Southwest Fisheries Center Administrative Report H-79-13

Subject to revision
File copy

Information Seeking in Markov Decision Processes

EDWARD J. SONDIK

National Heart and Lung Institute, National Institute of Health

Bethesda, MD 20014

ROY MENDELSSOHN

Southwest Fisheries Center, National Marine Fisheries Service, NOAA

Honolulu, HI 96812

In many decision and control problems the system's state can be observed only at considerable cost. Typical examples include: The control of a population of hypertensives in which the variables of interest are the awareness, attitudes, knowledge, and control status of the population [1], [2], [3]; machine inspection and replacement [4], [5], [6]; and the example which prompted this work, the management of a randomly varying fishery [7], [8], [9]. These problems share the common characteristics of Markovian-like dynamics with incompletely known state information. This paper derives efficient policy and value iteration algorithms for computing optimal control policies for such problems. The major assumptions in the formulation are that the underlying dynamic process is Markovian (discrete time) and that the state dynamics process is obscured between information seeking actions, which we term surveys. We also assume the surveys in general provide perfect information on the state of the process.

The primary problem which motivated this research relates to the management of a fishery resource. In essence, the problem is to outline either the fishing effort or the maximum permissible catch in a time period (1 year). The state of the fishery at a given time can be defined as the number of fish present or more generally, as the weight of the fish present, the biomass. The general relationship is that too large a catch in a period of low biomass can seriously deplete the resource--even eliminate it. The biomass can be accurately measured by surveying each year before the fishing season commences. The problem we address is how frequently should we survey?

This paper derives algorithms for two problems. The first problem is to determine the optimal survey interval s, $s \geq 1$, and the optimal control for the system as a function of some initial state of information. The second more general problem is to derive controls and survey times based on the time since a given state was last observed.

The algorithms are efficient in the sense that they outperform the naive approach to the problem. For a problem with N states and the same A actions per state, a simple approach is to take the s-fold Cartesian product of the action space as the new action space, and redefine the expected one-period reward as the expected reward from choosing any s-vector of actions. This reformulated problem is a completely observed Markov decision process that requires a search over $N \cdot A^s$ alternatives for each iteration of a successive approximation algorithm. The algorithms proposed in this paper require a search over at most $N \cdot A \cdot s$ alternatives and often much fewer alternatives than that. For example, in [10], a problem is solved with 25 states, 26 actions, and $s = 4$. For the naive approach, this would require a search over 11,424,400 alternatives each iteration, while the algorithms of this paper required a search over far fewer than the 2600 alternatives per iteration upper bound.

1. ANALYSIS AND THE BASIC ITERATION ALGORITHM

In this section we analyze the cost of a given policy and derive the value iteration and policy iteration algorithms. We assume the system can be described by an N state discrete time Markov process. Whenever the process is observed perfect information on the state is obtained, but the observations need not be taken each time period. The process is

controlled by choosing one of A alternatives each time period with the a^{th} alternative represented by an N x N state transition matrix P(a) and a reward vector $\gamma(a)$. The observation or survey process is assumed to take place instantaneously at the beginning of a time period, before a control alternative has been selected for that time period. We consider that the system will be observed every s time units using a survey which costs C(s). Costs are discounted by a discount factor β, $0 \leq \beta < 1$.

A control policy is defined by a sequence of functions $\delta(i, s', \pi)$, $0 \leq s' < s$, $1 \leq i \leq N$, where s' is the number of time units that have elapsed since the last survey was performed, i is the state that was observed at the last survey, and π is the current N-vector of state probabilities just prior to the application of $\delta(i, s', \pi)$. (The vector π is written $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ where π_i is the probability that the dynamic process actually is in state i.) We seek a control policy to maximize the expected discounted rewards of operating the system over an infinite horizon. A control or decision policy for fixed s is denoted by

$$\delta \equiv \Big(\delta(0), \ldots, \delta(s-1)\Big) .$$

For a fixed intersurvey interval s the cost of a given policy is developed by defining $f(i, s', \pi)$, $0 \leq s' \leq s$ as the expected discounted reward of operating the system given that s' time units have elapsed since the last survey, that i was the state observed at the last survey and that π is the state probability vector s' time units since the last survey. The functions $f(i, s', \cdot)$ satisfy the following system of equations:

$$f(i, s', \pi) = \pi\gamma\Big(\delta(i, s', \pi)\Big) + \beta f\Big(i, s' + 1, \pi P\Big(\delta(i, s', \pi)\Big)\Big)$$

$$0 \leq s' \leq s-1$$

$$f(i, s, \pi) = \pi f(0) - C(s) \tag{1}$$

where $f(0)$ is the column vector $\Big(f(1, 0, e_1), f(2, 0, e_2), \ldots, f(N, 0, e_N)\Big)$ and e_i is the state probability vector $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the i^{th} position representing perfect observation of state i. Note that for a given policy δ the state probability vector π is completely determined by i and s'; thus equation (1) can be more compactly represented by suppressing the dependence on i as follows:

$$f(s', \pi) = \pi\gamma\Big(\delta(s')\Big) + \beta f\Big(s' + 1, \pi P\Big(\delta(s')\Big)\Big) \quad 0 \leq s' \leq s-1$$

$$f(s, \pi) = \pi f(0) - C(s) \tag{2}$$

We note that for any policy δ the functions $\delta(i, s', \pi)$ simplify as follows:

$$\delta(i, 0, e_i) = \delta(i, 0)$$

$$\delta(i, 1, \pi) = \delta\Big(i, 1, \pi(1, i)\Big) = \delta(i, 1) \text{ where} \tag{3}$$

$$\pi(1, i) = e_i P\Big(\delta(i, 0)\Big)$$

$$\delta(i, 2, \pi) = \delta\Big(i, 2, \pi(2, i)\Big) = \delta(i, 2) \text{ where}$$

$$\pi(2,i) = \pi(1, i)P\Big(\delta(i, 1)\Big)$$

$$\begin{array}{ccc} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{array}$$

The state probability vectors $\pi(1, i), \pi(2, i), \ldots, \pi(s, i)$ are called the descendents of state i under policy δ, and are important to the algorithms to follow. Indeed, with the descendents identified, equation (2) is a linear system easily solved for the column vector f(0). It is straightforward to show that an optimal policy δ^* exists that maximizes $f(s', \pi)$ for all s' by noting that for fixed s the process is a partially observed Markov process with N • s states. Existence follows from [11].

An optimal policy is defined, in general, over all possible states of knowledge at each point in time; however, in fact it is sufficient to consider only those states of knowledge that are descendents of completely observed states. Any other states of knowledge are in effect transient, and once a survey is performed will never reoccur. Thus the problem is reduced from considering N states at time s' = 0 and the s-1 (N-1)-dimensional simplices of states of knowledge to simply N states at time s' = 0 and N (N)-vectors at each of the s-1 succeeding time periods between observations. Thus an optimal policy can be defined as that policy $\delta^*(s)$ that maximizes $f(i, 0)$ and $f\Big(i, s', \pi(s', i)\Big)$ $1 \leq s' \leq s$, $1 \leq i \leq N$.

The existence of an optimal policy allows us to concentrate on finding an efficient procedure for improving a policy in order to develop a policy iteration algorithm. We proceed by decomposing the functions $f(i, s', \pi)$ into $f(i, s', \pi) = \pi\alpha(i, s')$. From equations (1) and (2) we can show that the column vectors $\alpha(i, s')$ satisfy:

$$\alpha(i, s') = \gamma\Big(\delta(i, s')\Big) + \beta P\Big(\delta(i, s')\Big)\alpha(i, s' + 1) \quad 0 \leq s' \leq s-1$$

$$\alpha(i, s) = f(s, 0) - C(s)\underline{1} \tag{4}$$

where $\underline{1}$ is a column vector of ones.

We can use the return functions of a given policy δ to find a policy $\hat{\delta}$ that has rewards at least as great as the rewards from δ. This fact is proven in theorem 1 below. Before stating and proving the procedure, we need to examine more closely the function $f(i, s', \pi)$. In general, a given policy defines this function at N points in the space Π, the space of all possible states of knowledge. These N points are the descendents of each state i at time s'. In order to say one policy has an improved expected reward over another policy, we must be able to define $f(i, s', \pi)$ at values of π which are not one of the N descendents of the given policy. Note that for a given policy δ, $f(i, s', \pi) = \pi\alpha(i, s')$ corresponds to the cost of applying policy $\delta(i, s')$, $\delta(i, s' + 1)$, ..., $\delta(i, s-1)$ starting at time s' with state of information π. Clearly, we should choose the sequence $\delta(i, s')$, $\delta(i, s' + 1)$... that maximizes our expected reward. Thus we can let $f(i, s', \pi) = \max_{\ell} \pi\alpha(\ell, s')$. This definition allows $f(i, s', \pi)$ to be well defined over the entire space Π. (A rigorous proof for the sufficiency of this definition can be found in [11].)

We can identify a new policy $\hat{\delta}$ based on δ as follows:

$$\hat{\delta}(i, s) = \underset{a}{\operatorname{argmax}}[\pi(s, i)\gamma(a) + \beta\pi(s, i)P(a)f(0)]$$

$$\hat{\delta}(i, s') = \underset{a}{\operatorname{argmax}}[\pi(s', i)\gamma(a) + \beta\max_{\ell}\pi(s', i)P(a)\hat{\alpha}(\ell, s' + 1)] \qquad (5)$$

$$0 \leq s' \leq s - 1$$

where the α's are defined recursively

$$\hat{\alpha}(i, s) = \gamma\left(\hat{\delta}(s)\right) + \beta P\left(\hat{\delta}(s)\right)f(0)$$

$$\hat{\alpha}(i, s') = \gamma\left(\hat{\delta}(s')\right) + \beta P\left(\hat{\delta}(s')\right)\hat{\alpha}(\hat{\ell}, s' + 1)$$

where $\hat{\ell}$ achieved the inner maximum in (5).

In essence, a new policy is found for each state and each intersurvey time point. The points $\hat{\pi}(1, i)$, $\hat{\pi}(2, i)$, ..., $\hat{\pi}(s-1, i)$ together with $\hat{\pi}(s, i)$ are the descendents of i for the new policy $\hat{\delta}$. We now prove that $\hat{\delta}$ is an improved policy over δ.

Theorem 1. a) The expected value of following policy $\hat{\delta}(s)$ is greater than or equal to the expected value of following policy $\delta(s)$ having just observed state i. That is,

$$\hat{f}(i, 0) \geq f(i, 0) \quad \text{for each } i$$

b) If $\hat{\delta}$ maximized $f(i, s)$ for all i then $\hat{\delta} = \delta^*$ such that δ^* maximizes $f\left(i, s', \pi(s', i)\right)$ where $\pi(s', i)$ is the s'-th descendent of i under policy δ^*.

Proof. Let $\bar{f}\left(i, s', \pi(s', i)\right)$ be the "improved" value after applying (5). At $s' = s$,

$$\bar{f}\left(i, s, \pi(s, i)\right) - f\left(i, s, \pi(s, i)\right) = \pi(s, i)\Big(\gamma\left(\hat{\delta}(s)\right) - \gamma\left(\delta(s)\right) + P\left(\hat{\delta}(s)\right)f(0) - P\left(\delta(s)\right)f(0)\Big) \tag{6}$$

which is nonnegative since maximums are taken in (5). Suppose for $s' = s, s-1, s-2, \ldots, j$ it is true that $\bar{f}(i, s') - f(i, s) \geq 0$ for all states i. Then at $s' = j-1$:

$$\bar{f}\left(i, j-1, \pi(j-1, i)\right) = \pi(j-1, i)\gamma\left(\hat{\delta}(j-1)\right) + \beta\max_{\ell}\pi(j-1, i)P\left(\hat{\delta}(j-1)\right)\hat{\alpha}(\ell, j) \tag{7}$$

$$\geq \pi(j-1, i)\gamma\left(\delta(j-1)\right) + \beta\max_{\ell}\pi(j-1, i)P\left(\delta(j-1)\right)\hat{\alpha}(\ell, j)$$

the last inequality again because (5) implies finding a maximum. By definition, $\pi(j\text{-}1,\ i)P\Big(\delta(j\text{-}1)\Big) = \pi(j,\ i)$. At j, $\hat{\delta}(j)$ was found to have been an improved policy for $\pi(j,\ i)$. This implies

$$\max_{\ell}\pi(j\text{-}1,\ i)P\Big(\delta(j\text{-}1)\Big)\hat{\alpha}(\ell,\ j) \geq \pi(j\text{-}1,\ i)P\Big(\delta(j\text{-}1)\Big)\alpha(i,\ s) = f\Big(i,\ j,\ \pi(j,\ i)\Big) \tag{8}$$

Combining equations (7) and (8) we have

$$\bar{f}\Big(i,\ j\text{-}1,\ \pi(j\text{-}1,\ i)\Big) \geq \pi(j\text{-}1,\ i)\gamma\Big(\delta(j\text{-}1)\Big) + \beta\pi(j\text{-}1,\ i)P\Big(\delta(j\text{-}1)\Big)\alpha(i,\ j) = f\Big(i,\ j\text{-}1,\ \pi(j\text{-}1,\ i)\Big) \tag{9}$$

The inductive proof yields the desired result that $\bar{f}(i,\ 0) \geq f(i,\ 0)$ for all states i. The improvement algorithm (5) starts with an initial value f(0), and adds a value equal to the s-period expected reward of policy $\hat{\delta}$. Let $G(i,\ \hat{\delta})$ be the expected s-period reward when state i is observed at the survey, and policy $\hat{\delta}$ is followed. Then (5) and (9) imply:

$$\bar{f}(i,\ 0) = G(i,\ \hat{\delta}) + \beta^{s}\hat{\pi}(s+1,\ i)f(0) \tag{10}$$

Define $\Delta(i) = \bar{f}(i,\ 0) - f(i,\ 0)$, let $\hat{f}(i,\ 0)$ be the value of $\hat{\delta}$ calculated in the policy evaluation stage, and let $\Delta f(i) = \hat{f}(i,\ 0) - f(i,\ 0)$. Then it follows from standard arguments in Markov decision processes that

$$\Delta f(i) = \Delta(i) + \hat{\pi}(s+1,\ i)\Delta f \tag{11}$$

The remainder of the proof of parts a) and b) follow as in [

□

Substituting (6) into (7) we have

$$\hat{f}\left(i, s', \hat{\pi}(s', i)\right) - f\left(i, s', \hat{\pi}(s', i)\right) = \Delta(i, s')$$

$$+ \beta\left[\hat{\pi}(s', i)P\left(\delta(s')\right)\alpha(\ell, s'+1) - \hat{\pi}(s', i)P\left(\hat{\delta}(s')\right)\alpha(\hat{\ell}, s'+1)\right]$$

$$+ \beta\left[\hat{f}\left(i, s'+1, \hat{\pi}(s'+1, i)\right) - f\left(i, s'+1, \hat{\pi}(s', i)P\left(\delta(s)\right)\right)\right] \quad (8)$$

Noting that $f\left(i, s'+1, \hat{\pi}(s', i)P\left(\delta(s)\right)\right) = \max_{\ell} \hat{\pi}(s', i)P\left(\delta(s)\right)\alpha(\ell, s'+1)$

yields for $0 \leq s' \leq s-1$:

$$\hat{f}\left(i, s', \hat{\pi}(s', i)\right) - f\left(i, s', \hat{\pi}(s', i)\right) = \Delta(i, s')$$

$$+ \beta\left[\hat{f}\left(i, s'+1, \hat{\pi}(s'+1, i)\right) - f\left(i, s'+1, \hat{\pi}(s'+1, i)\right)\right] \quad (9)$$

and for s: $\hat{f}\left(i, s, \hat{\pi}(s, i)\right) - f\left(i, s, \hat{\pi}(s, i)\right) = \hat{\pi}(s, i) \cdot \left[\hat{f}(0) - f(0)\right]$.

Equation (9) and the fact that $\Delta(i, s') \geq 0$ imply that:

$$\hat{f}\left(i, s', \hat{\pi}(s', i)\right) - f\left(i, s', \hat{\pi}(s', i)\right) \geq 0 \qquad 0 \leq s' \leq s-1 \quad (10)$$

Letting $s' = 0$ (for which $\hat{\pi}(0, i) = \pi(0, i) = e_i$) proves part a) of the theorem.

Part b) of theorem guarantees that nonoptimal policies will always be improved. To prove the assertion we assume that δ cannot be improved by (5) but that $\delta^* \neq \delta$ is optimal. Optimality implies that:

$$f^*\left(i, s', \pi^*(s', i)\right) \geq f\left(i, s', \pi^*(s', i)\right) \quad (11)$$

Noting that for fixed s the set of all policies is finite, we now have the rudiments for a policy iteration algorithm. In order to determine the optimal survey interval--which may be unbounded--we would compute an optimal policy and its expected return for increasing intervals until increasing the survey interval further decreases the expected rewards for some state i immediately after observation. In using the algorithms summarized in the next section, the optimal policy and/or optimal value functions derived for an interval of length s are used to initialize the iterations for finding an optimal policy and value for an interval of length $s+1$. In practice, to date, only a few iterations are required to find the optimal policy and/or value function for each sampling interval $s > 1$.

2. POLICY AND VALUE ITERATION ALGORITHMS FOR FINDING δ^*

Fig. 1 The algorithm to find δ^* for fixed s is summarized in Figure 1. For $C(s) = 0$ and $s = 1$ the algorithm reduces to the standard policy iteration algorithm.

Fig. 2 Figure 2 summarizes an equivalent value iteration algorithm. For $C(s) = 0$ and $s = 1$, the algorithm described is Jacobi iterates of successive approximations. The usual upper bounds, as in [12] are still valid, however, the lower bounds given in [12] are not valid except when $C(s) = 0$ and $s = 1$.

3. MAXIMAL INFORMATION ALGORITHM

The algorithms of the preceding section do not make full use of the available information in that they do not allow surveys to occur--if warranted--before the obligatory survey interval s. In this section we relax this restriction, and augment the policy alternative space by a

survey alternative. The survey, if chosen, is assumed to occur immediately following a transition, i.e., at the start of a new time period, and does not consume a time period. If the survey is chosen at the start of a time period in effect two alternatives are chosen: first the survey alternative and then a regular alternative based on the perfectly observed state obtained from the survey. It is straightforward to modify these assumptions to allow for surveys that require a full time period.

We proceed by modifying equation (2) to include the survey alternative. We denote the survey alternative as alternative A+1. As above, we assume a survey will be performed following the s^{th} interval unless performed before that time. The cost of a given policy δ for these assumptions is given by (12), where the points $\pi(s', i)$ are the descendents of the policy δ.

For $0 \leq s' \leq s-1$:

$$f(i, 0) = e_i\gamma\Big(\delta(i, 0)\Big) + \beta f\Big(i, i, \pi(1, i)\Big)$$

$$f\Big(i, s', \pi(s', 1)\Big) = \begin{cases} \pi(s', 1)\gamma\Big(\delta(s', 1)\Big) + \beta f\Big(i, s', \pi(s'+1, i)\Big), & \delta(i, s') \neq A+1 \\ \pi(s, i)\, f(0) - C(s), & \delta(i, s') = A+1 \end{cases} \tag{12}$$

$$f\Big(i, s, \pi(s, i)\Big) = \pi(s, i)\, f(0) - C(s)$$

By defining $f(i, s', \pi) = \underset{\ell}{\text{maximum}}\ \pi\alpha(\ell, s')$ as in section 1, we construct the policy improvement algorithm as follows: An improved policy $\hat{\delta}$ is found by

$$\hat{\delta}(i, s) = \underset{a,\ A+1}{\operatorname{argmax}} \begin{cases} \pi(s, i)\gamma(a) + \beta\pi(s, i)P(\alpha)f(0) \\ \pi(s, i)f(0) - c(s) \end{cases} \tag{13}$$

$$\hat{\delta}(i, s') = \underset{a,\ A+1}{\operatorname{argmax}} \begin{cases} \pi(s, i)\gamma(a) + \beta\max_{\ell}\pi(s, i)P(a)\alpha(\ell, s' + 1) \\ \pi(s, i)f(0) - c(s) \end{cases}$$

$$1 \leq s' \leq s - 1$$

where the $\hat{a}$'s are defined recursively as

$$\hat{\alpha}(i, s) = \gamma\left(\hat{\delta}(s)\right) + \beta P\left(\hat{\delta}(s)\right)f(0)$$

$$\hat{\alpha}(i, s') = \gamma\left(\hat{\delta}(s')\right) + \beta P\left(\hat{\delta}(s)\right)\hat{\alpha}(\hat{\ell}, s' + 1)$$

$$1 \leq s' \leq s - 1$$

where $\hat{\ell}$ achieves the inner maximum in (13).

The improvement algorithm will converge to a possibly suboptimal policy that is to a policy at least optimal for some survey interval $s' < s$. Suppose that a policy is reached that will always survey at an interval s' less than s. Then the states $\pi(i, j)$, $s' < j \leq s$ are transient, in the sense that once a survey is performed, they will never be reached. The algorithm in (13) must be perturbed to consider the possibly transient policies. To do this, at $s' - 1$ an improved policy must be found by searching over all two-period policies, that is:

$$\hat{\delta}(i, s' - 1) = \underset{a,\ A+1}{\operatorname{argmax}} \begin{cases} \pi(s' - 1, i)\gamma(a) + \beta\max_{a',\ A+1}\pi(s' - 1)P(a)\hat{\alpha}(i, s', a') \\ \pi(s' - 1, i)f(0) - c(s) \end{cases}$$

where $\hat{\alpha}(i, s', a) = \gamma(a') + \beta P(a')\hat{\alpha}(\hat{\ell}, s' + 1)$

The result is a pair of policies $\hat{\delta}(i, s' - 1)$ and $\hat{\delta}(i, s') = a'$. With this pertubation added if the algorithm converges to a policy that will always survey within an interval $s' < s$, the improvement algorithm can be seen to guarantee that nonoptimal actions will be improved, as we now show.

Theorem 2. In the general case in which surveys (perfect state information) are allowed before time period s, the expected value of following the policy $\hat{\delta}$ given in equation (13), having just observed state i, is greater than or equal to the expected cost of following policy δ, i.e.

$$\hat{f}(i, 0) \geq f(i, 0) \text{ for all } i$$

In addition, if $\delta(s)$ cannot be improved, then $\delta = \delta^*$.

Proof: Essentially the same as theorem 1, with the following definition. Assume that $\delta(i, s') = A + 1$. Then nominally $\delta(i, s'')$, $s \geq s'' > s'$ is undefined. We define this alternative to be the A + 1st (survey) as follows: If $\delta(i, s') = A + 1$, then $\delta(i, s'') = A + 1$ for $s \geq s'' \geq s'$. With this definition, $f(i, s'', \pi) = f(i, s', \pi)$ and the proof proceeds as in theorem 1.

A policy iteration algorithm and a value iteration algorithm for the maximal information case are described in Figures 3 and 4.

Figs. 3,4

4. AN EXAMPLE

To illustrate the algorithms, we consider Howard's toymaker problem [4], revamped to allow for surveys to provide state information. The original problem has the following parameters:

State i	Alternative a	$P_{ij}(a)$	$\gamma(a, i)$
1. Successful toy	1. No advertising	0.5 0.5	6
	2. Advertising	0.8 0.2	4
2. Unsuccessful toy	1. No research	0.4 0.5	-3
	2. Research	0.7 0.3	-5

We employ a discount rate of $\beta = 0.9$; the optimal solution for the completely observed problem is $\delta(1, 0) = \delta(2, 0) = 2$, $f(1, 0) = 22.2$, $f(2, 0) = 12.3$ with $C(s) = 0$.

To address the problem without perfect information at each time period we expand the alternative space to consider four alternatives including one in which both advertising and research occur.

Alternative a	P(a)	$\gamma(a)$
Advertising/ no research	$\begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 4 \\ -5 \end{bmatrix}$
No advertising/ no research	$\begin{bmatrix} 0.5 & 0.5 \\ 0.7 & 0.3 \end{bmatrix}$	$\begin{bmatrix} 4 \\ -5 \end{bmatrix}$
No advertising/ research	$\begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 6 \\ -3 \end{bmatrix}$
Advertising/ research	$\begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix}$	$\begin{bmatrix} 2 \\ -7 \end{bmatrix}$
Survey	--	-C(s)

We arbitrarily choose C(s) = 0.20, and begin the algorithm with s = 1, equivalent to perfect observation each time period at a cost C(s). We begin with $\delta(i, 0) = 1$, $\delta(i, 1)$ = survey and find that an optimal policy and its returns are as follows:

i	s'	Policy $\delta^*(i, s')$	Return $f^*(i, s')$
1	0		
1	1	Survey	
2	0		
2	1	Survey	

We proceed to s = 2 beginning the iteration with an optimal policy for s = 1. Then $\delta(i, s') = \delta^*(i, s')$, s' = 0, 1 and we set $\delta(i, 2)$ = survey. After iterations the solution is:

i	s'	Policy $\delta^*(i, s')$	Cost $f^*(i, s')$
1	0		
1	1		
1	2	Survey	
2	0		
2	1		
2	2	Survey	

Proceeding to s = 3 and starting with $\delta^3(i, s') = \delta^{2*}(i, s')$, s' = 0, 1, 2, with $\delta^3(i, 3)$ representing survey, in ___ iterations we find the same optimal cost and control and conclude that s = 2 is optimal.

It is illustrative to consider the sensitivity of an optimal policy (and optimal value of s) to changes in C(s). Figure 5 shows these changes. Note that for $C(s) \geq 1.00$ it is optimal not to survey.

Fig. 5

5. SUMMARY AND CONCLUSIONS

The algorithms presented in this paper generalize policy iteration to the case of periodic perfect observation of a Markov process. The algorithms are based on an analysis of the process as a partially observable Markov decision process. The resulting algorithms appear to be highly efficient, requiring only a few extra iterations for each stage s.

The policy iteration algorithms have been programmed in APL and have been run interactively on the computer at the National Institute of Health/ a DEC 10) and on the University of Hawaii's IBM 370/158. The value iteration algorithms have been programmed in FORTRAN and have been run on the University of Hawaii's IBM 370/158. Further computational experience with these algorithms on real life fisheries problems are reported in [10]. (Reference to trade names does not imply endorsement by the National Marine Fisheries Service, NOAA.)

Another potential use for these algorithms is in solving machine inspection and replacement problems. The algorithms solve the renewal problem in a straightforward fashion with a minimum of computational effort. More importantly, the algorithm allows the machine inspection and replacement problem to be modeled and solved in more detail than the usual two or three state, two or three action problem. Using the value iteration algorithm, we have solved a 25 state problem with 26 actions per state for $s = 4$ in slightly over 2 minutes of CPU time, including time to compile the program and to calculate from a problem defined on a continuous state space the transition probabilities and reward vector on a discrete grid. A problem this size would be impractical to solve using either the naive approach of forming an equivalent completely observed problem, or else by using the algorithm for a full scale partially observed Markov decision

problem given in [11]. The algorithm can also solve larger problems with less restrictive assumptions than the regenerative stopping algorithm discussed in [

In all the examples we have solved so far, information seeking has been found to be costly. Over a broad range of values of s, little value is added to the expected return function by surveying more frequently, while the surveys themselves have been expensive. Yet decisionmakers appear to still favor frequent surveys. We can speculate two reasons for this. Firstly, managers may be incorrect in their valuation of information. Secondly, the information may be deemed important either to improve the transistion probability model, or else as a final safeguard or hedge against risk. Martin [13] considers surveying costs in a Bayesian context in order to determine more accurately the transition probabilities. An area of future research would be to combine our approach with his in order to estimate the tradeoffs of increased value by not surveying but decreased certainty about the estimates of the transistion probabilities.

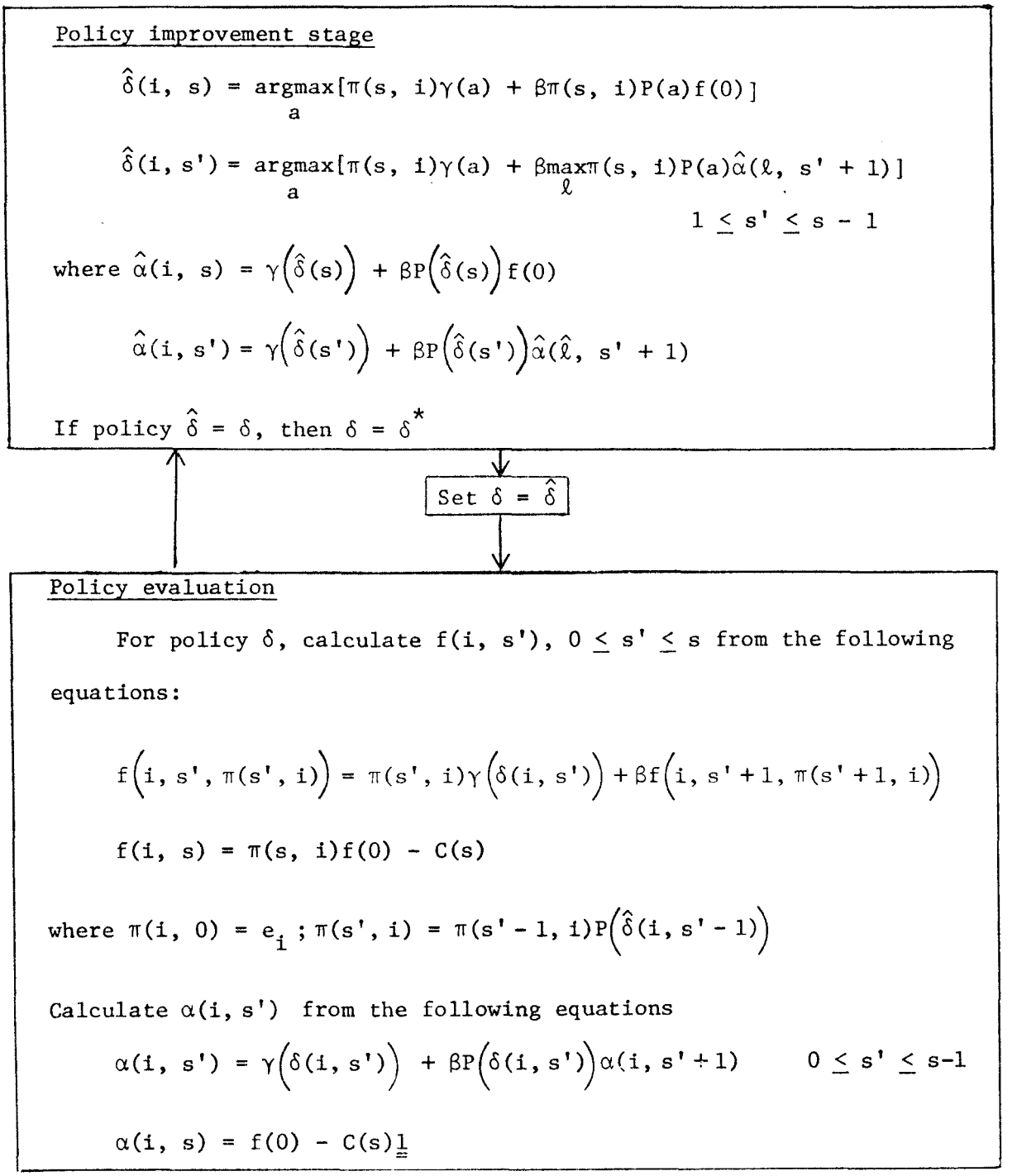

Figure 1. Policy iteration for fixed s.

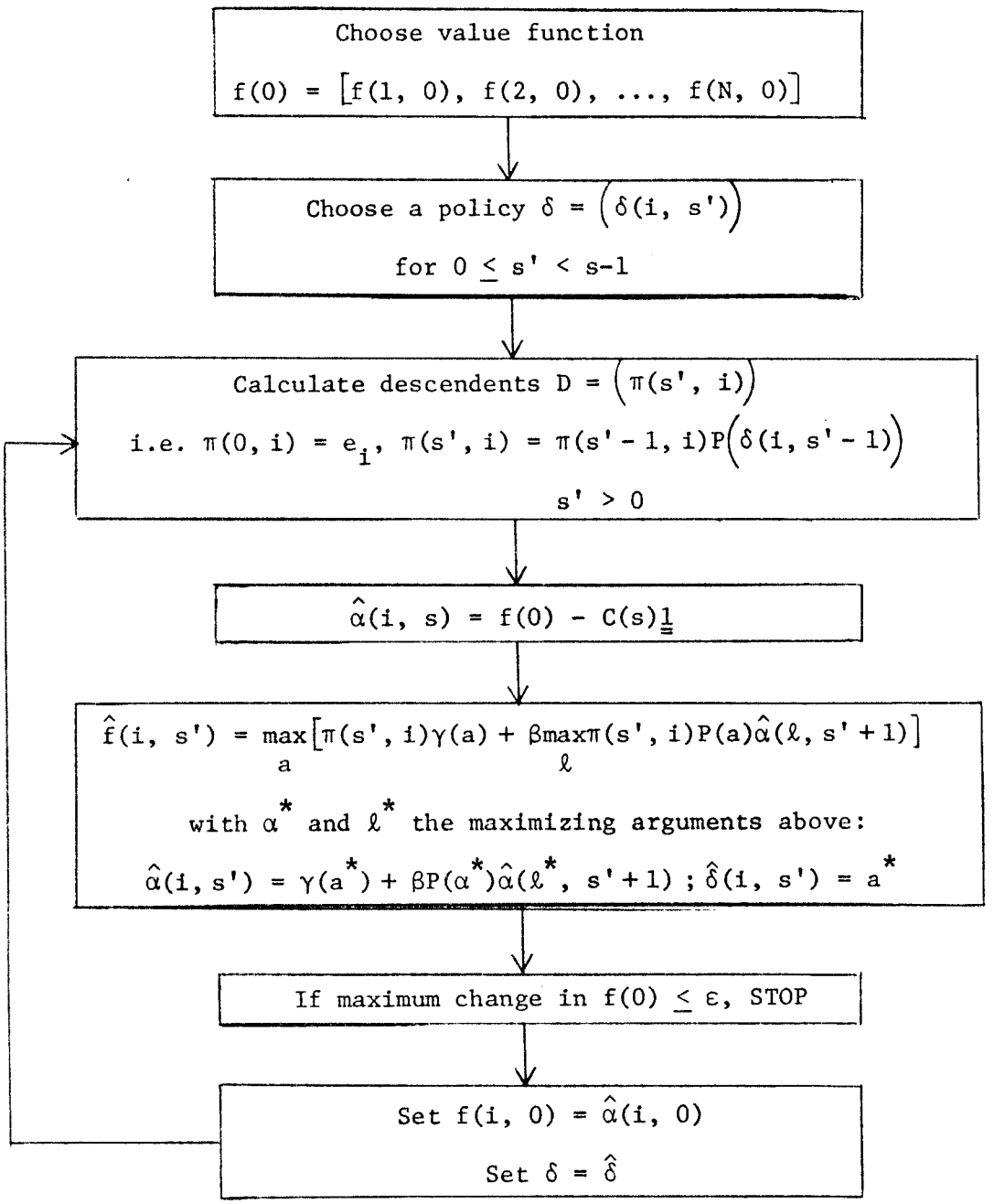

Figure 2. Value iteration algorithm for fixed s.

Policy evaluation

For policy δ, calculate $f(i, s')$, $0 \leq s' \leq s$ from the equations:

$$f(i, 0) = \pi(0, i)\gamma\Big(\delta(i, 0)\Big) + \beta f(i, 1)$$

for $1 \leq s' \leq s-1$:

$$f\Big(i, s', \pi(s', i)\Big) = \begin{cases} \pi(s', i)\gamma\Big(\delta(i, s')\Big) + \beta f\Big(i, s'+1, \pi_{\cdot}(s'+1, i)\Big) & \text{if } \delta(i, s') \neq A+1 \\ \pi(s', i)f(0) - C(s) & \text{if } \delta(i, s') = A+1 \end{cases}$$

$$f\Big(i, s, \pi(s, i)\Big) = \pi(s, i)f(0) - C(s)\underline{1}$$

where $\pi(0, i) = e_i$, $\pi(s', i) = \pi(s'-1, i)P\Big(\hat{\delta}(i, s'-1)\Big)$, $s' \geq 1$

calculate the $\alpha(i, s')$ from:

$$\alpha(i, 0) = \gamma\Big(\delta(i, 0)\Big) + \beta P\Big(\delta(i, 0)\Big)\alpha(i, 1)$$

$$\alpha(i, s') = \begin{cases} \gamma\Big(\delta(i, s')\Big) + \beta P\Big(\delta(i, s')\Big)\alpha(i, s'+1) & \text{if } \delta(i, s') \neq A+1 \\ f(0) - C(s)\underline{1} & \text{if } \delta(i, s') = A+1 \end{cases}$$

$$\alpha(i, s) = f(0) - C(s)\underline{1}$$

Set $\delta = \hat{\delta}$

Policy improvement

$$\hat{\delta}(i, s') = \underset{a,\, A+1}{\text{argmax}} \begin{cases} \pi(s, i)\gamma(a) + \beta \max_{\ell} \pi(s, i)P(a)\hat{\alpha}(\ell, s'+1) \\ \pi(s, i)f(0) - c(s) \end{cases}$$

where $\hat{\alpha}(i, s') = \begin{cases} \gamma\Big(\hat{\delta}(s')\Big) + \beta P\Big(\hat{\delta}(s')\Big)\hat{\alpha}(\hat{\ell}, s'+1) \\ f(0) - c(s) \end{cases}$

If $\hat{\delta} = \delta$, then $\delta = \delta^*$

Figure 3. Policy iteration for variable survey interval.

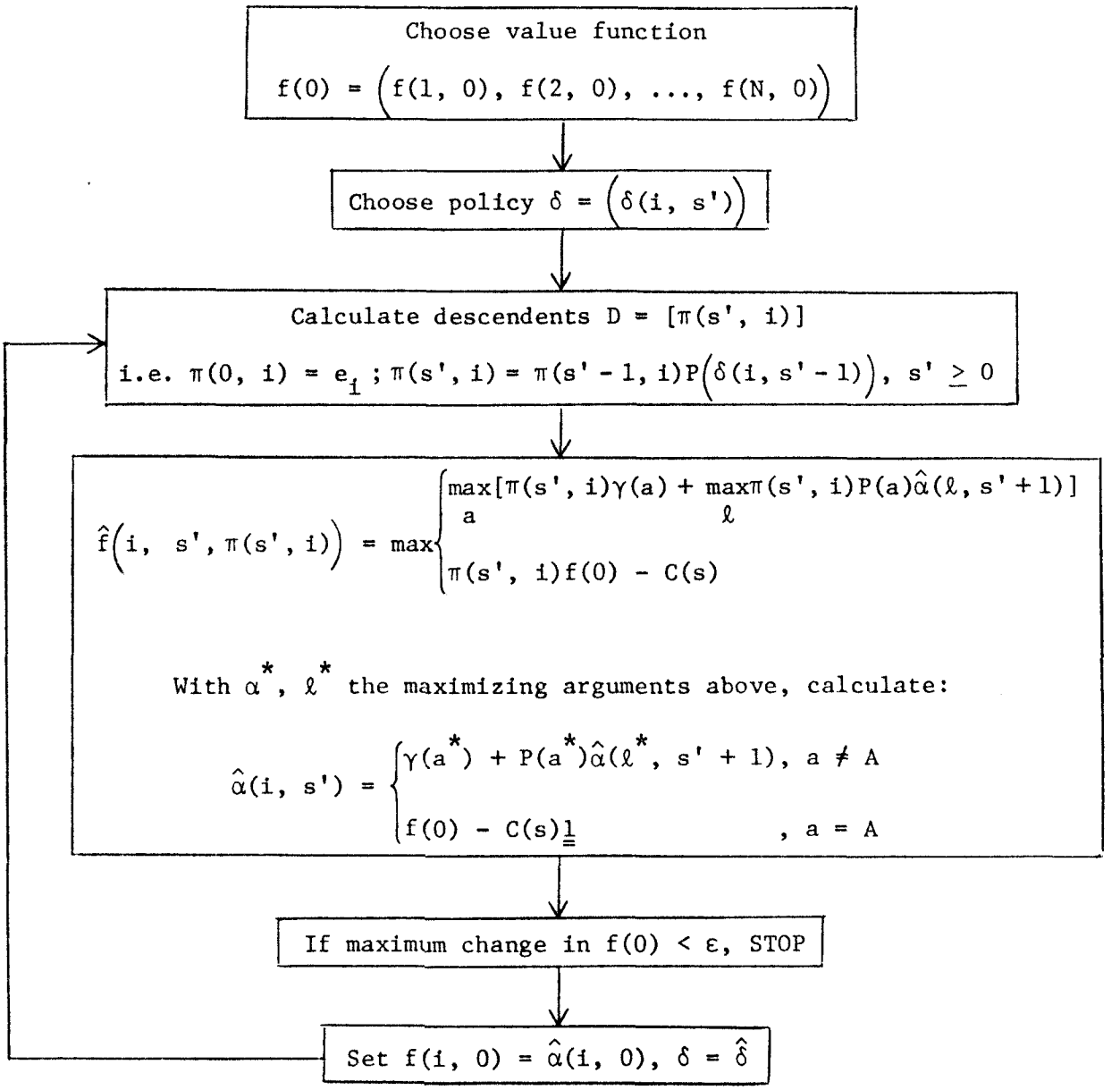

Figure 4. Value iteration for variable survey interval.