PARAMETER IMPRECISION IN FINITE STATE, FINITE ACTION DYNAMIC PROGRAMS

CHELSEA C. WHITE III and HANY K. EL-DEIB

University of Virginia, Charlottesville, Virginia

(Received January 1983; revised September 1983, April 1984; accepted November 1984)

In order to model parameter imprecision associated with a problem's reward or preference structure, we examine a finite state, finite action dynamic program having a one-step transition value-function that is affine in an imprecisely known parameter. For the finite horizon case, we also assume that the terminal value function is affine in the imprecise parameter. We assume that the parameter of interest has no dynamics, no new information about its value is received once the decision process begins, and its imprecision is described by set inclusion. We seek the set of all parameter-independent strategies that are optimal for some value of the imprecisely known parameter. We present a successive approximations procedure for solving the finite horizon case and a policy iteration procedure for determining the solution of the discounted infinite horizon case. These algorithms are then applied to a decision analysis problem with imprecise utility function and to a Markov decision process with imprecise reward structure. We also present conditions that guarantee the existence of a parameter-independent strategy that maximizes, with respect to all other parameter invariant strategies, the minimum value of its expected reward function over all possible parameter values.

Standard models of sequential decision making assume that model parameters are known precisely and do not vary over the problem horizon. These parameters are, for example, terminal node utility values and chance node probabilities for decision analysis problems (Keeney and Raiffa 1976) and single stage rewards, terminal stage rewards, the discount factor, and transition probabilities for Markov decision processes (Bertsekas 1976). Such parameters, however,

1. may be known only imprecisely,
2. may change value, perhaps in an uncertain manner, over the problem's planning horizon,
3. may be subject to an estimation process that is based on parameter value information which arrives sequentially.

A standard procedure for considering a dynamically changing parameter has been state augmentation (Bertsekas); probabilistic and nonprobabilistic state dynamics are described in Bertsekas and in Bertsekas and Rhodes (1973); Figueras (1972); and White (1984), respectively. Researchers have adopted both probabilistic (Van Hee 1978; Sagalovsky 1982; Cyert and DeGroot 1975) and nonprobabilistic (Sworder 1966; Bertsekas and Rhodes, Figueras) approaches to problem formulation and analysis that involve (non-dynamic) parameter imprecision and sequential estimation. Parametric programming (Viswanathan, Aggarwal and Nair 1977), successive approximations, and policy iteration (White and Kim 1980) have been applied to analyze the vector criterion Markov decision process, where the imprecisely known parameter is a vector of importance weights that has no dynamics, is not subject to sequential estimation, and is known only to have nonnegative components and satisfy a sum-to-one property.

Most sequential decision making or control models considered to date consider imprecise, estimated, and/or dynamic parameters that are associated with state dynamics and/or the observation mechanism. Notable exceptions are due to Cyert and DeGroot and White (1984) for parameters in the utility function and to Smallwood (1966) and Veinott (1969) for the discount factor. We remark that Kreps (1977) and Meyer (1977) have assumed that the utility function is allowed to change in an uncertain manner over time; however, these changes depend on a probabilistically described state. Both Kreps and Meyer assume that the functional relationship between utility and state is precisely known over the entire planning horizon.

In this paper, we examine a finite state, finite action dynamic program having a specially structured one-transition value function that is dependent on an imprecisely known parameter. Parameter imprecision is described by set inclusion. For finite horizon prob-

Subject classification: 116 finite state Markov dynamic programming.

0030-354X/86/3401-0120 $01.25

lems, we also assume that the terminal value function is affine (linear plus a constant) in the imprecise parameter. The specific functional form of the one-transition value function and the terminal value function have been selected in order to model parameter imprecision associated with the reward or preference structure of the problem for the case in which no new information regarding the parameter's value becomes available over the planning horizon and the parameter has no dynamics. Our motivation for examining such a problem formulation stems from the following perceptions:

1. rewards and preferences can be difficult to quantify precisely, particularly when multiple, conflicting, and noncommensurate objectives are being considered and/or preference assessment is involved;
2. set inclusion is a particularly simple description of parameter imprecision that is relatively easy to justify behaviorally in the context of preference assessment.

We remark that set inclusion is an often used model of parameter imprecision in the single stage decision-making literature; see for example, Sarin (1977a, b), Fishburn (1965), White, Dozono and Scherer (1983), and their references. The results presented in this paper represent to some extent an extension of this work to the sequential case and a generalization of the cited results for the vector criterion Markov decision process.

The paper is organized as follows. Results for the finite horizon case are presented in Section 1. Under suitable assumptions on the functional form of the one transition value function, we show that the optimal reward functions are all affine and convex in the unknown parameter. This result is used to develop a simple computational algorithm that is reminiscent of a successive approximation algorithm due to Smallwood and Sondik (1973) for the finite horizon, partially observed Markov decision process. The intent of this algorithm is to identify, on the basis of the given description of parameter imprecision, the set of all strategies that may be optimal, or equivalently, to eliminate all clearly suboptimal strategies. This algorithm is applied to a decision analysis problem and to a Markov decision process problem in Section 2. In Section 3 we examine the infinite horizon, discounted case. A policy iteration algorithm, similar to a policy iteration algorithm presented in Sondik (1978) is developed and applied to a Markov decision process problem. Again, the intent of this algorithm is to determine the set of all strategies that are optimal for some feasible parameter value. In Section 4, we present results for determining a best maximin, parameter invariant strategy for the infinite horizon, discounted case. Conclusions are presented in the final section.

1. The Finite Horizon Case

We now present a finite horizon, finite state and action dynamic program having a specially structured one-transition value function and terminal value function. Our approach to problem definition is similar to that of Denardo (1967). Consider the following definitions:

$K < \infty$ represents the number of *stages* or decision epochs of the problem,

S_k is the finite *state space* at stage $k = 0, 1, \ldots, K$,

$A_k(i)$ is the finite *action space* at stage $k = 0, 1, \ldots, K-1$ when the state at stage k is $i \in S_k$,

$\mathscr{P} \subseteq \mathbf{R}^N$ is the set of all possible *parameter values* (assumed to be convex),

V_k is the set of all bounded, real-valued functions on $S_k \times \mathscr{P}$, $k = 0, 1, \ldots, K$,

$h_k: S_k \times \mathscr{P} \times A_k \times V_{k+1} \to \mathbf{R}$ is the bounded one *transition value function* at stage $k = 0, 1, \ldots, K-1$.

$f_K \in V_K$ is the *terminal value function*,

Δ_k is the set of all *parameter dependent policies* δ: $S_k \times \mathscr{P} \to A_k$ such that $\delta(i, \rho) \in A_k(i)$, $k = 0, 1, \ldots, K-1$,

Λ_k is the (finite) set of all *parameter independent policies* $\lambda: S_k \to A_k$ such that $\lambda(i) \in A_k(i)$, $k = 0, 1, \ldots, K-1$,

$H_{k\delta}: V_{k+1} \to V_k$ is defined as $[H_{k\delta}v](i, \rho) = h_k[i, \rho, \delta(i, \rho), v]$ for each $\delta \in \Delta_k$, $k = 0, 1, \ldots, K-1$,

$H_k: V_{k+1} \to V_k$ is defined as $H_k v = \sup_\delta H_{k\delta} v$, $k = 0, 1, \ldots, K-1$,

$f_k: S_k \times P \to \mathbf{R}$ is the *optimal reward function* from stage k to the end of the planning horizon.

A sequence of parameter dependent policies $\pi = \{\delta_0, \ldots, \delta_{K-1}\}$, $\delta_k \in \Delta_k$, $k = 0, 1, \ldots, K-1$, is called a *parameter dependent strategy.*

A sequence of parameter independent policies $\xi = \{\lambda_0, \ldots, \lambda_{K-1}\}$, $\lambda_k \in \Lambda_k$, $k = 0, 1, \ldots, K-1$, is called a *parameter independent strategy.*

We remark that the dynamic programming equation for this problem is

$$f_k(i, \rho) = \max_{a \in A_k(i)} \{h_k(i, \rho, a, f_{k+1})\} \quad k = 0, 1, \ldots, K-1,$$

which we express more succinctly as $f_k = H_k f_{k+1}$ for $k = 0, 1, \ldots, k-1$. If the strategy $\pi^* = \{\delta_0^*, \ldots, \delta_{K-1}^*\}$

satisfies $f_k = H_{k\delta_k^*} f_{k+1}$ for all $k = 0, 1, \ldots, K-1$, then we call π^* an *optimal strategy*. Reference to the examples in Section 2 may help to explicate our notation.

Our objective is to determine the smallest set of parameter independent strategies that are optimal for some $\rho \in \mathscr{P}$. This collection of strategies represents the set of all strategies that may be optimal, given that parameter imprecision is described by the set $\mathscr{P}$. We determine this collection of strategies from the optimal strategy $\pi^* = \{\delta_0^*, \ldots, \delta_{K-1}^*\}$ in the following manner: $\{\lambda_k, k = 0, 1, \ldots, K-1\}$, $\lambda_k : S_k \to A_k$, $\lambda_k(i) \in A_k(i)$, may not be excluded as a candidate for an optimal parameter independent strategy if there exists a $\rho \in \mathscr{P}$ such that $\lambda_k(i) = \delta_k^*(i, \rho)$ for all $i \in S_k$, $k = 0, 1, \ldots, K-1$. We observe that it is sufficient to determine an optimal strategy and the optimal reward functions in order to achieve this objective. The determination of an optimal strategy and the optimal reward functions represents the primary emphasis of this section and of Section 3.

Assume throughout this section that the following assumptions hold:

A1. For each $k = 0, 1, \ldots, K-1$, there are functions $h_k^1 : S \times A \to \mathbf{R}$, $h_k^2 : S \times A \to \mathbf{R}^N$, and $h_k^3 : S^2 \times A \to R$ such that

$$h_k(i, \rho, a, v) = h_k^1(i, a) + \sum_{n=1}^{N} h_k^2(i, a)_n \rho_n + \sum_{j \in S_{k+1}} h_k^3(i, a)_j v(j, \rho).$$

A2. For each $i \in S_k$, $a \in A_k(i)$, and $j \in S_{k+1}$, $h_k^3(i, a)_j \geqslant 0$, $k = 0, 1, \ldots, K-1$.

A3. There are functions $\bar{h}^1$ and $\bar{h}^2$ such that

$$f_K(i, \rho) = \bar{h}^1(i) + \sum_{n=1}^{N} \bar{h}^2(i)_n \rho_n.$$

We remark that these assumptions appear to be quite reasonable for a broad class of decision analysis problems and Markov decision processes having imprecisely known parameters associated with their preference or reward structure, a claim supported by the examples in Section 2. We now present two results that will lead to the development of a computational algorithm for this class of dynamic programs.

Lemma 1. *Let $k \in \{0, 1, \ldots, K-1\}$. Assume $v \in V_{k+1}$ is piecewise affine and convex in ρ on $\mathscr{P}$ for each $i \in S_{k+1}$. Then, $H_k v$ is piecewise affine and convex in ρ on $\mathscr{P}$ for each $i \in S_k$.*

Proof. Note that it is sufficient to show that $h_k(i, \rho, a, v)$ is piecewise affine and convex in ρ for each pair $(i, a) \in S_k \times A_k$. Since $v \in V_{k+1}$ is piecewise affine and convex in ρ for each $i \in S_{k+1}$, then for each $i \in S_{k+1}$, there is a set $\mathscr{A}(i)$ such that

$$v(i, \rho) = \max\{\alpha + \gamma\rho : (\alpha, \gamma) \in \mathscr{A}(i)\}.$$

It then follows that

$$h_k(i, \rho, a, v) = h_k^1(i, a) + h_k^2(i, a)\rho + \sum_j h_k^3(i, a)_j \max\{\alpha + \gamma\rho : (\alpha, \gamma) \in \mathscr{A}(j)\}. \tag{1}$$

Since the sum of nonnegatively weighted piecewise affine and convex functions is piecewise affine and convex, $h_k(i, \rho, a, v)$ is piecewise affine and convex in ρ for each pair (i, a).

Proposition 1. *For each $k = 0, 1, \ldots, K$, f_k is piecewise affine and convex in ρ on $\mathscr{P}$ for each $i \in S_k$.*

Proof. The result is true by assumption for $k = K$. Backward induction and the result in Lemma 1 imply that the result is true for $k = 0, 1, \ldots, K-1$.

These results suggest the following computational procedure for determining the f_k:

0. Define $\mathscr{A}_K(i) = \{(\bar{h}^1(i), \bar{h}^2(i)\}$; set $k = K-1$.
1. Define $\mathscr{A}_k(i, a)$ as the set of all pairs (α', γ'), where

$$a' = h_k^1(i, a) + \sum_j h_k^3(i, a)_j \alpha(j) \tag{2a}$$

$$\gamma' = h_k^2(i, a) + \sum_j h_k^3(i, a)_j \gamma(j) \tag{2b}$$

and where $(\alpha(j), \gamma(j)) \in \mathscr{A}_{k+1}(j)$. Eliminate all pairs (α', γ') in $\mathscr{A}_k(i, a)$ that do not achieve the maximum in $\max\{\alpha' + \gamma'\rho : (\alpha', \gamma') \in \mathscr{A}_k(i, a)\}$ for some value of $\rho \in \mathscr{P}$. (See Smallwood and Sondik for a related elimination procedure.)
2. Define $\mathscr{A}_k(i) = U_a \mathscr{A}_k(i, a)$. Eliminate all pairs (α', γ') in $\mathscr{A}_k(i)$ that do not achieve the maximum in $\max\{\alpha' + \gamma'\rho : (\alpha', \gamma') \in \mathscr{A}_k(i)\}$ for some value of $\rho \in \mathscr{P}$.
3. If $k = 0$, stop; if not, set $k = k - 1$, and go to Step 1.

We remark that

$$f_k(i, \rho) = \max\{\alpha + \gamma\rho : (\alpha, \gamma) \in \mathscr{A}_k(i)\}$$

and that optimal strategy $\delta_k^*(i, \rho) = a$ if $f_k(i, \rho) = \alpha^* + \gamma^*\rho$ and if $(\alpha^*, \gamma^*) \in \mathscr{A}_k(i, a)$. Thus, this algorithm can be used to provide both $\{f_k, k = 0, 1, \ldots, K\}$ and $\{\delta_k^*, k = 0, 1, \ldots, K-1\}$. Note also that (2) is easily derived from (1) by

a. replacing $\mathscr{A}(i)$ with $\mathscr{A}_{k+1}(i)$ in (1),

b. replacing $\max\{\alpha + \gamma\rho : (\alpha, \gamma) \in \mathscr{A}_{k+1}(j)\}$ by $\alpha(j) + \gamma(j)\rho$ for $(\alpha(j), \gamma(j)) \in \mathscr{A}_{k+1}(j)$,
c. collecting terms, and
d. considering every combination of pairs in $\mathscr{A}_{k+1}(j)$ for $j \in S$.

2. Examples: Finite Horizon Case

We now consider two areas of application of the results presented in Section 1, decision analysis and Markov decision processes, and illuminate these results with two associated numerical examples.

Decision Analysis. Assume that the given decision tree has a maximum of K stages. Add the appropriate number of decision nodes with single actions and chance nodes with single outcomes to branches having less than K stages in order to insure all branches have exactly K stages.

Let z_{k+1} be the outcome received after having chosen action a_k, $k = 0, 1, \ldots, K-1$. Define $s_k = \{a_0, z_1, a_1, z_2, \ldots, a_{k-1}, z_k\}$, and assume all probabilities of the form $p(z_{k+1} \mid s_k, a_k)$, and hence $p(s_{k+1} \mid s_k, a_k)$, are known. Let $f_K(s_K) = u(s_K)$, where $u: S_K \to R$ is a utility function. The function u ascribes a utility value to all possible terminal nodes in the decision tree, each branch of which is uniquely associated with an element in S_K.

We consider two cases. Case 1 assumes that all there is known about u is that the collection $\{u(s_K)\}$ is a member of a set $U \subseteq \mathbf{R}^N$, where $N = \#S_K$. For this case, $h_k^1 = h_k^2 = \bar{h}^1 = 0$, $h_k^3(i, a)_j = P(s_{k+1} = j \mid s_k = i, a_k = a)$, and $\bar{h}^2(i)_n = 1 (= 0)$ if $i = n$ (if $i \neq n$). Thus, ρ_n corresponds to the appropriate value of $u(s_k)$.

With respect to Case 2, we assume that $u(s_K) = \sum_{m=1}^{M} w_m u_m(s_K)$, where $u_m: S_K \to \mathbf{R}$ is the utility function associated with the m^{th} attribute, M is the number of attributes under consideration, the attributes are assumed additive and independent, $w_m \geq 0$ is the important weight of attribute m, and $\sum_m w_m = 1$. Assume each u_m is known exactly and all that is known about the vector of importance weights $w = \{w_m\}$ is that it is a member of the set $W \subseteq \mathbf{R}^M$. For this case, $h_k^1 = h_k^2 = \bar{h}^1 = 0$, $h_k^3(i, a) = P(s_{k+1} = j \mid s_k = i, a_k = a)$, $N = M$, $\bar{h}^2(i)_m = u_m(i)$, and $\rho_m = w_m$.

Example 1

This example illustrates the Case 1 decision analysis model. Consider the decision tree in Figure 1, which is based on the ore buying example presented in Chapter 2 of Brown, Kahr and Peterson (1974). Assume that all that is known about the single attribute terminal utility values u_i for $i = 1, \ldots, 9$ is

$u_1 \in [0.7, 1.0]$, $u_2 = 0$, $u_3 = 0.43$,

$u_4 = u_1$, $u_5 = 0$, $u_6 = 0.46$,

$u_7 \in [0.7, 1.0]$, $u_7 \geq u_1$, $u_8 = 0.01$, $u_9 = 0.43$.

Let $\mathscr{P} = \{(\rho_1, \rho_2) : \rho_1, \rho_2 \in [0.7, 1.0] \text{ and } \rho_2 \geq \rho_1\}$, $S_0 = \{01\}$, $S_1 = \{11, \ldots, 15\}$, and $S_2 = \{21, \ldots, 29\}$, where these states are defined in Figure 1. Associate ρ_1 with u_1 and ρ_2 with u_7. The functions f_k and the optimal strategy π^*, as a function of state and parameter value, are given in Figure 1. We see there are two strategies, λ^1 and λ^2, that are possibly preferred, given the available utility value information:

$\lambda^1(01) = a_0^1$, $\lambda^1(11) = a_1^1$, $\lambda^1(12) = a_1^4$,

$\lambda^2(01) = a_0^2$.

Markov Decision Processes. Consider the following stationary, finite horizon Markov decision process. Let $p_{ij}(a)$ be the probability of making transition from state i to state j at the next stage, given action a was just selected. If at stage k the system is in state i and action a was just selected, a reward of $r(i, a)$ is assumed to be accrued, for $k = 0, 1, \ldots, K-1$. If at the terminal stage K, the system is in state i, then a terminal reward $\bar{r}(i)$ is accrued. The case in which the terminal reward is imprecisely known can be treated in a manner analogous to the above decision analysis results. We therefore assume $\bar{r}(i)$ is precisely known, and hence $\bar{h}^1(i) = \bar{r}(i)$ and $\bar{h}^2(i)_n = 0$ for all i and n. Assume all that is known about $r = \{r(i, a)\}$ is that it is a member of the given set $R \subseteq \mathbf{R}^N$, where $N = {}^{\#}S \times {}^{\#}A$. Then, $h^2(i, a)_n = 1 (= 0)$ if $n = (i, a)$ (if $n \neq (i, a)$).

Example 2

Consider the following Markov decision process, which is based on the maintenance-model example presented in Chapter 13 of Hillier and Lieberman (1980). Let $K = 2$, $S = \{1, \ldots, 4\}$, $A = \{1, 2, 3\}$, $\bar{r}(i) = 0$, $r(i, a) = h^1(i, a) + h^2(i, a)_1\rho_1 + h^2(i, a)_2\rho_2$, where

$$h^1(i, a) = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ -3 & 0 & 0 \\ -\infty & -\infty & 0 \end{bmatrix}, \quad h^2(i, a)_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$h^2(i, a)_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\{p_{ij}(1)\} = \begin{bmatrix} 0 & 7/8 & 1/16 & 1/16 \\ 0 & 3/4 & 1/8 & 1/8 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$p_{i2}(2) = 1$ for all i, $p_{i1}(3) = 1$ for all i, $\rho_1 \in [-6, -2]$,

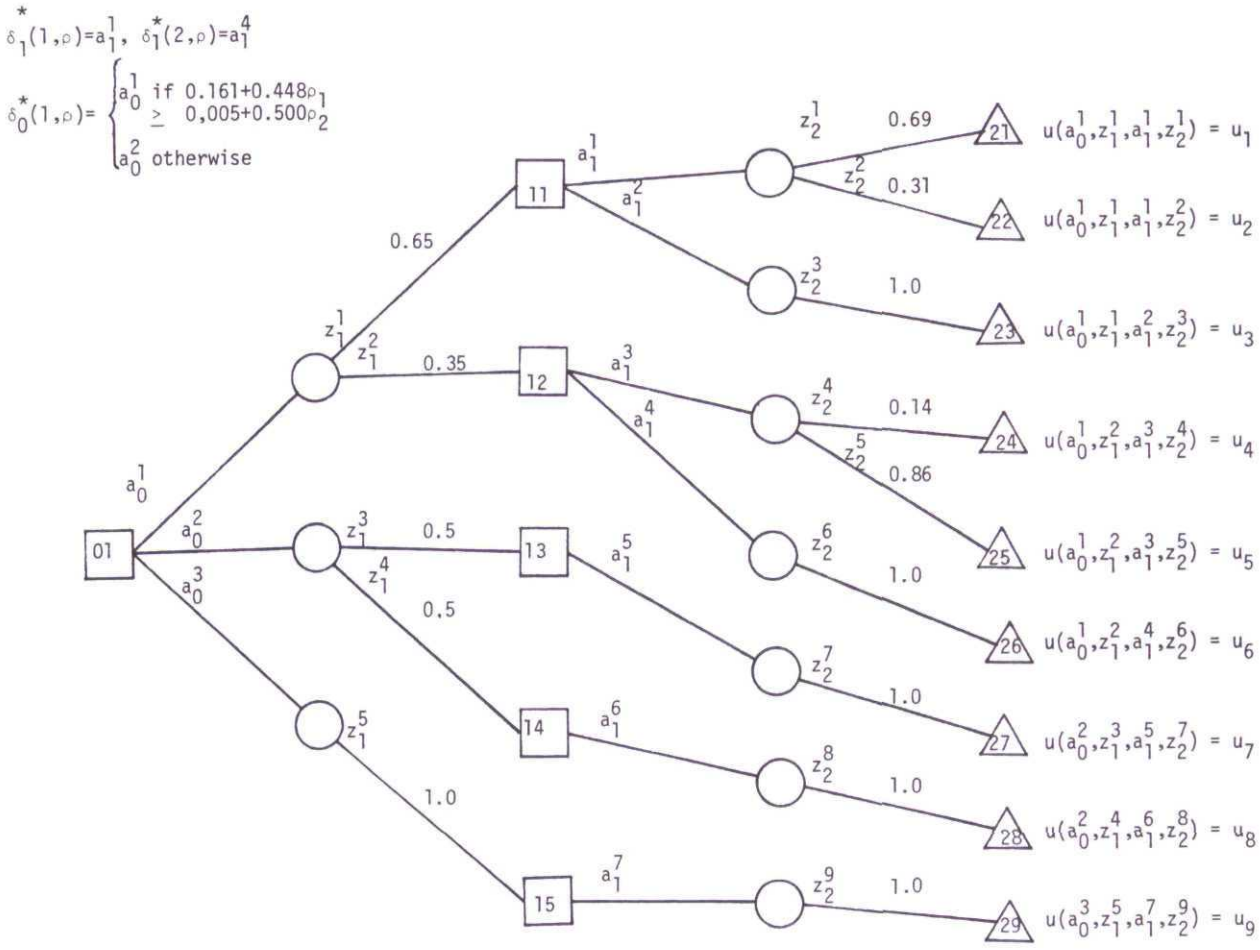

Figure 1. Decision tree and possibly optimal strategies for Example 1.

$\rho_2 \in [-7, -4]$, and $\rho_2 \leq \rho_1$. Table I presents the f_k and π^* for this example. We note that there are three strategies that are possibly preferred, based on the given reward structure information, all three of which select action 1 in states 1 and 2 and action 3 in state 4 for both $k = 0$ and $k = 1$:

$$\lambda_0^1(3) = 3, \quad \lambda_1^1(3) = 1$$
$$\lambda_0^2(3) = 2, \quad \lambda_1^2(3) = 1$$
$$\lambda_0^3(3) = 2, \quad \lambda_1^3(3) = 2$$

where the subscript designates stage and the superscript designates strategy. The fourth possibility, $\lambda_0^4(3) = 3$ and $\lambda_1^4(3) = 2$, produces an optimal expected total reward function identical to these strategies only when $\rho_1 = -4$ and $\rho_2 = -3$, is inferior to at least one of these strategies otherwise, and thus has been deleted. The analysis has therefore eliminated all but three of the 81 possible strategies for this Markov decision process.

3. The Infinite Horizon, Discounted Case

We now consider the case in which $K = \infty$ and S_k, A_k, h_k, and hence Δ_k, Λ_k, V_k, $H_{k\delta}$, and H_k are all stage invariant. We continue to assume that assumptions A1 and A2 hold and additionally assume the following:

A4. There is a β, $0 \leq \beta < 1$, such that $\sum_{j \in S} h^3(i, a)_j \leq \beta$, for all $i \in S$ and $a \in A(i)$.

Assumptions A2 and A4 imply that H_δ, for all $\delta \in \Delta$, and H are isotone contractions on V with respect to the supremum norm, and hence there exist unique fixed points f_δ and f of the operators H_δ and H, respectively. Our objective is to determine f and δ^* such that $f = Hf = H_{\delta^*} f$.

A variety of computational procedures have been used to determine f and δ for the $h^2 = 0$ case: linear

programming (d'Epenoux 1960), successive approximations (White 1963), policy iteration (Howard 1960), and various modifications and combinations of the latter two procedures (Puterman and Shin 1978, 1982; Schweitzer 1971; Platzman, White and Popyack 1984). Viswanathan, Aggarwal and Nair have applied parametric linear programming and White and Kim have applied successive approximations and policy iteration in order to determine f and δ for the case where $h^1 = 0$, $\mathscr{P} = \{\rho : \rho_n \geq 0, \sum_n \rho_n = 1\}$, and $\sum_j h^3(i, a)_j = \beta$ for all i and a, i.e., the vector criterion Markov decison process. The latter approaches were based on algorithms due to Smallwood and Sondik, and Sondik. We remark that Henig (1978) has shown that for the vector criterion Markov decision process, the set of all stationary, parameter invariant policies generated by δ produces all extreme points of the convex hull of the nondominated set. We conjecture that the parametric programming approach presented by Viswanathan et al. is easily extended to consider our more general problem formulation, as is the successive approximations procedure presented earlier. We now motivate and present a generalized policy iteration algorithm for the solution of the infinite horizon case.

Let f_λ be the fixed point of H_λ, where $\lambda \in \Lambda$. Note that

$$f_\lambda(\cdot, \rho) = (I - h_\lambda^3)^{-1}(h_\lambda^1 + h_\lambda^2 \rho)$$

is affine in ρ, where

$$h_\lambda^1 = \{h^1[i, \lambda(i)],\ i \in S\},$$

$$h_\lambda^2 = \{h^2[i, \lambda(i)]_n,\ i \in S,\ n = 1, \ldots, N\}, \quad \text{and}$$

$$h_\lambda^3 = \{h^3[i, \lambda(i)]_j,\ i, j \in S\}.$$

The following result provides an important characterization of the fixed point of H that serves to motivate the algorithm presented below for determining this fixed point.

Proposition 2. *The fixed point of H, f, is piecewise affine and convex and is given by*

$$f(i, \rho) = \max_{\lambda \in \Lambda} f_\lambda(i, \rho)$$

for all $i \in S$ *and* $\rho \in \mathscr{P}$.

Proof. Proposition 2, p. 229, in Bertsekas implies that $\max\{f_\lambda(\cdot, \rho) : \lambda \in \Lambda\}$ is the fixed point of the operator H for any ρ. The result follows from the fact that the maximum of a finite number of affine functions is piecewise affine and convex.

We remark that the piecewise affine and convex nature of f allows us to avoid issues related to the concept of finite transcience found in Sondik.

Proposition 2 suggests that an appropriate procedure for calculating f is to determine a set $\Lambda^* \subseteq \Lambda$, containing as small a number of parameter independent policies as possible, such that

$$f(i, \rho) = \max_{\lambda \in \Lambda^*} f_\lambda(i, \rho)$$

for all $i \in S$ and $\rho \in \mathscr{P}$. The intent of the following policy iteration algorithm is to accomplish this objective:

0. Let $\Lambda_0 \subseteq \Lambda$ be given, and set $n = 0$.
1. Let $\overline{f}_n(i, \rho) = \max_{\lambda \in \Lambda_n} f_\lambda(i, \rho)$, $i \in S$, $\rho \in \mathscr{P}$.
2. Remove any λ from Λ_n that does not achieve the maximum in Step 1 some $i \in S$ and $\rho \in \mathscr{P}$.

TABLE I

Optimal Expected Utility Functions and Possibly Optimal Strategies for Example 2

$$f_2(i, \rho) = 0 \quad \text{for all} \quad i \in S, \qquad \rho \in \mathscr{P}$$

$$f_1(1, \rho) = 0, \qquad \delta_1^*(1, \rho) = 1$$

$$f_1(2, \rho) = -1, \qquad \delta_1^*(2, \rho) = 1$$

$$f_1(3, \rho) = \max\{-3, \rho_1\}, \qquad \delta_1^*(3, \rho) = \begin{cases} 1 & \text{if} \quad -3 \geq \rho_1 \\ 2 & \text{if} \quad \rho_1 \geq -3 \end{cases}$$

$$f_1(4, \rho) = \rho_2, \qquad \delta_1^*(4, \rho) = 3$$

$$f_0(1, \rho) = \max\{(-17 + \rho_2)/16, (-14 + \rho_1 + \rho_2)/16\}. \qquad \delta_0^*(1, \rho) = 1$$

$$f_0(2, \rho) = \max\{(-17 + \rho_2)/8, (-14 + \rho_1 + \rho_2)/8\}, \qquad \delta_0^*(2, \rho) = 1$$

$$f_0(3, \rho) = \max\{-1 + \rho_1, \rho_2\}, \qquad \delta_0^*(3, \rho) = \begin{cases} 2 & \text{if} \quad -1 + \rho_1 \geq \rho_2 \\ 3 & \text{if} \quad \rho_2 \geq -1 + \rho_1 \end{cases}$$

$$f_0(4, \rho) = \rho_2, \qquad \delta_0^*(4, \rho) = 3$$

3. Let $\Lambda_{n+1} \subseteq \Lambda$ be composed of the set of all λ that for some $\rho \in \mathscr{P}$ achieves the maximum in $H\bar{f}_n$ for all $i \in S$. Set $\Lambda_{n+1} = \Lambda_n$ if possible.
4. If $\Lambda_{n+1} = \Lambda_n$, then stop. Otherwise, set $n = n + 1$, and go to Step 1.

We remark that by choosing Λ_{n+1} in Step 3 to equal Λ_n whenever possible, we eliminate the possibility of cycling due to any nonuniqueness in achieving the maximum in $H\bar{f}_n$. We also remark that Λ_{n+1} can be determined from $\delta^* \in \Delta$, where $H_{\delta^*}\bar{f}_n = H\bar{f}_n$, as follows: $\lambda \in \Lambda_{n+1}$ if and only if $\delta^*(\cdot, \rho) = \lambda(\cdot)$ for some $\rho \in \mathscr{P}$. The parameter dependent policy δ^* and the function $\bar{f}_{n+1}$ can be determined in a manner analogous to the procedure for determining δ_k^* and f_k presented in Section 1.

An alternative approach to describing this algorithm is to eliminate Step 2 and to ask in Step 3 if there exists a Λ_{n+1} such that $\Lambda_{n+1} \subseteq \Lambda_n$. Step 4 would then be modified to use $\Lambda_{n+1} \subseteq \Lambda_n$ as the stopping rule. We remark that the algorithm viewed in this manner guarantees convergence in one iteration if Λ_0 is chosen to equal Λ since $\Lambda_1 \subseteq \Lambda_0$ for any Λ_1. Of course, the computational impracticality of this choice of Λ_0 limits its usefulness. With regard to the selection of Λ_0, we note that if $\Lambda_0 = \Lambda^*$, then convergence occurs in one step. We suggest, therefore, selecting Λ_0 as the best a priori estimate of Λ^*. We now verify that the algorithm converges to Λ^* and that this convergence is achieved in a finite number of steps.

Proposition 3. *The policy iteration algorithm converges to the fixed point of H in a finite number of iterations.*

Proof of this result follows the proof of two lemmas. The first lemma indicates that the stopping rule, $\Lambda_{n+1} = \Lambda_n$, is a valid one. The second lemma guarantees that if $\Lambda_{n+1} \neq \Lambda_n$, then Λ_{n+1} produces a strictly better policy than does Λ_n.

Lemma 2. *If $\Lambda_{n+1} = \Lambda_n$, then $\bar{f}_n$ is the fixed point of H.*

Proof. The proof of Lemma 1 and the assumption $\Lambda_{n+1} = \Lambda_n$ guarantee that for fixed ρ, if $\lambda \in \Lambda_n$ is such that $f_\lambda(\cdot, \rho) \geq f_\sigma(\cdot, \rho)$ for all $\sigma \in \Lambda_n$, then $f_\lambda(\cdot, \rho) = [Hf_\lambda](\cdot, \rho) = [H_\lambda f_\lambda](\cdot, \rho)$. Proof of Lemma 2 then follows from the fact that this result holds for all $\rho \in \mathscr{P}$ and from the definition of $\bar{f}_n$.

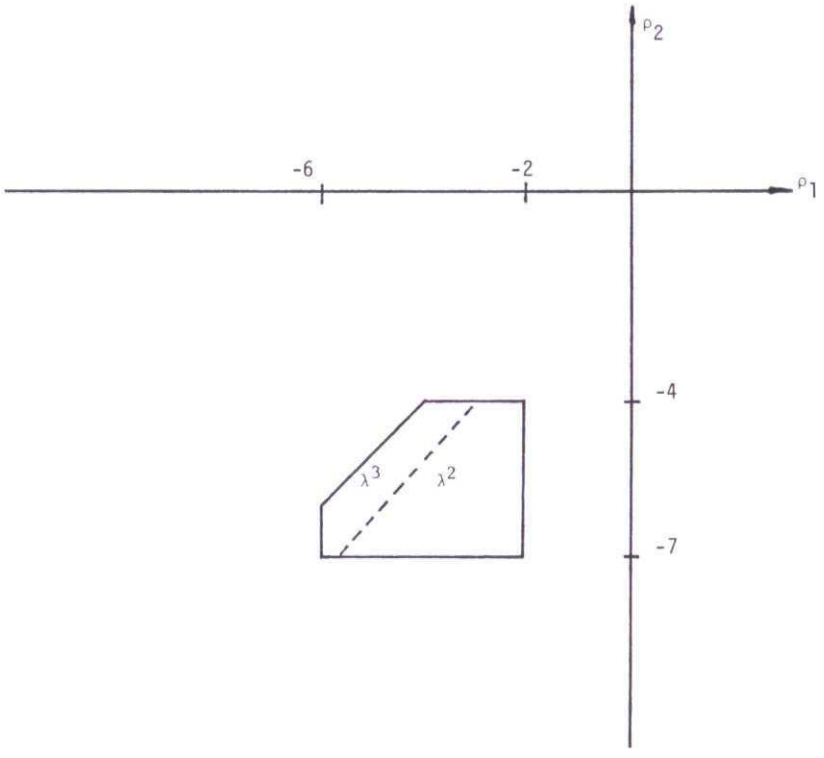

Figure 2. The regions in $\mathscr{P}$ where λ^2 and λ^3 are optimal for Example 3.

TABLE II
Expected Total Discounted Reward Functions for Example 3

$$f_\lambda(\cdot, \rho) = \begin{bmatrix} -9.12945 \\ -10.13630 \\ -12.17710 \\ -8.21650 \end{bmatrix} + \begin{bmatrix} 0 & 1.29447 \\ 0 & 1.36260 \\ 0 & 1.77139 \\ 0 & 2.16503 \end{bmatrix} \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$$

$$f_\lambda 2(\cdot, \rho) = \begin{bmatrix} -6.57031 \\ -7.55243 \\ -6.69818 \\ -5.91327 \end{bmatrix} + \begin{bmatrix} 0.83782 & 0.83782 \\ 0.88192 & 0.88192 \\ 1.79373 & 0.79373 \\ 0.75404 & 1.75404 \end{bmatrix} \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$$

$$f_\lambda 3(\cdot, \rho) = \begin{bmatrix} -5.93779 \\ -6.77663 \\ -5.34402 \\ -5.34402 \end{bmatrix} + \begin{bmatrix} 0 & 1.61169 \\ 0 & 1.69651 \\ 0 & 2.45052 \\ 0 & 2.45052 \end{bmatrix} \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$$

Lemma 3. *If* $\Lambda_{n+1} \neq \Lambda_n$, *then* $\overline{f}_{n+1} \geq \overline{f}_n$, $\overline{f}_{n+1} \neq \overline{f}_n$.

Proof. Step 3 of the algorithm implies that if $\Lambda_{n+1} \neq \Lambda_n$, then $H\overline{f}_n \geq \overline{f}_n$, $H\overline{f}_n \neq \overline{f}_n$. Let $\delta \in \Delta$ satisfy $H_\delta \overline{f}_n = H\overline{f}_n$. Note that δ is comprised of λ's in Λ_{n+1}. By the isotonicity of H_δ, for all $k \geq 1$, $H_\delta^k \overline{f}_n \geq \overline{f}_n$, $H_\delta^k \overline{f}_n \neq \overline{f}_n$; thus, $f_\delta \geq f_n$, $\overline{f}_\delta \neq \overline{f}_n$, where f_δ is the fixed point of H_δ. It is easy to show that $f_\delta = f_\lambda$ whenever $\delta = \lambda$, $\lambda \in \Lambda_{n+1}$. However, $\overline{f}_{n+1}(i, \rho) = \max_\lambda f_\lambda(i, \rho)$, $\lambda \in \Lambda_{n+1}$; thus, $\overline{f}_{n+1} \geq f_\delta$.

Proof of Proposition 3. Since Λ is finite, it has only a finite number of subsets that can be examined. Each of these subsets can be examined at most once since

1. if $\Lambda_{n+1} = \Lambda_n$, the algorithm stops (by Lemma 2);
2. if $\Lambda_{n+1} \neq \Lambda_n$, then $\overline{f}_{n+1} \geq \overline{f}_n$, $\overline{f}_{n+1} \neq \overline{f}_n$ (by Lemma 3), ensuring that Λ_n will not be considered further.

The algorithm determines the set of all parameter independent policies that may be optimal, given $\mathscr{P}$. It may be useful to be able to identify those regions in $\mathscr{P}$ where these parameter independent policies are optimal. For example, if further assessment information becomes available which indicates that the parameter of interest is now known to be in the set $\mathscr{P}' \subseteq \mathscr{P}$, it may be desirable to know if any of the nonexcluded parameter independent policies can be eliminated. It is easy to show that if Λ_n is the nonexcluded subset of parameter independent policies, then the optimal parameter dependent policy δ^* is such that for $\lambda \in \Lambda_n$, $\delta^*(i, \rho) = \lambda(i)$, for all $i \in S$, for all $\rho \in \mathscr{P}$ such that $f_\lambda(\cdot, \rho) \geq f_\sigma(\cdot, \rho)$, for any $\sigma \in \Lambda_n$.

Example 3

Consider again the maintenance model examined in Example 2, assuming the planning horizon is infinite, the criterion is the expected total discounted reward, and the discount factor is 0.9. Standard results (Bertsekas) indicate that it is sufficient to examine only strategies that are stage invariant. Let $\Lambda_0 = \{\lambda^1\}$, where $\lambda^1(1) = 1$, $\lambda^1(2) = 1$, $\lambda^1(3) = 1$, and $\lambda^1(4) = 3$. The function f_{λ^1} is given in Table II. Let $\delta^* \in \Delta$ achieve the maximum in Hf_{λ^1}. It is easy to show that $\delta^*(1, \rho) = \delta^*(2, \rho) = 1$ and $\delta^*(4, \rho) = 3$ for all $\rho \in \mathscr{P}$. Straightforward calculations show that

$$h(3, \rho, 1, f_{\lambda^1}) = -13.1968 + 1.96821\rho_2$$

$$h(3, \rho, 2, f_{\lambda^1}) = -10.13630 + \rho_1 + 1.36260\rho_2$$

$$h(3, \rho, 3, f_{\lambda^1}) = -9.12945 + 2.2944\rho_2$$

indicating for $\rho \in \mathscr{P}$,

$$\delta^*(3, \rho) = 2 \quad \text{if} \quad \rho_1 - 0.9684\rho_2 \geq 1.00685$$
$$= 3 \quad \text{otherwise.}$$

Thus, $\Lambda_1 = \{\lambda^2, \lambda^3\} \neq \Lambda_0$, where

$$\lambda^2(1) = \lambda^3(1) = 1$$
$$\lambda^2(2) = \lambda^3(2) = 1$$
$$\lambda^2(4) = \lambda^3(4) = 3$$
$$\lambda^2(3) = 2, \qquad \lambda^3(3) = 3.$$

The resulting functions f_{λ^2} and f_{λ^3} are also given in Table II. Further calculations indicate that $\Lambda_2 = \Lambda_1$, and hence the algorithm can stop. The regions of $\mathscr{P}$ where λ^2 and λ^3 are optimal are presented in Figure 2 and are associated with the statement: $\lambda^2(\lambda^3)$ is optimal if and only if

$$\rho_1 \geq (\leq) 0.92\rho_2 + 0.75.$$

4. Maximin Policies

In earlier sections, we have concentrated on determining the set of all parameter independent strategies that are optimal for some value of the imprecise parameter. If further parameter value assessment can eliminate all but one of these strategies, then strategy selection

is trivial. If, however, such elimination is not possible, we are still confronted with the problem of which parameter-independent strategy to choose. A likely candidate, and the candidate of interest in this section, is the parameter-independent strategy that maximizes, with respect to all other strategies, the minimum value of its expected reward function, where the minimum is taken over all possible parameter values. That is, for the infinite horizon case, we seek a $\tilde{\lambda} \in \Lambda$ such that $\inf\{f_{\tilde{\lambda}}(i, \rho): \rho \in \} \geq \inf\{f_{\lambda}(i, \rho): \rho \in \mathscr{P}\}$ for all $i \in S$ and $\lambda \in \Lambda$. We remark that such a $\tilde{\lambda}$ may not be a member of the set of all parameter-independent strategies that are optimal for some value of the imprecise parameter. The infinite horizon problem defined in Section 3 represents the problem of interest; therefore, we will assume assumptions A1, A2, and A4 hold throughout. Extension of the following result to the finite horizon case is straightforward and left to the reader.

Proposition 4. *Assume there is a* $\tilde{\rho} \in \mathscr{P}$ *such that* $h^2(i, a)\tilde{\rho} = inf\{h^2(i, a)\rho: \rho \in \mathscr{P}\}$ *for all* $i \in S$ *and* $a \in A(i)$, *and let* $\tilde{\lambda} \in \Lambda$ *satisfy* $f_{\tilde{\lambda}}(i, \tilde{\rho}) = max\{f_{\lambda}(i, \tilde{\rho}): \lambda \in \Lambda\}$ *for all* $i \in S$. *Then,* $inf\{f_{\tilde{\lambda}}(i, \rho): \rho \in \mathscr{P}\} \geq inf\{f_{\lambda}(i, \rho): \rho \in \mathscr{P}\}$ *for all* $i \in S$ *and* $\lambda \in \Lambda$.

Proof. Note that for all $\lambda \in \Lambda$, $\rho \in \mathscr{P}$, $i \in S$, and $f \in V$, $h[i, \tilde{\rho}, \lambda(i), f] \leq h[i, \rho, \lambda(i), f]$. Assumptions A2 and A4 then guarantee that iterating both sides of the previous inequality produces convergent sequences with limits satisfying $f_{\lambda}(i, \tilde{\rho}) \leq f_{\lambda}(i, \rho)$ for all $\lambda \in \Lambda$, $\rho \in \mathscr{P}$, and $i \in S$. Thus, for any $\lambda \in \Lambda$ and all $i \in S$,

$$\inf\{f_{\tilde{\lambda}}(i, \rho): \rho \in \mathscr{P}\} = f_{\tilde{\lambda}}(i, \tilde{\rho}) \geq f_{\lambda}(i, \tilde{\rho})$$

$$= \inf\{f_{\lambda}(i, \rho): \rho \in \mathscr{P}\}.$$

We remark that although the existence of a $\tilde{\rho} \in \mathscr{P}$ in the hypotheses of Proposition 4 appears to be a strong assumption, there are interesting problem formulations that satisfy it. For example, consider the case in which $h^2(i, a)_n = 1$ if $i = n$ (= 0 if $i \neq n$), $\rho_n \in [LB_n, UB_n]$, $n = 1, \ldots, N$. Then, we would select $\tilde{\rho} = \{LB_1, \ldots, LB_N\}^T$, where T = transpose. As another example, observe that in Example 3, $\tilde{\rho} = (-6, -7)$ satisfies this condition and hence by Proposition 4 $\inf\{f_{\tilde{\lambda}}(i, \rho): \rho \in \mathscr{P}\}) \geq \inf\{f_{\lambda}(i, \rho): \rho \in \mathscr{P}\}$ for all $i \in$ S and $\lambda \in \Lambda$.

5. Conclusions

In this paper, we considered a specially structured class of finite state, finite action dynamic programs having an imprecisely described parameter. We assumed that the parameter has no dynamics, no new information about its value is received once the problem begins, and its imprecision is described by set inclusion. We developed computational procedures for determining the set of all strategies that may be optimal, given a description of the parameter's imprecision. Applications to decision analysis and Markov decision processes were presented. A condition that was also presented guaranteed that a parameter-independent strategy maximizes, with respect to all other parameter independent strategies, the minimum value of its expected reward function over all possible parameter values. Consideration of different functional forms of the one transition value function and different assumptions regarding the dynamics and timing of information availability regarding the unknown parameter are topics for future study. Two other topics for future study are (1) an analysis of the computational feasibility of the algorithms presented in Sections 1 and 3 (related discussions can be found in Smallwood and Sondik, and Sondik) and (2) the development of parametric programming and successive approximation procedures for the infinite horizon case and their comparison to the policy iteration algorithm presented in Section 3.

Acknowledgment

This research has been supported by ONR Contract N00014-80-C-0542, Work Unit No. N197-065, and NSF Grants ECS-8018266 and ECS-8319355.

References

BERTSEKAS, D. P. 1976. *Dynamic Programming and tochastic Control.* Academic Press, New York.

BERTSEKAS, D. P., AND I. B. RHODES. 1973. Sufficiently Informative Functions and the Minimax Feedback Control of Uncertain Dynamic Systems. *IEEE Trans. Aut. Control*, **AC-18,** 117–124.

BROWN, R. V., A. S. KAHR, AND C. PETERSON. 1974. *Decision Analysis for the Manager.* Holt, Rinehart & Winston, New York.

CYERT, R. M., AND DEGROOT, M. H. 1975. Adaptive Utility. In *Adaptive Economics*, pp. 223–246, R. H. Day (ed.). Academic Press, New York.

DENARDO, E. V. 1967. Contraction Mappings in the Theory Underlying Dynamic Programming. *SIAM Rev.* **9,** 165–177.

D'EPENOUX, F. 1960. Sur un Probleme de Production et de Stockage dans l'a Léatorie. *Rev. Fr. Inform. Rech. Opnl.* **14,** 3–16. English translation: 1963 *Mgmt. Sci.* **10,** 98–108.

FIGUERAS, J. 1972. The Restriction Set Approach to Stochastic Decision Systems. Ph.D. thesis, University of Michigan, Ann Arbor, Mich.

FISHBURN, P. C. 1965. Independence in Utility Theory with Whole Product Sets. *Opns. Res.* **13,** 28–45.

HENIG, M. I. 1978. Multicriteria Dynamic Programming. Ph.D. thesis, Yale University, New Haven, Conn.

HILLIER, F. S., AND G. J. LIEBERMAN. 1980. *Introduction to Operations Research* (3/E). Holden-Day, San Francisco, Calif.

HOWARD, R. 1960. *Dynamic Programming and Markov Processes.* MIT Press, Cambridge, Mass.

KEENEY, R. L., AND H. RAIFFA. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, Wiley, New York.

KREPS, D. M. 1977. Decision Problems with Expected Utility Criteria; II. Stationarity. *Math. Opns. Res.*, **2,** 266–274.

MEYER, R. F. 1977. State-Dependent Time Preference. In *Conflicting Objectives in Decisions*, pp. 232–243, D. E. Bell, R. L. Keeney and H. Raiffa (eds.). Wiley, Chichester, England.

PLATZMAN, L. K., C. C. WHITE, AND J. L. POPYACK. 1984. Optimally Damped Successive Approximation Algorithms for Markov Decision Programming. In preparation.

PUTERMAN, M. L., AND SHIN, C. S. 1978. Modified Policy Iteration Algorithms for Discounted Markov Decision Problems. *Mgmt. Sci.* **24,** 1127–1137.

PUTERMAN, M. L., AND C. S. SHIN. 1982. Action Elimination Procedures for Modified Policy Iteration Algorithms. *Opns. Res.* **30,** 301–318.

SAGALOVSKY, B. 1982. Adaptive Control and Parameter Estimation in Markov Chains: A Linear Case. *IEEE Trans. Aut. Control*, **AC-27,** 414–419.

SARIN, R. K. 1977a. Screening of Multiattribute Alternatives. *Omega* **13,** 481–489.

SARIN, R. K. 1977b. Interactive Evaluation and Bound Procedure for Selecting Multi-attributed Alternatives. *TIMS Stud. Mgmt. Sci.* **6,** 211–224.

SCHWEITZER, P. J. 1971. Iterative Solution of the Functional Equations of Undiscounted Markov Renewal Programming. *J. Math. Anal. Appl.* **34,** 495–501.

SMALLWOOD, R. D. 1966. Optimum Policy Regions for Markov Processes with Discounting. *Opns. Res.* **14,** 658–669.

SMALLWOOD, R. D., AND E. J. SONDIK. 1973. The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon. *Opns. Res.* **21,** 1300–1322.

SONDIK, E. J. 1978. The Optimal Control of Partially Observable Markov Processes Over the Infinite Horizon: Discounted Costs. *Opns. Res.* **26,** 282–304.

SWORDER, D. D. 1966. *Optimal Adaptive Control Systems.* Academic Press, New York.

VAN HEE, K. M. 1978. Bayesian Control of Markov Chains. In *Mathematical Centre Tracts, 95.* Mathematisch Centrum, Amsterdam.

VEINOTT, A. F., JR. 1969. Discrete Dynamic Programming with Sensitive Discount Optimality Criteria. *Ann. Math. Stat.* **40,** 1635–1660.

VISWANATHAN, B., V. V. AGGARWAL, AND K. P. K. NAIR. 1977. Multiple Criteria Markov Decision Processes. *TIMS Stud. Mgmt. Sci.* **6,** 263–272.

WHITE, C. C. 1984. Sequential Decisionmaking under Uncertain Future Preferences. *Opns. Res.* **23,** 148–168.

WHITE, C. C., AND K. W. KIM. 1980. Solution Procedures for Vector Criterion Markov Decision Processes. *Large Scale Syst.* **1,** 129–140.

WHITE, C. C., S. DOZONO, AND W. T. SCHERER. 1983. An Interactive Procedure for Aiding Multiattribute Alternative Selection. *Omega* **11,** 212–214.

WHITE, D. J. 1963. Dynamic Programming, Markov Chains, and the Method of Successive Approximations. *J. Math. Anal. Appl.* **6,** 373–376.