Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Optimal Control of Partially Observable Markov Processes over the Infinite Horizon: Discounted Costs

Edward J. Sondik,

To cite this article:
Edward J. Sondik, (1978) The Optimal Control of Partially Observable Markov Processes over the Infinite Horizon: Discounted Costs. Operations Research 26(2):282-304. https://doi.org/10.1287/opre.26.2.282

OPERATIONS RESEARCH
Vol. 26, No. 2, March–April 1978

0030-364X/78/2602-0282 $01.25

The Optimal Control of Partially Observable Markov Processes over the Infinite Horizon: Discounted Costs

EDWARD J. SONDIK

Stanford University, Stanford, California

(Received July 1973; accepted May 1977)

This paper treats the discounted cost, optimal control problem for Markov processes with incomplete state information. The optimization approach for these partially observable Markov processes is a generalization of the well-known policy iteration technique for finding optimal stationary policies for completely observable Markov processes. The state space for the problem is the space of state occupancy probability distributions (the unit simplex). The development of the algorithm introduces several new ideas, including the class of finitely transient policies, which are shown to possess piecewise linear cost functions. The paper develops easily implemented approximations to stationary policies based on these finitely transient policies and shows that the concave hull of an approximation can be included in the well-known Howard policy improvement algorithm with subsequent convergence. The paper closes with a detailed example illustrating the application of the algorithm to the two-state partially observable Markov process.

THIS PAPER studies the control of Markov processes for which only partial or incomplete state information is available. Partial information in our context is a fairly general structure, ranging from complete state information (a completely observable process) to no state information at all (a completely unobservable process). Various examples of such processes were presented in Smallwood and Sondik [14] in their discussion of the optimal control of partially observable Markov processes over the finite horizon. That work, encompassing both cases of discounted and undiscounted costs, is extended in this paper to the infinite planning horizon for the case of discounted costs. The optimal control of partially observable Markov processes for undiscounted costs over the infinite horizon will be considered in a future paper.

As discussed in Smallwood and Sondik [14] and in Sondik [15], much of the work on the optimal control of Markov processes stems from Howard [6]. Over the previous decade there has been extensive research into the completely observable control problem; related work over arbitrary, rather than discrete, state spaces has been presented by Blackwell [1], Ross [11],

and others. As we shall point out below, our work is based on results from both research areas since the state space for the partially observable Markov process is continuous. The control of the partially observable process over the infinite horizon has been discussed by Matheson [10], Smallwood [13], and Eckles [5] from the standpoint of teaching machines. Kakalik [7] and Satia [12] have considered approximate solutions to the general problem. Of special significance to this paper is the work of Drake [4], who considered the decoding of a Markov source over a noisy channel. He formulated a three-alternative optimization problem and found the optimal control minimizing the expected cost of decoding the channel. Drake also considered a phenomenon closely related to what we term here a finitely transient policy. Finitely transient policies, which are discussed extensively in a later section, are a particular class of control policies for partially observable Markov processes that generate dynamics equivalent to those of a completely observable process. They play a key role in our results.

This paper develops a general solution method for determining the optimal control of partially observable Markov processes with discounting. The method presented here finds the exact optimal controls in some cases, while in other cases it leads to controls arbitrarily close to the optimum. As noted below, this is the best that can be expected. Our development is based on an analysis of finitely transient policies, a subclass of stationary policies possessing special properties that allow straightforward computation of expected costs. It can be shown that for a completely observable Markov process, any stationary policy is finitely transient. Thus, this class of policies provides a convenient conceptual bridge between the completely observable and the partially observable Markov processes. The next section describes the model and notation, Section 2 describes the optimization problem, and Section 3 defines and derives several properties of finitely transient policies. Subsequent sections develop approximations to stationary policies based on finitely transient policies, derive a policy improvement procedure, present a policy iteration algorithm, and illustrate the algorithm with an example.

1. THE MODEL AND NOTATION

Following the definitions and notation presented in [14], the partially observable Markov process consists first of the core process, an N-state discrete-time (time invariant) Markov process with states labeled $1, \cdots, N$. The core process is described by an $N \times N$ transition matrix $P = [p_{ij}]$, where p_{ij} is the probability that the process will occupy state j at time $t+1$ given that the state at time t is i. An observer does not directly observe the process; he sees instead one of M outputs where the conditional probability of viewing output θ, given that the current state of the process is i, is denoted $r_{i\theta}$. The corresponding matrix is given by R.

To introduce control into the formulation we assume that at each instant of time we are free to choose one of some finite set S of alternative state and observation dynamics. The ath alternative is described by a state transition matrix P^a, an output probability matrix R^a and a (column) vector of expected costs $q^a = [q_i^a]$, where q_i^a is the expected cost of making one transition from state i under alternative a. Without loss of generality, we assume that each alternative has N states and M outputs and that the costs incurred are viewed only at process termination. In general, costs observed during process operation must be considered as additional outputs and included in the formulation of the observation matrix R^a.

We assume a sequence of operation as follows: A Markov process, say the ath, is selected to govern the next state transition. When the state has changed (possibly no change if a transition occurs from i to i) and a cost has been incurred, an output is generated according to probabilities governed by R^a. Note that the rules governing the observed output depend on the alternative that governed the last transition. (It is a straightforward operation to modify the following analysis to accom nodate the case in which the outputs are generated before the state transition.)

We define $\pi(t)$ as the (row) vector of state probabilities $\pi_i(t)$ where $\pi_i(t)$ is the probability (based on all past information) that the current state (i.e., at time t) of the process is i. We shall refer to the vectors $\pi(t)$ as information or state vectors, and the space of all $\pi(t)$, Π, as the information or state space for the partially observable Markov process. (That $\pi(t)$ is an appropriate definition of state for this problem can be readily verified.) Now, if the current information vector is $\pi(t)$ and alternative a is selected, and if output θ results, then the new information vector $\pi(t+1)$ is given by $T(\pi(t)|\theta, a)$ where by Bayes' rule

$$T(\pi|\theta, a) = \sum_j \pi_j p_{ji}^a r_{i\theta}^a / \sum_{i,j} \pi_j p_{ji}^a r_{i\theta}^a = \pi P^a R_\theta^a / \{\theta|\pi, a\}$$

where $R_\theta^a = \text{diag}\ [r_{1\theta}^a, \cdots, r_{N\theta}^a]$, $\{\theta|\pi,a\} = \sum_{i,j} \pi_j p_{ij}^a r_{i\theta}^a = \pi P^a R_\theta^a \mathbf{1}$, and where $\mathbf{1}$ is a column vector of 1's. The scalar quantity represented by $\{\theta|\pi, a\}$ is simply the probability that the next output will be θ, given that the current information vector is π, and that alternative a is next selected. Note that for a given sequence of alternatives $a_1, a_2, \cdots$ the subsequent information vectors form a Markov process over the continuous state space Π where the probability that the next information vector will be $T(\pi|\theta, a)$ is given by $\{\theta|\pi, a\}$. Throughout this paper we make only one assumption about the structure of the Markov process—that $T(\pi|\theta, a)$ is 1 to 1.

2. THE INFINITE HORIZON PROBLEM WITH DISCOUNTING

We shall assume that costs are discounted by discount factor β, $0 \leqq \beta < 1$, and that we seek to minimize the total expected cost of operating the

partially observable Markov process over all time. We require some preliminary definitions.

Definitions

A finite connected partition of Π, $[V_1, \cdots, V_l]$, is a finite collection of mutually exclusive and exhaustive connected subsets of Π. A control function $\delta: \Pi \rightarrow S$ is *admissible* if there exists a finite connected partition of Π such that $\delta(\pi)$ is constant over each set V_j. We denote the set of admissible controls by Δ and assume that all control functions δ considered here are members of Δ.

We use the notation $\delta_t(\pi)$ to represent the alternative to be used if the information vector at time t is $\pi(t)$. Then the discounted expected value control problem for fixed $\pi(0)$ can be written as

$$\min_{(\delta_0, \delta_1, \cdots)} [E_{\pi(0)} \sum_{t=0}^{\infty} \beta^t \pi(t) q^{\delta_t(\pi(t))}]$$

where the sequence of control functions $(\delta_0, \delta_1, \cdots)$ must be selected to minimize the expression. Such a sequence of control functions is termed a *policy*. If a policy consists of a single control function used at each time period, then the policy is termed *stationary*. A stationary policy is denoted by $(\delta)^\infty = (\delta, \delta, \cdots)$.

From Blackwell [1], it follows that the minimum expected cost as a function of the initial information vector π, $C^*(\pi)$, must satisfy

$$C^*(\pi) = \min_a [\pi q^a + \beta \sum_\theta \{\theta|\pi, a\} C^*(T(\pi|\theta, a))]. \tag{1}$$

Furthermore, a measurable policy (with respect to π) exists that achieves this minimum cost (Maitra [9]). This optimal policy is in fact stationary and is denoted by $(\delta^*)^\infty$, where $\delta^*(\pi)$ is the minimizing alternative in (1). Thus, (1) can be written as

$$C^*(\pi) = \pi q^{\delta^*} + \beta \sum_\theta \{\theta|\pi, \delta^*\} C^*(T(\pi|\theta, \delta^*)), \tag{2}$$

where the specific dependence of $\delta^*(\pi)$ on π has been suppressed.

We can solve (1) for δ^* by solving a finite horizon problem over successively larger horizons until we obtain convergence of the expected cost to C^*. Sondik [15] gives bounds on the distance to C^* from any horizon. This paper focuses on a second solution technique, a generalization of Howard's policy iteration algorithm for completely observable processes [6]. This generalization is expressed by the following theorems based on the work of both Blackwell [1] and Howard [6].

We define $C(\pi|\delta)$ to be the expected cost of following the stationary policy $(\delta)^\infty$ for all time.

THEOREM 1 (*Howard-Blackwell policy improvement*). *Let $\delta'(\pi)$ be the control function defined as the index a minimizing $U_a[\pi, C(\cdot|\delta)]$ where*

$$U_a[\pi, C(\cdot|\delta)] = \pi q^a + \beta \sum_\theta \{\theta|\pi, a\} C[T(\pi|\theta, a)|\delta]. \tag{3}$$

Then $C(\pi|\delta') \leqq C(\pi|\delta)$, *for every* $\pi \in \Pi$. *Furthermore, if* $C(\cdot|\delta) \neq C^*$, *there is some* π *such that* $C(\pi|\delta') < C(\pi|\delta)$.

For the proof of this theorem (presented in a slightly more general framework), see [1].

It is straightforward to develop a measure of the distance between $C(\cdot|\delta)$ and $C^*(\cdot)$ [15]. Any such measure is, effectively, a stopping rule for the iterative application of Theorem 1, and the measure and the theorem together form a policy iteration algorithm. We note that one cannot expect the policy iteration algorithm to converge in a finite number of iterations since the space of all stationary policies is uncountable.

Although Theorem 1 lays the theoretical basis for a policy iteration algorithm, there are two inherent difficulties in the implementation of such a procedure. First, the theorem requires the computation of $C(\cdot|\delta)$, the cost of a stationary policy $(\delta)^\infty$. One method for performing these computations proceeds by noting that the operator U_δ, defined for some control function $\delta(\pi)$ as

$$U_\delta(\pi, f) = \pi q^\delta + \beta \sum_\theta \{\theta|\pi, \delta\} f(T(\pi|\theta, \delta)),$$

is a contraction mapping under the supremum norm. (The supremum norm of a function f is defined as $\|f\| = \sup_{\pi \in \Pi} |f(\pi)|$.) It then follows that $C(\cdot|\delta)$ can be co mputed iteratively by defining a sequence of functions $f^n(\cdot)$, where $f^{n+1}(\pi) = U_\delta(\pi, f^n)$, $n \geqq 1$. By elementary properties of contraction mappings the sequence f^n converges (in the sense of the supremum norm) to $C(\cdot|\delta)$. In fact, $C(\cdot|\delta)$ is the unique bounded solution [3] to the equation

$$C(\pi|\delta) = \pi q^\delta + \beta \sum_\theta \{\theta|\pi, \delta\} C(T(\pi|\theta, \delta)|\delta). \tag{4}$$

The implementation of the operator U_δ, however, is far from trivial. It may be conjectured that the methods of [14] may be applied here in an iterative fashion to calculate $C(\cdot|\delta)$; those methods require, however, that $C(\cdot|\delta)$ be a piecewise linear concave function of π. It can be fairly easily demonstrated that a stationary policy exists whose cost function lacks these properties; thus, the methods of [14] are not directly applicable here. The methods we employ below compute or approximate $C(\cdot|\delta)$ from properties based on the fundamental structure of the partially observable Markov process and basic properties of the contraction mapping U_δ.

Whatever method is used to compute the expected cost $C(\cdot|\delta)$, there still remains an impediment in the path of using Theorem 1 to compute $C^*(\pi)$ and $\delta^*(\pi)$. To iterate with the theorem requires that a minimization operation be performed over all alternatives for *every point* in the space Π. By using certain fundamental properties of the process, we are able to reduce this excessive computation to a manageable level. The methods for performing this minimization consist of using the concave hull of $C(\cdot|\delta)$

in Theorem 1 in place of $C(\cdot|\delta)$. The justification for this substitution is presented below, where it is shown that, indeed, policies with successively lower expected costs can be found.

3. THE EXPECTED COST OF FINITELY TRANSIENT POLICIES

We begin this section with a characterization of solutions to the finite horizon problem.

DEFINITION. *A real-valued function $f(\cdot)$ over Π is termed "piecewise linear" if it can be written $f(\pi)=\pi\alpha_j$ for all $\pi\in V_j\subset\Pi$, where $V_1, \cdots, V_l$ is a finite connected partition of Π and α_j is a constant (column) vector for $\pi\in V_j$.*

It is demonstrated in [14] that the minimum expected cost of operating the partially observable Markov process for a *finite time period* is a piecewise linear function over Π. In general, C^* is not piecewise linear; but if the optimal policy belongs to a class of policies we term finitely transient, then C^* is piecewise linear and, furthermore, is a continuous function.

We define a finitely transient policy as a stationary policy $(\delta)^\infty$ where after some finite period of ti ne the information vector will not lie at a point in Π at which $\delta(\pi)$ changes value (i.e., where $\delta(\pi)$ is discontinuous). Thus, the points at which $\delta(\pi)$ is discontinuous are essentially "transient" in that after some finite time they can never recur. We will state this definition analytically, but first require the following definitions.

Definition

(i) For a stationary policy $(\delta)^\infty$ let T_δ be the set function defined for any set $A\subset\Pi$ by: $T_\delta(A)=$closure $[T(\pi|\theta,\ \delta):\pi\in A,\ \forall\theta]$ where the notation [:] is used to denote a set. (Thus $T_\delta(A)$ is the set of all possible next states given that the current state lies in A.)

(ii) Consider the sequence $S_\delta{}^n$ defined as $S_\delta{}^0=\Pi$, with the succeeding elements $S_\delta{}^n$ defined recursively as $S_\delta{}^n=T_\delta(S^{n-1})$, $n\geqq 1$. Note that $S_\delta{}^n$ contains all states of knowledge that can occur at the nth time period after the system begins operation. (Actual occurrence depends on the initial state of knowledge.)

(iii) We define $D_\delta=$closure $[\pi:\delta(\pi)$ is discontinuous at $\pi]$.

We may now define a finitely transient policy.

DEFINITION. *A stationary policy $(\delta)^\infty$ is "finitely transient" if and only if there is an integer $n<\infty$ such that $D_\delta\cap S_\delta{}^n=\Phi$, where Φ is the null set. The smallest such integer is the "index" of the finitely transient policy and is labeled n_δ.*

The following lemma verifies the earlier verbal definition of a finitely transient policy.

LEMMA 1. *If* $(\delta)^\infty$ *is finitely transient with index* n_δ, *then* $D_\delta \cap S_\delta^n = \Phi$ *for all* $n \geqq n_\delta$.

Proof. The proof follows directly if $S_\delta^{n+1} \subset S_\delta^n$, $n \geqq 0$. To show this note that $S_\delta^1 \subset S_\delta^0 \equiv \Pi$ by the fact that T_δ maps the space Π into itself. To show that $S_\delta^{n+1} \subset S_\delta^n$, suppose $S_\delta^n \subset S_\delta^{n-1}$ and consider a point $\pi^{n+1} \in S_\delta^{n+1}$. By definition there exists $\pi^n \in S_\delta^n$ such that $\pi^{n+1} = T(\pi^n|\theta, \delta)$ for some θ. By hypothesis $\pi^n \in S_\delta^{n-1}$; therefore, by definition of S_δ^n, $T(\pi^n|\theta, \delta) \in S_\delta^n$. Since $T(\pi^n|\theta, \delta) = \pi^{n+1}$, we have that $\pi^{n+1} \in S_\delta^n$. Therefore, if $\pi^{n+1} \in S_\delta^{n+1}$, $\pi^{n+1} \in S_\delta^n$, which by induction proves the nesting of the sets S_δ^n.

We now develop a series of results that lead to a proof that the expected cost of a finitely transient policy is piecewise linear. These results provide us with insight into the structure of partially observable Markov processes and are fundamental for calculation of $C(\pi|\delta)$ for arbitrary $(\delta)^\infty$.

Definition

(i) A partition $V = [V_1, V_2, \cdots]$ of Π that possesses the following properties with respect to a stationary policy $(\delta)^\infty$ is said to be a Markov partition.

Property (a) All points in the set V_j are assigned the same control alternative by δ [i.e., if $\pi^1, \pi^2 \in V_j$, then $\delta(\pi^1) = \delta(\pi^2)$].

Property (b) Under the mapping $T(\cdot|\theta, \delta)$ all points in V_j map into the same set. The relationship between the sets V_j and the mappings $T(\cdot|\theta, \delta)$ is given by the Markov mapping $\nu(j, \theta)$ such that if $\pi \in V_j$ then $T(\pi|\theta, \delta) \in V_{\nu(j,\theta)}$. If this property holds, we shall say that one set maps completely into another for each output.

(ii) The Markov mapping ν and Markov partition V satisfying (b) are said to be "equivalent to δ."

(iii) The control alternative over V_j is defined δ_j, i.e., if $\pi \in V_j$ then $\delta(\pi) \equiv \delta_j$.

The following lemma relates Markov partitions and mappings to finitely transient policies.

LEMMA 2. *For every finitely transient policy* $(\delta)^\infty$ *there exists a Markov partition of* Π, $V = [V_1, V_2, \cdots]$ *and Markov mapping* ν *satisfying properties* (*a*) *and* (*b*).

Before proving this lemma we first define a D-sequence of sets and a V-sequence of partitions.

(i) A D-sequence of sets $D^0, D^1, \cdots$ is defined as follows: $D^0 = D_\delta$, $D^n = [\pi : T(\pi|\theta, \delta) \in D^{n-1}$, for some $\theta]$, $n > 1$, where D_δ is defined earlier.

(ii) A V-sequence of partitions $V^0, V^1, \cdots$ is defined as follows: V^0 is the smallest collection of mutually exclusive connected sets in Π satisfying property (a). We say that V^0 is *induced* by δ, and note that the set D^0 forms the boundaries of the sets of V^0. Analogously for $k \geqq 1$, V^k is the refinement of V^{k-1} where the set $D^0 \cup \cdots \cup D^k$ forms the set boundaries of V^k. (Note that each sequence of k outputs maps every point in a set of V^k into the same set of V^0.)

Proof. The proof is by contradiction. By construction $V^{n_\delta-1}$ must satisfy property (a). Suppose some set in $V^{n_\delta-1}$, say $V_1^{n_\delta-1}$, violates property (b). Then there exists $\pi^1, \pi^2 \in V_1^{n_\delta-1}$ such that $T(\pi^1|\theta, \delta) \in V_i^{n_\delta-1}$ but that $T(\pi^2|\theta, \delta) \in V_j^{n_\delta-1}$, $i \neq j$. Since $T(\cdot|\theta, a)$ is a continuous function and since $\delta(\pi^1) = \delta(\pi^2)$ by property (a), there exists some $\pi^* \in V_1^{n_\delta-1}$ such that $T(\pi^*|\theta, \delta)$ lies on a boundary between $V_i^{n_\delta-1}$ and some other set. We claim that $T(\pi^*|\theta, \delta)$ must lie in the set $D^{n_\delta-1}$. To verify this note that $V^{n_\delta-1}$ is constructed such that any set in $V^{n_\delta-1}$, say $V_j^{n_\delta-1}$, maps completely into a set in $V^{n_\delta-2}$. Thus if $T(\pi^1|\theta, \delta)$ lies on a set boundary, it must be a boundary that makes $V^{n_\delta-1}$ different from $V^{n_\delta-2}$, hence a boundary contained in $D^{n_\delta-1}$. Thus we have $T(\pi^1|\theta_1, \delta) \in D^{n_\delta-1}$. Therefore, $T(\pi^1|(\theta_1, \theta_2), \delta) \in D^{n_\delta-2}$ for so ne θ_2 by the definition of $D^{n_\delta-1}$. (The state as a result of a sequence of outputs and stationary policy $(\delta)^\infty$ is defined as $T(\pi|(\theta_1, \theta_2), \delta) = T(T(\pi|\theta_1, \delta)|\theta_2, \delta)$, for two outputs and similarly for more than two successive outputs.) Now for some sequence of n_δ outputs we have by definition $T(\pi^1|(\theta_1 \cdots \theta_{n_\delta}), \delta) \in S_\delta^{n_\delta}$ and also $T(\pi^1|(\theta_1 \cdots \theta_{n_\delta}), \delta) \in D_\delta$; hence $D_\delta \cap S_\delta^{n_\delta} \neq \Phi$, which is a contradiction to the index assumption n_δ. Thus every set in $V^{n_\delta-1}$ maps completely into another set in $V^{n_\delta-1}$ (property (b)) and, therefore, $V^{n_\delta-1}$ is the required partition.

The mapping ν_1 is easily constructed from $V^{n_\delta-1}$. Note that for every point $\pi \in V_j^{n_\delta-1}$, the points $T(\pi|\theta, \delta)$ are contained in the same set $V_k^{n_\delta-1}$, where k depends solely on j and θ. This relationship is given by the function ν such that if $\pi \in V_j$, then $T(\pi|\theta, \delta) \in V_{\nu(j,\theta)}^{n_\delta-1}$.

Note that we have also proved

LEMMA 3. *The policy $(\delta)^\infty$ is finitely transient with $n_\delta = n$ if and only if D^n is the first empty set in the sequence $D_\delta, D^1, D^2, \cdots$.*

Example. This example illustrates a finitely transient policy with index $n_\delta = 5$. The parameters for the example are

$$\begin{array}{lll} & P^a & R^a \\ a=1, & \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix} & \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} \\ a=2, & \begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix} & \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array} \qquad \delta(\pi_1, 1-\pi_1) = \begin{cases} 1, & \pi_1 \leqq 0.7 \\ 2, & \pi_1 > 0.7. \end{cases}$$

The transition diagram and mapping ν for the example are given in Figure 1. The figure is plotted for π_1 and $T_1(\pi_1|\theta, \delta)$, where $\pi=(\pi_1, 1-\pi_1)$ and $T(\pi|\theta, \delta)=(T_1, 1-T_1)$. The set $D_\delta=D^0=[0.7]$ or more precisely $[(0.7, 0.3)]$. The set D^1 is obtained by reflecting $D^0=[0.7]$ as indicated by the arrows. In this way, those points in Π that can reach D^0 in one time period are determined. These points comprise D^1, which is then reflected to obtain D^2. In this instance the set consists of two points. These points are reflected

Set V_j	$\nu(j,\theta)$	
	$\theta=1$	$\theta=2$
1	4	2
2	4	3
3	4	4
4	5	4
5	4	2
6	3	2
7	2	1

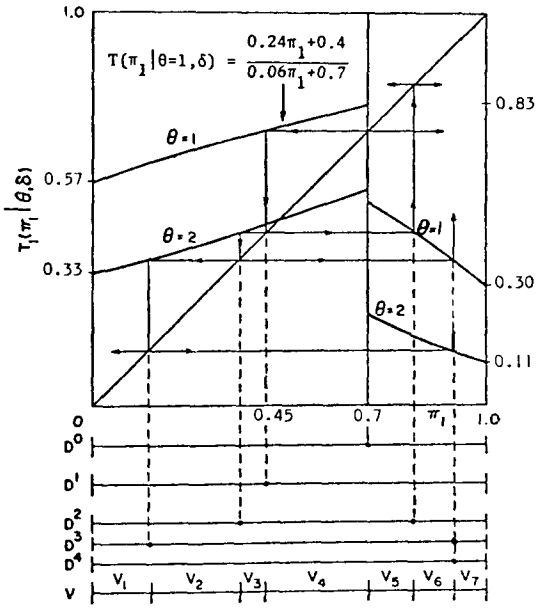


Figure 1. An example of a Finitely Transient Policy with $n_\delta=5$.

to obtain D^3 and D^4. Note that it is impossible to find any point that can reach D^4; that is, $D^5=\Phi$. Thus by Lemma 3, δ^∞ is finitely transient with $n_\delta=5$.

The partition V consists of seven sets and is constructed using the set $D^0 \cup \cdots \cup D^4$ as the set boundaries. The mapping is calculated as follows: Consider any point in V_1. Under outcome $\theta=1$, the point maps into V_4; under outcome $\theta=2$, the point maps to V_2. Thus $\nu(1, 1)=4$ and $\nu(1, 2)=2$. Continuing in this way we obtain the entire mapping ν as shown in the figure.

We next show that $C(\pi|\delta)$ assumes a particularly simple form when $(\delta)^\infty$ is finitely transient.

LEMMA 4. *Let $\delta\in\Delta$; then $C(\pi|\delta)$ can be written as*

$$C(\pi|\delta)=\pi\alpha(\pi|\delta), \tag{5}$$

where $\alpha(\pi|\delta)$ is a column N-vector that is the unique bounded solution to the vector equation

$$\alpha(\pi|\delta)=q^{\delta(\pi)}+\beta\sum_\theta P^{\delta(\pi)}R_\theta^{\delta(\pi)}\alpha[T(\pi|\theta,\delta)|\delta]. \tag{6}$$

Proof. We first note that a unique bounded solution to (6) exists by properties of contraction mappings. If we next multiply each side of (6) on the left by the row vector π, we have

$$\begin{aligned}\pi\alpha(\pi|\delta)&=\pi q+\beta\sum_\theta \pi PR_\theta\alpha[T(\pi|\theta,\delta)|\delta]\\&=\pi q+\beta\sum_\theta\{\theta|\pi,\delta\}[\pi PR_\theta/\{\theta|\pi,\delta\}]\alpha[T(\pi|\theta,\delta)|\delta]\\&=\pi q+\beta\sum_\theta\{\theta|\pi,\delta\}T(\pi|\theta,\delta)\alpha[T(\pi|\theta,\delta)|\delta].\end{aligned}$$

Since the solution to (4) is unique, it follows that $C(\pi|\delta)=\pi\alpha(\pi|\delta)$.

LEMMA 5. *If $(\delta)^\infty$ is a finitely transient policy, the vector function $\alpha(\pi|\delta)$ assumes only a finite number of values over* Π; *furthermore, these values are related by a set of linear equations.*

Proof. Let $(\delta)^\infty$ be a finitely transient policy. From Lemma 2 we can partition the state space into sets $V_1,\cdots,V_m$ with properties (a) and (b). Now if we define $\alpha(\pi|\delta)=\alpha_j$, for $\pi\in V_j$, and select any set of m points $\pi^i\in V_i$, $1\leqq i\leqq m$, then (6) becomes the following set of vector equations

$$\alpha_i=q^{\delta(\pi^i)}+\beta\sum_\theta P^{\delta(\pi^i)}R_\theta^{\delta(\pi^i)}\alpha_{\nu(i,\theta)},\qquad 1\leqq i\leqq m. \tag{7}$$

Recalling that $\delta_i\equiv\delta(\pi)$, $\pi\in V_i$ we have

$$\alpha_i=q^{\delta_i}+\beta\sum_\theta P^{\delta_i}R_\theta^{\delta_i}\alpha_{\nu(i,\theta)}. \tag{8}$$

This set of linear equations can then be solved, uniquely, for $\alpha_1,\cdots,\alpha_n$.

As a simple example of Lemma 5, note that if $\delta(\pi)$ is constant and we assume that $\alpha(\pi|\delta)=\alpha$ for every π, then substitution into (4) yields $\alpha=q+\beta P\alpha$ or $\alpha=[I-\beta P]^{-1}q$. Thus $C(\pi|\delta)=\pi\alpha$. Clearly, this is also the expected cost of the completely observable process with a probabilistic initial state denoted by π. Note that α_i is the support function of $C(\pi|\delta)$; $C(\pi|\delta)$ has the same gradient for all $\pi\in V_i$.

THEOREM 2. *If a policy $(\delta)^\infty$ is finitely transient, then $C(\pi|\delta)$ is piecewise linear.*

Proof. Follows immediately from Lemmas 4 and 5.

It might be conjectured that the converse to Theorem 2 is true; that is, every policy $(\delta)^\infty$ with a piecewise linear cost function is finitely transient. This conjecture is false and a counter-example is presented in [15, Appendix B].

4. THE APPROXIMATION OF $C(\pi|\delta)$

Since all policies are not finitely transient, their costs cannot be computed as neatly as expressed by the solution of (8). The cost of an arbitrary

stationary policy is an important element of the policy iteration algorithm we seek. In this section we present a bound on the difference between $C(\pi|\delta)$ and the cost of an approximation based on our study of finitely transient processes.

The Approximation Mapping $\hat{\nu}$

In the preceding section the partition V was calculated by first finding the sets D_δ, D^1, D^2, $\cdots$ and then using these sets as boundaries of the sets V_j in the partition. Suppose that for some k the sequence D_δ, $\cdots$, D^k is found for an arbitrary policy $(\delta)^\infty$. If $D^k=\Phi$ then it follows from Lemma 3 that $(\delta)^\infty$ is finitely transient with degree $n_\delta \leqq k$. If D^k is nonvoid the question of whether $(\delta)^\infty$ is finitely transient is unanswered. Let the boundaries of the sets in the partition of Π, $V^k=[V_j^k]$ be formed by the set $D_\delta \cup D^1 \cup \cdots \cup D^k$ as for a finitely transient policy.

Using V^k we now construct a mapping $\hat{\nu}$ that will be used to approxi nate $(\delta)^\infty$. Since $\hat{\nu}$ will be constructed from V^k, the integer k will be called the *degree of the approximation*. We arbitrarily select any point π^j in each set V_j^k. Then the mapping $\hat{\nu}$ is defined as follows

$$\text{if } T(\pi^j|\theta, \delta) \in V_l^k, \quad \text{then } \hat{\nu}(j, \theta)=l. \tag{9}$$

Note that the mapping is similar in form to the Markov-mapping induced by a finitely transient process; but since V^k does not satisfy property (b) there is some set V_j^k, output θ, and $\pi \in V_j^k$ such that $T(\pi|\theta, \delta) \notin V_{\nu(j,\theta)}^k$.

The mapping $\hat{\nu}$ will now be used to construct a piecewise linear approximation to $C(\pi|\delta)$, which we shall see later has a bound on error proportional to β^k. This approximation to $C(\pi|\delta)$, denoted $\hat{C}(\pi|\delta)$, is defined by

$$\hat{C}(\pi|\delta)=\pi\hat{\alpha}_j, \quad \pi \in V_j^k, \tag{10}$$

where the vectors $\hat{\alpha}_j$ are chosen to satisfy the set of linear equations

$$\hat{\alpha}_j=q^{\delta_j}+\beta \sum_\theta P^{\delta_j} R_\theta^{\delta_j} \hat{\alpha}_{\hat{\nu}(j,\theta)} \tag{11}$$

where $\delta_i=\delta(\pi)$ for $\pi \in V_j^k$. (By elementary properties of contraction mappings, the solution of (11) exists for $0 \leqq \beta < 1$.)

An Error Bound on the Difference between $C(\pi|\delta)$ and $\hat{C}(\pi|\delta)$

The bound on the difference between $C(\pi|\delta)$ and $\hat{C}(\pi|\delta)$ depends on the following observation based on the construction of V^k. First we need some additional notation. If $\pi \in V_j^k$, $(\theta_1, \cdots, \theta_n)$ is any sequence of outputs of length $n \leqq k$ and $\hat{\nu}$ is constructed from a partition of degree k, then

$$\delta[T(\pi|(\theta_1, \cdots, \theta_n), \delta)]=\delta(\hat{\pi}), \tag{12}$$

where

$$\hat{\pi} \in V_{\hat{\nu}(j,(\theta_1,\ldots,\theta_n))} \tag{13}$$

and $\hat{p}(j, (\theta_1, \cdots, \theta_n))$ is defined as the composition of $\hat{p}(j, \theta)$ n times such that

$$\hat{p}(j, (\theta_1, \cdots, \theta_n)) = \hat{p}[\cdots \hat{p}[\hat{p}(j, \theta_1), \theta_2], \cdots, \theta_n]. \tag{14}$$

We can now derive a bound on the difference between $C(\pi|\delta)$ and $\hat{C}(\pi|\delta)$.

THEOREM 3. *If $\hat{p}$ is constructed from a partition of degree k, V^k, then*

$$\|C(\cdot|\delta) - \hat{C}(\cdot|\delta)\| \leqq [\beta^k/(1-\beta^k)]K/(1-\beta), \tag{15}$$

where $K = \max_{a,i} q_i^a - \min_{a,i} q_i^a$.

Proof. From (11) $\hat{C}(\pi|\delta)$ can be written for $\pi \in V_j^k$ as

$$\begin{aligned} \hat{C}(\pi|\delta) = \pi\hat{\alpha}_j &= \pi[q^{\delta_j} + \beta \textstyle\sum_\theta P^{\delta_j} R_\theta^{\delta_j} \hat{\alpha}_{\hat{p}(j,\theta)}] \\ &= \pi q^{\delta_j} + \beta \textstyle\sum_\theta \{\theta|\pi, \delta\} T(\pi|\theta, \delta)\hat{\alpha}_{\hat{p}(j,\theta)} \end{aligned} \tag{16}$$

where δ_j is the control alternative specified by δ over V_j^k. Now by iterating (16) and (4) m times for $m \leqq k$, we may write

$$\begin{aligned} C(\pi|\delta) - \hat{C}(\pi|\delta) = \beta^m [&\textstyle\sum_{\theta_1,\ldots,\theta_m} \{(\theta_1, \cdots, \theta_m)|\pi, \delta) \\ &\cdot [C[T(\pi|(\theta_1, \cdots, \theta_m), \delta)|\delta] \\ &- T(\pi|(\theta_1, \cdots, \theta_m), \delta)\hat{\alpha}_{\hat{p}(j,(\theta_1,\ldots,\theta_m))}]]. \end{aligned}$$

Defining $\mu(\pi)$ as the index of the set in V^k containing π, we add and subtract the quantity

$$\textstyle\sum_{\theta_1,\ldots,\theta_m} \{(\theta_1, \cdots, \theta_m)|\pi,\delta\} T(\pi|(\theta_1, \cdots, \theta_m),\delta)\hat{\alpha}_{\mu[T(\pi|(\theta_1,\ldots,\theta_m),\delta)]}$$

and rearrange terms to yield

$$(1-\beta^m)\|C(\cdot|\delta) - \hat{C}(\cdot|\delta)\| \leqq \beta^m \sup_{i,j,\pi\in\Pi} |\pi(\hat{\alpha}_i - \hat{\alpha}_j)|.$$

From (16) it follows that $\sup_{i,j,\pi\in\Pi} \pi(\hat{\alpha}_i - \hat{\alpha}_j) < K/(1-\beta)$. Combining this equation with the previous equation completes the proof.

The construction of $\hat{p}$ is discussed in detail in [16].

5. POLICY IMPROVEMENT

We now turn to the policy improvement operation to develop an algorithm to improve a stationary policy.

Use of the Concave Hull of $C(\pi|\delta)$ in Policy Improvement

The general proof of policy i nprovement (Theorem 1) shows that the cost of the stationary policy $(\delta')^\infty$ found by minimizing $U_a(\pi, f)$ is less than f, where f is the cost of some policy. In this section we will prove Theorem 4, which states that the concave hull of $C(\pi|\delta)$ (a continuous function) can be used in policy improvement in conjunction with the meth-

ods of [14] dealing with piecewise linear functions. The proof of this assertion depends on the fact that the concave hull of $C(\pi|\delta)$ is the cost of a policy that is, in general, nonstationary.

Strauch [17] has defined a class of policies called semi-Markov. Briefly, a *semi-Markov policy* is a sequence of functions $(\xi^1, \xi^2, \cdots)$ where $\xi^t: \Pi \times \Pi \rightarrow A$ and $\xi^t(\pi^0, \pi)$ is a control to be used at the tth time period of system operation if the state at time t is π and the initial state was π^0. It will be shown that the concave hull of $C(\pi|\delta)$ is the cost function of a semi-Markov policy and thus can be used in policy improvement. We now make these notions precise.

DEFINITION. *The "concave hull" of* $C(\pi|\delta)$ *is denoted* $\bar{C}(\pi|\delta)$ *and defined as* $\bar{C}(\pi|\delta) = min_{\pi'} \pi\alpha(\pi'|\delta)$, *where* $\alpha(\pi|\delta)$ *satisfies* (6).

An example of the concave hull of a function is given in Figure 2 for the case $N=2$. In the example $C(\pi|\delta)$ consists of three linear discontinuous segments. The vector function $\alpha(\pi|\delta)$ for $\frac{1}{4} \leqq \pi_1 < \frac{1}{2}$ is simply (8, 4) We note that the function $\alpha(\pi|\delta)$ is essentially the support hyperplane of $C(\pi|\delta)$; and, consequently, $\bar{C}(\pi|\delta)$ is formed from these hyperplanes. (If $C(\pi|\delta)$ were not linear, we would simply draw the tangent hyperplanes at each point and take the minimum hull; however, the algorithm developed in Section 6 need consider only piecewise linear functions.) In the remainder of this paper we assume that $C(\pi|\delta)$ is piecewise linear and consists of a finite number of linear segments, a structure that is sufficient to develop the optimization algorithm.

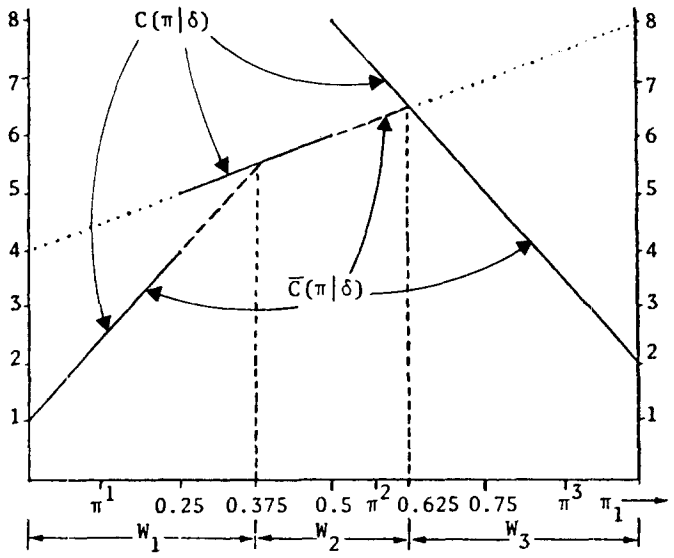
THEOREM 4. *If* $\bar{C}(\pi|\delta)$ *is substituted for* $C(\pi|\delta)$ *in* (3), *then* $C(\pi|\delta') \leqq C(\pi|\delta)$ *for every* $\pi \in \Pi$, *where* δ' *is defined as in Theorem* 1.

Proof. If $\bar{C}(\pi|\delta)$ is the cost of some policy and if $\bar{C}(\pi|\delta)$ is substituted for $C(\pi|\delta)$ in (1), then by a generalization of Theorem 1 (Blackwell [1]), $C(\pi|\delta') \leqq \bar{C}(\pi|\delta)$ and by definition, $\bar{C}(\pi|\delta) \leqq C(\pi|\delta)$. Thus we must show that a policy exists for which the expected cost is $\bar{C}$.

The function $\alpha(\pi|\delta)$ of (6) assumes a finite number of values α_i satisfying (8). We define a collection of mutually exclusive and exhaustive sets $W_1, W_2, \cdots$ and one point in each set $\pi^i \in W_i$ such that $W_i = [\pi: \min_j \pi\alpha_j = \pi\alpha_i]$. Assignment of π^i on the set W_i is arbitrary (see Figure 2). Now consider a semi-Markov policy $(\xi^1, \xi^2, \cdots)$ defined as follows: If the initial state π^0 is such that $C(\pi^0|\delta) = C(\pi|\delta)$, then $\xi^t(\pi^0, \pi) = \delta(\pi)$. If π^0 is such that $C(\pi^0|\delta) \neq \bar{C}(\pi^0|\delta)$ and that $\pi^0 \in W_i$, then the control at time t after output sequence $(\theta_1, \cdots, \theta_t)$ is observed is the same control as if the process had started in state $\pi^i \in W_i$ and the same output sequence had occurred.

The semi-Markov policy $(\xi^1, \xi^2, \cdots)$ has expected cost $\bar{C}(\pi|\delta)$. To prove this fact, we note that the expected cost of beginning in π^0 such that

$C(\pi^0|\delta) = \bar{C}(\pi^0|\delta)$ is simply $\bar{C}(\pi^0|\delta)$. On the other hand, if $C(\pi^0|\delta) \neq \bar{C}(\pi^0|\delta)$ and $\pi^0 \in W_j$, the expected cost of the policy is simply the expected cost of choosing alternatives as if the process began in state π^j. Thus the expected



$$C(\pi|\delta) = \begin{cases} (1-\pi_1) + 13\pi_1 & 0 \leq \pi_1 < 1/4 \\ 4(1-\pi_1) + 8\pi_1 & 1/4 \leq \pi_1 < 1/2 \\ 1/4(1-\pi_1) + 2\pi_1 & 1/2 \leq \pi_1 \leq 1 \end{cases}$$

$$\bar{C}(\pi|\delta) = \begin{cases} (1-\pi_1) + 13\pi_1 & 0 \leq \pi_1 \leq 3/8 \\ 4(1-\pi_1) + 8\pi_1 & 3/8 \leq \pi_1 \leq 5/8 \\ 1/4(1-\pi_1) + 2\pi_1 & 5/8 \leq \pi_1 \leq 1 \end{cases}$$

Figure 2. Example of $C(\pi|\delta)$ and $\bar{C}(\pi|\delta)$ for $N = 2$.

cost of this semi-Markov policy is given by

$$\pi^0\alpha(\pi^j|\delta) = \pi^0 q^{\delta(\pi^j)} + \beta \sum_{\theta_1} \{\theta_1|\pi^0, \xi^1\} T(\pi^0|\theta_1, \xi^1) q^{\delta(T(\pi^j|\theta_1,\delta))}$$
$$+ \beta^2 \sum_{\theta_1,\theta_2} \{(\theta_1, \theta_2)|\pi^0, \xi^1, \xi^2\} \cdot T(\pi^0|(\theta_1, \theta_2), \xi^1, \xi^2) \cdot q^{\delta(T(\pi^j|(\theta_1,\theta_2),\delta))} + \cdots$$

where $\{\theta_1|\pi^0, \xi^1\} = \pi^0 P^{\delta(\pi^j)} R_{\theta_1}{}^{\delta(\pi^j)}\,\mathbf{1}$, etc. Iteratively expanding $\alpha(\pi^j|\delta)$

as follows

$$\alpha(\pi^j|\delta) = q^{\delta(\pi^j)} + \beta \sum_{\theta_1} P^{\delta(\pi^j)} R_{\theta_1}^{\delta(\pi^j)} q^{\delta(T(\pi^j|\theta_1,\delta))}$$
$$+ \beta^2 \sum_{\theta_1,\theta_2} P^{\delta(\pi^j)} R_{\theta_1}^{\delta(\pi^j)} P^{\delta(T(\pi^j|\theta_1,\delta)} R_{\theta_2}^{\delta\ T(\pi^j|\theta_1,\delta)} q^{\delta(T(\pi^j|(\theta_1,\theta_2),\delta))} + \cdots,$$

we see the expected cost of the semi-Markov policy is $\pi^0\alpha(\pi^j|\delta) = \bar{C}(\pi^0|\delta)$. Thus, for any π^0 the expected cost of the semi-Markov policy is $\bar{C}(\pi|\delta)$, from which the theorem follows.

In the proof $\bar{C}(\pi|\delta)$ need not be piecewise linear; however, piecewise linearity allows us to represent C rather simply by $\bar{C}(\pi|\delta) = \min_l \pi\alpha_l$, where $[\alpha_1, \alpha_2, \cdots]$ is some set of vectors. If so, then the methods of [14] can be used to implement the policy improvement function $\min_a U_a(\pi, \bar{C})$. As discussed extensively in [14], changes in δ determined by the minimization operation will occur only at those points in Π where there are intersections of the linear segments. Thus, the complete space Π need not be searched, a major saving in time for many-state problems.

Use of $\bar{\hat{C}}$ in Policy Improvement

The use of the concave hull of C in policy improvement leads us to believe that the concave hull of an approximation to C (denoted $\bar{\hat{C}}$) may lead to improved policies if the approximation is sufficiently close to C. In this section we establish that policy improvement based on $\bar{\hat{C}}$ does occur by first showing that $\hat{C}(\cdot|\delta)$ represents the expected cost of a policy based on selecting controls according to the output sequences. It is then easy to show that $\bar{\hat{C}}(\cdot|\delta)$ represents the expected cost of a semi-Markov policy based on the output sequence.

In Section 4 we developed $\hat{\nu}$, a kth degree approximation mapping to $(\delta)^\infty$. We denoted the cost of this policy as $\hat{C}(\pi|\delta) = \pi\hat{\alpha}_j$ for $\pi \in V_j^k$ where V_j^k is a set in the kth partition of Π, induced by δ. Recall that for $\pi \in V_j^k$, $\delta(\pi) \equiv \delta_j$. We may define a semi-Markov policy based on the approximation $\hat{\nu}$ as follows: The initial alternative is chosen as δ_j if j minimizes $\pi\hat{\alpha}_j$, regardless of the set in V^k in which π may lie. Succeeding alternatives follow according to $\hat{\nu}(j, \theta)$. The following lemma describes the cost of this semi-Markov policy. Its proof follows the proof of Theorem 4.

LEMMA 6. *The expected cost of following the semi-Markov policy described above based on the kth degree approximation mapping $\hat{\nu}$ is $\bar{\hat{C}}(\cdot|\delta)$.*

We now show that the piecewise linear function $\bar{\hat{C}}(\cdot|\delta)$ can be used effectively in place of $C(\cdot|\delta)$ in policy improvement.

THEOREM 5. *If $\bar{\hat{C}}(\cdot|\delta)$ is used in policy improvement in place of $C(\cdot|\delta)$, then $C(\cdot|\delta')$, the expected cost of the resultant policy $(\delta')^\infty$, satisfies for all $\pi \in \Pi$*

$$C(\pi|\delta') \leqq \bar{\hat{C}}(\pi|\delta) \leqq \hat{C}(\pi|\delta) \leqq C(\pi|\delta) + \epsilon_k, \tag{17}$$

where ϵ_k depends on the degree of the approximation to $(\delta)^\infty$ and ϵ_k approaches zero as $k \to \infty$.

Proof. The right-hand inequality follows directly from Theorem 3. The next inequality follows from the definition of concave hull. The last inequality $C(\pi|\delta') \leqq \bar{\bar{C}}(\pi|\delta)$ follows from Theorem 1 and Lemma 6.

We now present a bound on the distance of $\bar{\bar{C}}$ from C^*.

THEOREM 6.

$$\| \bar{\bar{C}}(\cdot|\delta) - C^*(\cdot) \| \leqq (1-\beta)^{-1} \| \bar{\bar{C}}(\cdot|\delta) - \min_a U_a[\cdot, \bar{\bar{C}}(\cdot|\delta)] \|. \quad (18)$$

Proof. The proof of (18) follows directly from properties of contraction mappings in [3], noting that $\bar{\bar{C}}(\pi|\delta) \geqq C^*(\pi)$ for all $\pi \in \Pi$ and that $U_a[\pi, \bar{\bar{C}}(\pi|\delta)] \geqq C^*(\pi)$ for all a, and for all $\pi \in \Pi$.

6. THE POLICY ITERATION ALGORITHM

We now describe the policy iteration algorithm shown schematically in Figure 3. The central idea of the algorithm is to iterate through a succession of approximations to stationary policies, using the expected cost of each approximation policy as a basis for policy improvement. This algorithm differs from the one implied by Theorem 1 by performing policy improvement with the easily computed expected costs of approximations to stationary policies instead of the cost functions of stationary policies. These approximations are given by the mappings $\hat{\nu}$, where $\hat{\nu}$ is created from a partition of degree k induced by a stationary policy $(\delta)^\infty$. (If the policy is finitely transient with degree less than k, then the mapping $\hat{\nu}$ equals ν, the Markov-mapping equivalent to the finitely transient policy.)

At Step 1 a bound on the distance of the expected cost of the previous approximation policy is given by ϵ. The desired degree of accuracy of the final solution (times $(1-\beta)$) is labeled ϵ^*. The inner circuit (Steps 2 through 7) finds a new approximation policy that is at least as good as the previous approximation policy. (We show below that this is always possible.) If the new approximation policy is not within the bound ϵ^*, then the bound ϵ is tightened (to $\hat{\epsilon}$ computed in Step 6) and the algorithm returns to Step 1 for the development of another approximation policy that will be closer than $\hat{\epsilon}$ to the optimal cost. If $\hat{\epsilon} \leqq \epsilon^*$, then the algorithm stops and the approximation policy is within the desired degree of accuracy.

THEOREM 7. *The policy iteration algorithm defined in Figure 3 converges to a policy arbitrarily close to the optimal policy.*

Proof. We need only show that Steps 2 through 7 develop an approximation policy, say $\hat{\nu}$, better than the previous approximation policy, say $\hat{\nu}^0$. We note that in the last passage through Step 8, we replace policy δ_0 (used to calculate $\hat{\nu}^0$) by δ (found in Step 5 and there labeled as δ'). Theo-

rem 5 assures us that $C(\pi|\delta) \leqq \bar{C}(\pi|\delta^0) \forall \pi \in \Pi$. Theorem 3 guarantees we can traverse the inner circuit a sufficient number of times so that $\hat{p}$ de-

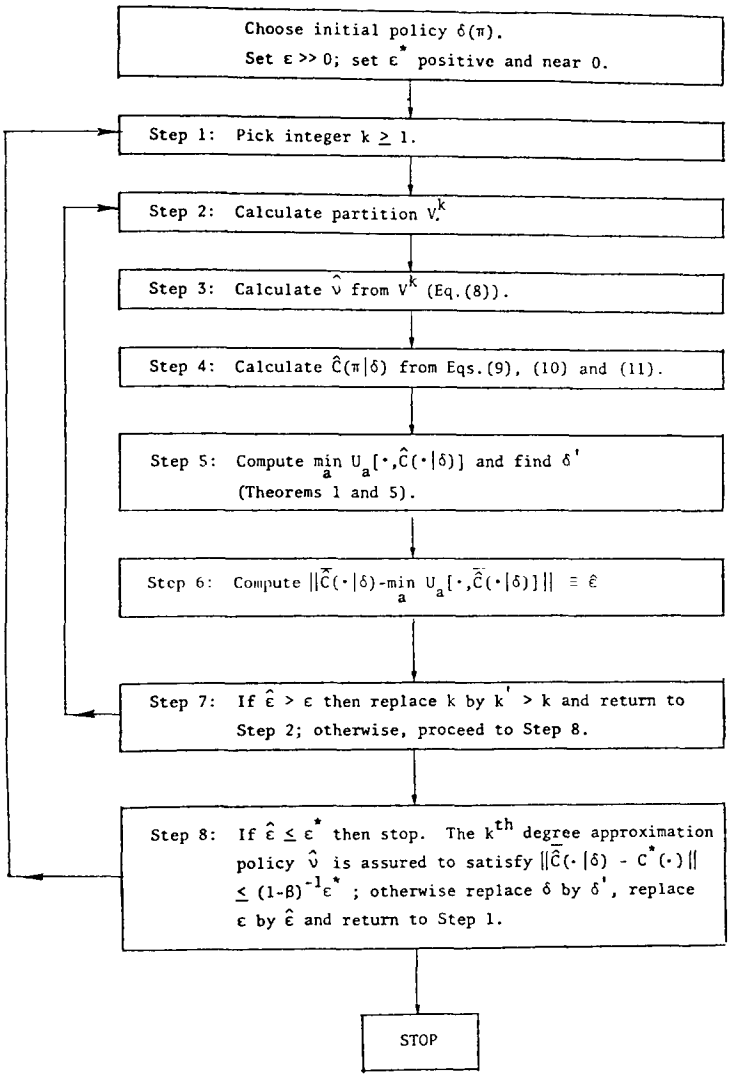


Figure 3. The policy iteration algorithm.

veloped from $(\delta)^\infty$ will have expected cost as close as desired to $C(\pi|\delta)$. Thus, we guarantee that $\bar{C}(\pi|\delta) \leqq \bar{C}(\pi|\delta^0)$, or equivalently, that $\hat{p}$ is closer to optimal than is $\hat{p}^0$.

The quantity k in Step 1 should be chosen as small as practical to re-

duce the computation of V^k. In practice, values of k near 1 early in the algorithm have been found to be quite practical.

Discussion

In one sense the algorithm is similar to the grid approximations to Π of Eckles [5] and Kakalik [7]. Such methods, however, are difficult if not impossible to use for $N \geqq 4$. In our case the partition V^k can be considered to represent a set of points and the vectors $\hat{\alpha}_j$ each represent a vector cost over the states. Computation is minimized since the accuracy of the approximations needs to be increased (corresponding to traversing the inner loop in Figure 3) only when policy improvement does not occur.

Brown [2] has proposed a recursive set of control "rules" based on finite sequences of past outputs and controls (finite histories). He also describes an iterative procedure for refining the set to an optimal policy. The procedure basically requires a search of all points in Π. His rules are similar to our kth degree approximations to stationary policies defined over π, the state of information. Our development shows the relationship between such rules (based on finite sequences of outputs and controls) to controls based on π. In particular, policies based on finite histories can be equivalent in expected cost to stationary policies based on π; in such cases the policies are finitely transient.

We have programmed the algorithm and solved several problems, including the example in the next section. Computation time for the example was on the order of a few seconds. A discounted version of the machine maintenance problem in [14] was solved to within a bound of 10% of the optimal expected cost in about 6 seconds, less time than for an equivalent finite horizon solution. All times are on the Stanford Computation Center IBM 360/67. For this discounted problem we conjecture that a minimum computation procedure is a combination of successive value iterations for policy improvement followed by a policy iteration cycle to determine optimality.

Although the example is a two-state problem, the algorithm can be applied to the case $N>2$ by using the results in [16]. It seems that the algorithm is most suited to problems with either few states and many alternatives and outputs, or few alternatives and outputs and many states. The characteristics of transient processes and sparse matrices may lead to simplifications in the algorithm for such special cases.

7. EXAMPLE

In this section we apply the algorithm of Figure 3 to solve a two-state, two-alternative, and two-output problem. Each alternative changes the state transition matrix P, the output matrix R, and the cost vector q.

The problem parameters are given in Table I. This problem might represent a marketing problem where the states are attitudes for purchase of a particular brand; these attitudes are translated to actual purchases through

TABLE I

THE PARAMETERS FOR THE EXAMPLE

Control a	p^a	R^a	q^a
1	$\begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix}$	$\begin{bmatrix} 4 \\ -4 \end{bmatrix}$
2	$\begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 0 \\ -3 \end{bmatrix}$

the R matrix where the outputs are purchases. A similar model dependent on partially observable attitudes has been proposed by Lipstein [8].

Defining $\pi = (\pi_1, 1-\pi_1)$ the optimal policy can be shown to be (i) use Control 1 if $\pi_1 < 0.1188$, or (ii) use Control 2 if $\pi_1 \geqq 0.1188$. The total expected discounted cost of using this policy is shown in Figure 4.

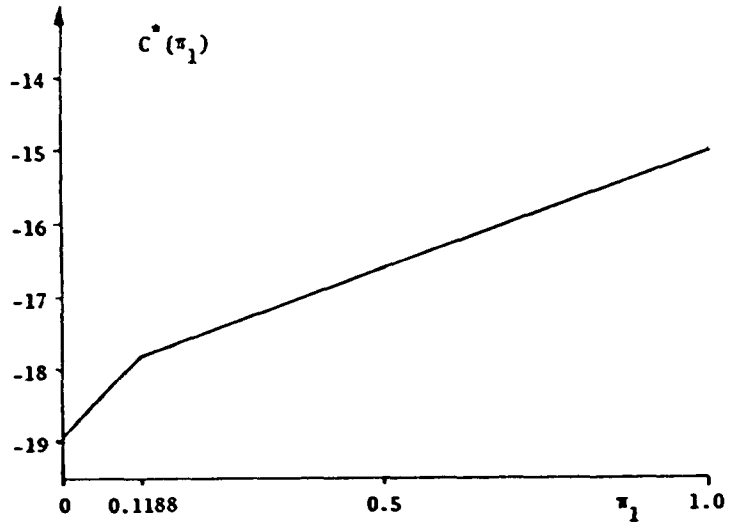


Figure 4. The optimal cost $C^*(\pi_1)$ for the example.

To use this control, the controller must update π_1 after each time period using the observed output. Another form of this control in which these calculations need not be performed is represented by the block diagram of Figure 5. Each block corresponds to a range of π_1 and also to some control alternative. The changes between the blocks are controlled entirely by the

output records. Thus, after an initial determination of π_1, the controller need use only his output record and the current block in the diagram to choose his strategy optimally.

The form of the control in Figure 5 is not an approximation; if the block diagram is followed, the minimum expected cost will be achieved. This result is a consequence of the fact that this control policy is finitely transient.

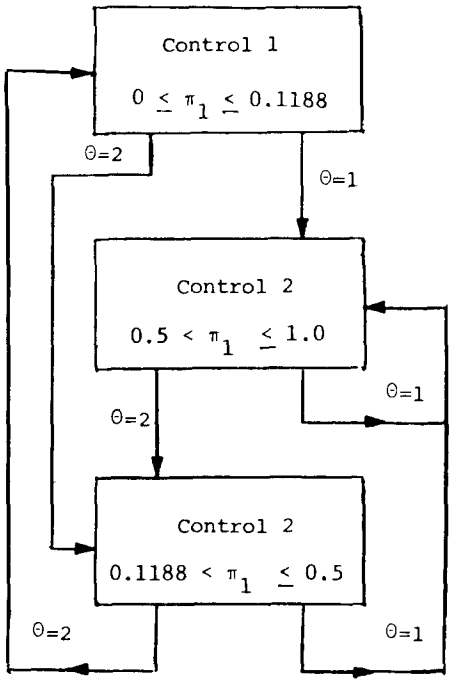


Figure 5. Block diagram of the optimal control for the example.

Solution of the Example

The optimal control can be found in three iterations. We begin by choosing a stationary policy. We arbitrarily choose this policy as that policy minimizing the expected immediate costs of operating the process. This policy $(\delta^0)^\infty$ is simply the alternative a minimizing πq^a; thus, $\delta^0(\pi)=1$ if $\pi_1 \leqq 0.2$ and $\delta^0(\pi)=2$ if $\pi_1 > 0.2$. When we construct V^0 as shown in Figure 6, we find this policy to be finitely transient with index 1. Solving (6) we find $\bar{C}(\pi|\delta^0)$ to be given by $\bar{C}(\pi|\delta^0)=\min_i \pi\alpha_i$, where $\alpha_1=(-9.86, -18.78)^T$ and $\alpha_2=(-14.76, -18.14)^T$. We note that $\bar{C}(\pi|\delta^0) \neq C(\pi|\delta^0)$.

Performing policy improvement we find policy $(\delta^1)^\infty$ where $\delta'(\pi)=1$ if

$\pi_1 \leqq 0.1155$ and $\delta'(\pi) = 2$ if $\pi_1 > 0.1155$. In addition, we find from Theorem 6 that $\|\bar{C}(\pi|\delta^0) - C^*(\pi)\| \leqq 0.4$.

We proceed with the algorithm returning to Step 1 to pick k and calcu-

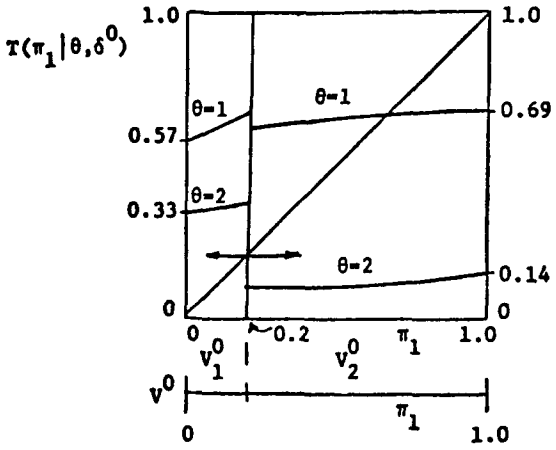


Figure 6. The transition diagram for policy $(\delta^0)^\infty$ of the example.

late V^k for $(\delta^1)^\infty$. Choosing $k=1$ we find $(\delta^1)^\infty$ to be finitely transient with index 2. Policy improvement and Theorem 6 show that $\|\bar{C}(\pi|\delta^1) - C^*(\pi)\| = 0$; thus, $\bar{C}(\cdot|\delta^1) = C^*$. Now, by Theorem 1 there must be a stationary policy $(\delta^*)^\infty$ such that $C(\cdot|\delta^*) = C^*$. We find this policy by

TABLE II

THE MAPPING ν

j	θ	$\nu^*(i,\theta)$	$\delta_j^* = \delta^*(\pi);\ \pi \epsilon V_j^1$	V_j^1
1	1	3	1	$0 \leqq \pi_1 \leqq 0.1188$
1	2	2	1	
2	1	3	2	$0.1188 < \pi_1 \leqq 0.5$
2	2	1	2	
3	1	3	2	$0.5 < \pi_1 \leqq 1.0$
3	2	2	2	

performing policy improvement once more (based on $\bar{C}(\pi|\delta^1)$, which yields $\delta^*(\pi) = 1$ if $\pi_1 \leqq 0.1188$ and $\delta^*(\pi) = 2$ if $\pi_1 > 0.1188$.

This optimal policy turns out to be finitely transient with index 2, the partition V^1 consists of three sets and the mapping ν^* is given in Table II.

The block diagram form of the optimal control of Figure 5 fol-

lows from the mapping ν^*, which is equivalent to $(\delta^*)^\infty$ as discussed in an earlier section under finitely transient policies. Using ν^* and (6) we find $C^*(\pi) = \min_i \pi\alpha_i^*$, where $\alpha_1^* = (-10.03, -18.93)^T$, $\alpha_2^* = (-14.89, -18.27)^T$, $\alpha_3^* = (-14.93, -18.23)^T$.

ACKNOWLEDGMENT

The author most gratefully acknowledges the guidance and encouragement of R. D. Smallwood during the course of this work. Also worthy of acknowledgment is the extensive, very helpful review by one of the referees. His many constructive comments are incorporated throughout this paper. This research was partially supported by the Joint Services Electronics Program through Stanford University.

REFERENCES

1. D. Blackwell, "Discounted Dynamic Programming," *Ann. Math. Stat.* **36,** 226–235 (1965).
2. G. W. Brown, "Recursive Sets of Rules in Statistical Decision Processes," in *Statistical Papers in Honor of George W. Snedecor*, pp. 59–75, University of Iowa Press, Ames, Iowa, 1972.
3. E. V. Denardo, "Contraction Mappings in the Theory Underlying Dynamic Programming," *SIAM Rev.* **9,** 165–177 (1967).
4. A. W. Drake, Observation of a Markov Process through a Noisy Channel. Sc.D. thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1962.
5. J. E. Eckles, Optimum Replacement of Stochastically Failing Systems. Ph.D. thesis, Department of Engineering-Economic Systems, Stanford University, Stanford, Calif., September 1966.
6. R. A. Howard, *Dynamic Probabilistic Systems*, John Wiley & Sons, New York, 1971.
7. J. S. Kakalik, "Optimum Policies for Partially Observable Markov Systems," Technical Report TR-18, Operations Research Center, Massachusetts Institute of Technology, Cambridge, Mass., October 1965.
8. B. Lipstein, "A Mathematical Model of Consumer Behavior," *J. Marketing Res. II*, 259–265 (1965).
9. A. Maitra, "Discounted Dynamic Programming on Compact Metric Spaces," *Sankhya, Ser. A*, **30,** 211–216 (1968).
10. J. E. Matheson, Optimum Teaching Procedures Derived from Mathematical Learning Models. Ph.D. thesis, Department of Engineering-Economic Systems, Stanford University, Stanford, California, August 1964.
11. S. M. Ross, "Arbitrary State Markovian Decision Processes," *Ann. Math. Stat.* **39,** 2118–2122 (1968).
12. J. K. Satia, Markovian Decision Processes with Uncertain Transition Matrices or/and Probabilistic Observation of States. Ph.D. thesis, Department of Industrial Engineering, Stanford University, Stanford, Calif., March 1968.
13. R. D. Smallwood, "The Analysis of Economic Teaching Strategies for a Simple Learning Model," *J. Math. Psychol.* **8,** 285–301 (1971).

14. R. D. SMALLWOOD AND E. J. SONDIK, "Optimal Control of Partially Observable Processes over the Finite Horizon," *Opns. Res.* **21,** 1071–1088 (1973).
15. E. J. SONDIK, The Optimal Control of Partially Observable Markov Processes. Ph.D. thesis, Department of Electrical Engineering, Stanford University, Stanford, Calif., May 1971.
16. E. J. SONDIK, "Construction of $\hat{v}$ for $N \geq 2$," Internal Memorandum, Department of Engineering-Economic Systems, Stanford University, Stanford, Calif., April 1976.
17. R. STRAUCH, "Negative Dynamic Programming," *Ann. Math. Stat.* **37,** 871–890 (1966).