 illustrates the complete optimal control policy for three regions of the information vector space with eleven control intervals remaining; the three regions are labeled A, B, and C in Fig. 4d. Region A includes the case in which both internal components are known to be working; in this case the optimal policy is to manufacture five items with no examination, inspect and replace any internal components that have failed, and then manufacture five more items without examination. This policy is rather simple; however, the optimal policy for regions B and C, as portrayed in Figs. 5(b) and 5(c), are more complex, with succeeding actions dependent upon the quality of examined products. Such representations can be very useful in translating the complete specification of the optimal control policy into a practical format for implementation.

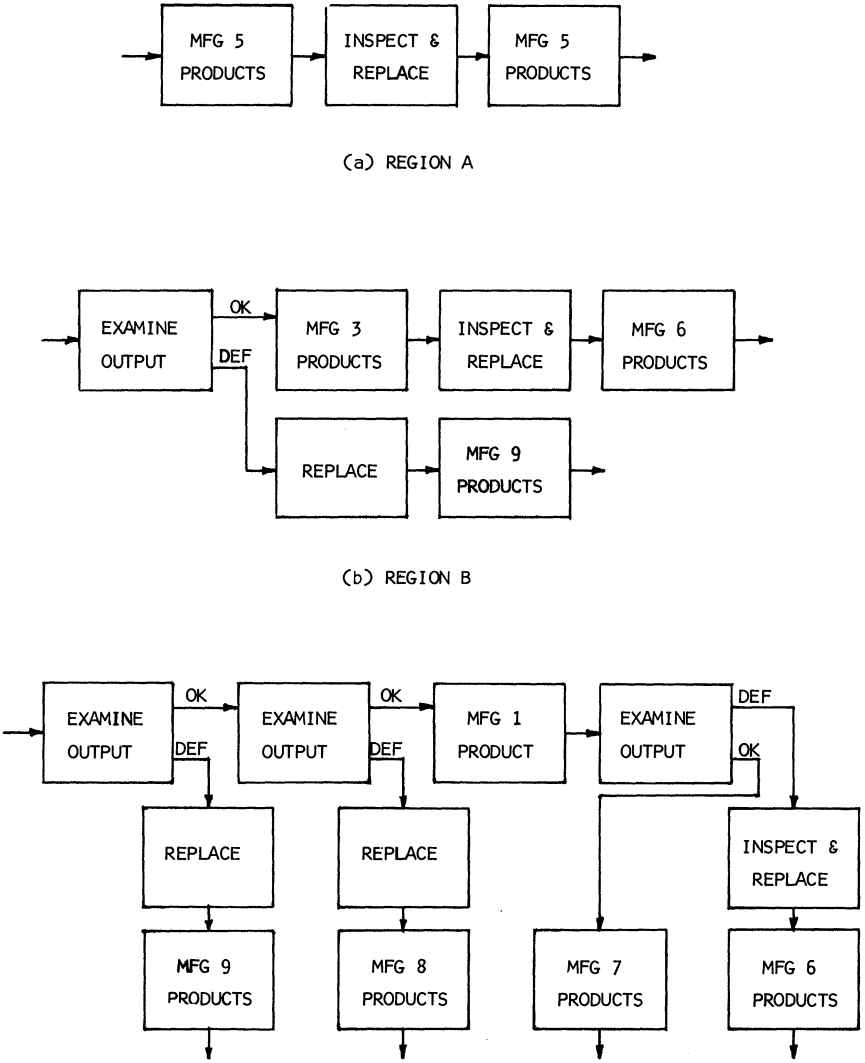
III. AN ALGORITHM FOR COMPUTING $V_n(\pi)$

Having discovered the relatively simple form of the solution to the optimal control problem, it only remains to construct an orderly practical procedure for calculating this control policy. In other words, we require an algorithm for computing the α-vectors and the corresponding mapping of these vectors onto the set of alternative controls. In the succeeding discussion, we shall assume that the α-vectors $\alpha^k(n-1)$ for the case of $n-1$ control intervals have been calculated. The problem then is to find an algorithm for calculating the vectors $\alpha^k(n)$ from this information.

One scheme for accomplishing this would lay out a grid of information vectors and then use (10) to calculate $\alpha^k(n)$ and the corresponding optimal control alternative for each of these information vectors. Unfortunately, this procedure does not guarantee the detection of all the α-vectors. This section describes an algorithm that uses this idea of calculating the appropriate α-vector at a number of distinct points in the space of information vectors. However, in choosing the grid, the algorithm is guaranteed to find *all* of the vectors $\alpha^k(n)$; and furthermore, the number of points chosen equals only the number of α-vectors.

To begin the algorithm, we pick an information vector, say π^0, and then calculate using (10) the optimum control alternative and corresponding α-vector if there are n control intervals remaining in the process. We shall denote these two quantities by a^* and $\alpha^*(n)$, respectively. Now the algorithm proceeds by identifying the region of the information-vector space over which $\alpha^*(n)$ is the appropriate α-vector. With reference to Fig. 3, we can imagine moving an information vector π away from π^0 and calculating [from (10)], for every π, $V_n(\pi)$, and the corresponding α-vector α. We keep moving π until $\alpha \neq \alpha^*(n)$. Since the quantity multiplying π_i in (10) is just $\alpha_i(n)$, it is clear that there are only two ways for α to change. Either the quantity $l(\pi, a^*, \theta)$ will change from $l(\pi^0, a^*, \theta)$ for some output θ or else the optimum control alternative will change.

The first case occurs when we move π away from π^0 until one of the corresponding points $T(\pi|a^*, 1)$, $T(\pi|a^*, 2)$, $\cdots$ in the space of information vectors at time



(a) REGION A

(b) REGION B

(c) REGION C

Fig. 5. The complete optimal policy for regions A, B, and C in Fig. 4d with $n=11$.

$n-1$ eventually crosses the boundary of a region. If we focus for a minute on one value of θ, then the condition for the point $T(\pi|a^*, \theta)$ to remain in the same region specified by $l(\pi^0, a^*, \theta)$ is

$$T(\pi|a^*, \theta)\cdot\alpha^l(n-1)\geqq T(\pi|a^*, \theta)\cdot\alpha^k(n-1) \quad \text{for all} \quad k, \tag{11}$$

where $l(\pi^0, a^*, \theta)$ has been abbreviated l. Substituting (1) into (11) yields the following linear form for these inequalities

$$\sum\nolimits_{i,j} \pi_i p_{ij}^{a^*} r_{j\theta}^{a^*} [\alpha_j^{\,l}(n-1) - \alpha_j^{\,k}(n-1)] \leqq 0 \quad \text{for all} \quad k \tag{12}$$

There will be a set of these inequalities for each possible output θ. Actually, the situation is not quite as severe as stated in (12). In fact, the inequality in (12) can be limited to only the values of k for which the region of $\alpha^k(n-1)$ forms a boundary with the region for $\alpha^l(n-1)$. This simplification represents a considerable computational saving and is easily incorporated into the algorithm by storing, along with the α-vectors, the indices of the hyperplanes that define the region for the vector.

The second condition for the region of $\alpha^*(n)$ is that the control alternative a^* must be optimal. In calculating the optimal control alternative at π^0, we must calculate the expression multiplying $\pi_i^{\,0}$ in (10) for each alternative a. If $\alpha_a(n)$ denotes the α-vector calculated for alternative a at the point π^0, then the condition for a^* to remain the optimal control alternative is

$$\pi \cdot \alpha^*(n) \geqq \pi \cdot \alpha_a(n) \quad \text{for all} \quad a \epsilon A(n), \tag{13}$$

or

$$\sum\nolimits_i \pi_i [\alpha_i^{\,*}(n) - \alpha_{a,i}(n)] \geqq 0 \quad \text{for all} \quad a \epsilon A(n). \tag{14}$$

If we add to the sets of inequalities in (12) and (14) the conditions

$$\pi_i \geqq 0 \quad \text{for} \quad 1 \leqq i \leqq N \quad \text{and} \quad \sum\nolimits_{i=1}^{i=N} \pi_i = 1, \tag{15}$$

then (12), (14), and (15) together specify the region in the space of information vectors over which the α-vector $\alpha^*(n)$ defines the optimal payoff function $V_n(\pi)$. In addition, of course, the control alternative a^* is optimal in this region.

Generally, only some subset of these linear constraints will be necessary to define the region; that is, some of the hyperplanes in (12), (14), and (15) will not be boundaries of the region. Appendix B describes a linear-programming algorithm for identifying the constraints that are the defining ones for the region. The application of this procedure to the linear inequalities in (12), (14), and (15) yields a minimum subset of inequalities that define the region.

Each defining inequality of the form in (12) will identify a new α-vector whose value can be calculated by substituting the index k from (12) for $l(\pi^0, a^*, \theta)$ in (10). This new α-vector must be added to a list of α-vectors whose defining regions will be calculated later; the control alternative for this new α-vector is still a^*. The defining inequalities of the type in (14) will produce both a new α-vector $\alpha_a(n)$ and a new optimal control alternative a. This new α-vector must also be added to the list for later examination. In this way, the algorithm calculates the appropriate regions for each α-vector and, in the process, discovers additional α-vectors whose defining regions must be identified later. When the list of new α-vectors is exhausted, the specification of the optimal control policy is complete.

The following four steps summarize the complete algorithm for calculating the optimal control policy for the n-horizon case if the α-vectors for the $(n-1)$-horizon case are known:

1. Pick an initial state vector and calculate the optimal control alternative a^* and corresponding α-vector $\alpha^*(n)$.

2. Construct the complete list of inequalities for the region using (10), (12), (14), and (15).

3. Use the linear-programming procedure of Appendix B to calculate the minimum set of inequalities that define the region for $\alpha^*(n)$. From each boundary of the region, construct, using (10), a new α-vector and store the following in a list for each: the vector, its corresponding optimal control alternative, and one information vector for which it is the α-vector (easily obtained from the linear program).

4. Store the indices of the α-vectors that are neighbors to the region under consideration to limit the number of inequalities of the type in (12) during the $(n+1)$-horizon calculations. If there are any α-vectors on the list whose region has not been calculated, pick a new α-vector and return to Step 2. Otherwise, the complete specification of the optimal control policy has been calculated.

This simple, four-step procedure can be used successively to calculate the optimal control policy and payoff function for any finite number of control intervals.

IV. CONCLUSIONS AND DISCUSSION

As STATED EARLIER, previous authors have attacked the calculation of the optimal control policy for partially observable Markov processes by quantizing the space of information vectors. This technique essentially converts the continuous-state Markov process (*see* Appendix A) to a finite-state Markov process, and the problem can then be handled by traditional techniques, as developed by Howard.[3] However, the number of states in this approximate finite-state Markov process becomes prohibitively large for any but the smallest problems. For example, if a quantization interval of 0.05 is used, then a five-state internal process will require 6.2 million states in the quantized process. This is, of course, a completely impractical problem from a computational point of view. Since the technique in this paper does not require this quantization, it provides a significant increase in the size of the problem for which an optimal control policy can be calculated.

In working with the algorithm of Section III, we have found that most of the computation time is spent on using the linear programming technique to find the defining hyperplanes for the optimal policy regions. Therefore, the computation time for the algorithm is dependent, not only upon the number of states, but also upon the number of policy regions, i.e., the number of distinct α-vectors. In an attempt to circumvent this dependence on the number of policy regions, we have developed a second algorithm that does not require this linear-programming step and which appears superior to the algorithm above for small numbers of states, i.e., less than four (this algorithm is described in more detail by SONDIK.[9]) As a specific illustration, the first algorithm, when applied to the machine-maintenance example required approximately 50 seconds to calculate the optimal policy regions for 8 time periods, while the second algorithm required 50 seconds to calculate the optimal policy for 13 time periods. [The calculations were done on the Stanford Computation Center 360/67 using WATFIV in the batch partition.]

Although the formulation of the control problem in (5) has not assumed any discounting, it is a trivial matter to multiply the second term in (5) by a discount factor. The remainder of the development, the results, and the algorithm then follow as before. The results of the preceding sections are also applicable, with minor algebraic changes, when the sequence of operation in Fig. 2 is changed from (control, transition, output) to (control, output, transition). In this case, $r^a_{j\theta}$ must be changed to $r^a_{i\theta}$ in (1) through (10), but the form of the solution in (7) and the algorithm in Section III remain unchanged.

In summary, this paper has formulated the optimal control problem for partially observable Markov processes, has shown that the optimal payoff function is piecewise linear and convex, and has presented an algorithm that uses this property to calculate the optimal control policy for the finite-horizon case. Finding the optimal stationary control policy over an infinite horizon is more complicated; a later paper will extend these results to that case.

APPENDIX A

IN THIS APPENDIX we show that a sufficient statistic for the past history of observations of a partially observable Markov process is just the current information-state vector π. In demonstrating this property, we derive the rule for updating the information-state vector from one control interval to the next. To make this explicit, we define $\epsilon(t)$ as the total available information about the process at the end of control interval t. Notice that in this appendix the time variable t increases with increasing time, whereas in the main body of the paper the time variable n, which is equal to the number of remaining control intervals, decreased with increasing time. For the process as defined in this paper, the only information that we obtain during a control interval is the fact that the application of a particular control alternative produced the observed output. If $a(t)$ and $z(t)$ denote the control alternative and corresponding output, respectively, during control interval t, then we can write

$$\epsilon(t)=[z(t),\ a(t),\ \epsilon(t-1)] \tag{A1}$$

That is, $\epsilon(t)$ represents our state of information prior to control interval t plus the additional information that a particular control alternative and output were recorded.

By the definition of the information state vector,

$$\pi_j(t)=\mathrm{pr}\{s(t)=j|\epsilon(t)\}, \tag{A2}$$

where $s(t)$ is a discrete-valued random variable equal to the internal state of the process at the conclusion of control interval t. The substitution of (A1) into (A2) plus the application of Bayes' rule yields

$$\pi_j(t)=\mathrm{pr}\{s(t)=j,\ z(t)=\theta|a(t),\ \epsilon(t-1)\}/\mathrm{pr}\{z(t)=\theta|a(t),\ \epsilon(t-1)\}, \tag{A3}$$

where, to make things explicit, we have assumed that the output during control interval t was observation θ. The expansion of the numerator in (A3) over all possible internal states of the process at the end of $t-1$ plus the expansion of the joint probability as a product of conditional probabilities produces

$$\begin{aligned}\pi_j(t)=\textstyle\sum_i \mathrm{pr}\{s(t-1)=i|a(t),\ \epsilon(t-1)\}\mathrm{pr}\{s(t)=j|s(t-1)=i,\ a(t),\ \epsilon(t-1)\}\\ \mathrm{pr}\{z(t)=\theta|s(t)=j,\ s(t-1)=i,\ a(t),\ \epsilon(t-1)\}/\mathrm{pr}\{z(t)=\theta|a(t),\ \epsilon(t-1)\}.\end{aligned} \tag{A4}$$

The first probability in the numerator of (A4) will be independent of $a(t)$, since the control alternative is completely under our control and thus does not provide any information about the previous state of the process. The remaining two probabilities in the numerator of (A4) are just transition probabilities and response probabilities for the process, while the denominator in the equation is just the numerator summed over all values of j. Thus, we have

$$\pi_j(t)=[\textstyle\sum_i \pi_i(t-1)p_{ij}^{a(t)}r_{j\theta}^{a(t)}]/[\textstyle\sum_{i,j} \pi_i(t-1)p_{ij}^{a(t)}r_{j\theta}^{a(t)}], \qquad \text{(A5)}$$

which is (1) of the main body of the paper.

The important feature of (A5) is that the calculation of the information-state vector after control interval t requires only $\pi(t-1)$, the information-state vector after control interval $t-1$; thus, $\pi(t-1)$ summarizes all the information gained prior to control interval t and represents a sufficient statistic for the complete past history of the process $\epsilon(t-1)$.

In fact, (A5) describes the possible transitions for a continuous-state Markov process in which the state of the process is the information state vector $\pi(t)$. For this process, the denominator of (A5) is the transition probability of the transition $\pi(t-1)\rightarrow T[\pi(t-1)|a(t), \theta]$. This is rather a special case of a continuous-state Markov process, since the state is continuous but the state transition probabilities are discrete.

APPENDIX B

EQUATIONS (12), (14), AND (15) define a region in the space of information vectors, and we require a technique for determining which of these constraints are the defining ones for the region. That is, which inequalities form actual boundaries for the region and which ones can be discarded? We can represent the set of linear inequalities as

$$\pi\cdot b^m \geqq 0, \qquad (m=1, 2, 3, \cdots) \quad \text{(B1)}$$

where the index m ranges over the set of constraints defined by (12) and (14). The solution to the linear program

$$\min_\pi \pi\cdot b^k,$$

subject to

$$\begin{aligned} &\pi\cdot b^m \geqq 0, && (m=1, 2, 3, \cdots) \\ &\pi_i \geqq 0, && (i=1, 2, \cdots, N) \quad \text{(B2)} \\ &\textstyle\sum_i \pi_i = 1, \end{aligned}$$

will yield a solution that has the slack variable for the kth inequality equal to zero if and only if this inequality forms a part of the boundary of the region. Thus, by solving a linear program of the form in (B2) for each of the constraints, we can identify the constraints that define the region and the ones that can be discarded.

The procedure can be made more efficient if, for each iteration of the linear programming problem with the kth inequality as the objective function, all other constraints are tested as objective functions to see if they are optimized at the current feasible solution. If a constraint is optimized at any point, then either this constraint forms a boundary of the region (a zero slack variable) or is a superfluous constraint (a nonzero slack variable). Once a constraint has been optimized, it need not be used as the objective function. We have found that this procedure typically decreases the number of linear programming iterations by approximately 50 percent.

In Section III, Step 3 of the algorithm requires that we have an information vector on each boundary of the region for later use when the α-vector for the bordering region is investigated. This information vector can be obtained easily from the linear programming problem, since π will lie precisely on the appropriate boundary.

ACKNOWLEDGMENT

THIS WORK WAS supported in part by NASA, and in part by the Joint Services Electronics Programs, Stanford Electronic Laboratories.

REFERENCES

1. ALVIN DRAKE, "Observation of a Markov Process through a Noisy Channel," Sc.D. Thesis, Electrical Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts, June 1962.
2. JAMES E. ECKLES, "Optimum Maintenance with Incomplete Information," *Opns. Res.* **16,** 1058–1067 (1968).
3. RONALD A. HOWARD, *Dynamic Programming and Markov Processes*, Wiley, New York, N. Y., 1960.
4. JAMES MATHESON, "Optimum Teaching Procedures Derived from Mathematical Learning Models," *EES Technical Report* No. CCS-2, Department of Engineering-Economic Systems, Stanford University, Stanford, California, 1964.
5. STEPHEN M. POLLOCK, "A Simple Model of Search for a Moving Target," *Opns. Res.* **18,** 883–903 (1970).
6. RICHARD D. SMALLWOOD, I. WEINSTEIN, AND J. ECKLES, "Quantitative Methods in Computer-directed Teaching Systems," Final Report Nonr-225(84), Department of Engineering-Economic Systems, Stanford University, Stanford, California, March 15, 1967.
7. ———, "Optimum Policy Regions for Computer-directed Teaching Systems," Ch. 6 in Wayne H. Holtzman (ed.) *Computer Assisted Instruction Testing and Guidance*, Harper and Row, New York, N. Y., 1970.
8. ———, "The Analysis of Economic Teaching Strategies for a Simple Learning Model," *J. of Math. Psych.* **8,** 285–301 (May 1971).
9. EDWARD J. SONDIK, "The Optimal Control of Partially Observable Markov Processes," Ph.D. Dissertation, Department of Engineering-Economic Systems, Stanford University, Stanford, California, June 1971.