DISCUSSION PAPER NO. 70

AN ITERATIVE PROCEDURE FOR NON-DISCOUNTED DISCRETE-TIME MARKOV DECISIONS

Gary J. Koehler,* Andrew B. Whinston,** and Gordon P. Wright**

January 25, 1974

* Northwestern University
Evanston, Illinois

** Purdue University
West Lafayette, Indiana

Introduction

This paper utilizes the linear programming procedure given by Denardo [1] for finding a 1-optimal policy (by repeated application) in discrete-time Markov Decisions with an infinite horizon and no discounting. The problem to be addressed is the efficient solution of the sequence of linear programming problems.

Consider the following linear program and its dual:

Primal:	Min $c'x$	Dual:	Max $b'\pi$
	subject to:		subject to:
	$Ax = b$		$A'\pi \leqq c$
	$x \geqq 0$		$b \geqq 0$
	$b \geqq 0$		

where A' is the transpose of A, A is m by n, b and π are m by 1, and x and c are m by 1. Let J be a set of column identifications of A and A_J a submatrix of A consisting of the columns of A listed in J.

Consider a basis J. Let

$$A_J = R_J - Q_J$$

such that R_J is non-singular and

(1) $x \geqq y$ implies that $R_J^{-1}Q_J x \geqq R_J^{-1}Q_J y$

(2) $\rho(R_J^{-1}Q_J) < 1$

where $\rho(\cdot)$ is the spectral radius. If there exists an R and Q for each feasible basis J then iterative methods may be used to solve linear programs.

Using superscripts on vectors to denote iteration counts and superscripts on matrices to denote a power an iterative procedure can be specified. Let π^o be an arbitrary vector, J an initial basis, and parameters K and δ

be specified. The iterative procedure is:

Step One:

$$\pi^{n+K+1} = (R_J^{-1}Q_J)^{K+1}\pi^n + \sum_{\ell=0}^{K}(R_J^{-1}Q_J)^{\ell} R_J^{-1}c_J$$

Step Two:

Choose another basis, M, such that:

$$R_M^{-1}Q_M\pi^{n+K+1} + R_M^{-1}c_M \geq R_J^{-1}Q_J\pi^{n+K+1} + R_J^{-1}c_J$$

Step Three:

If $||R_M^{-1}Q_M\pi^{n+K+1} + R_M^{-1}c_M - R_J^{-1}Q_J\pi^{n+K+1} - R_J^{-1}c_J|| \leq \delta$ stop; otherwise set π^n to $R_M^{-1}Q_M\pi^{n+K+1} + R_M^{-1}c_M$ and label M by J. Go to Step One.

where δ is a tolerance and K is the number of refinements in Step One.

The advantages of using iterative methods in linear programming stem primarily from deletion of the need for storing and maintaining a basis inverse. Storage requirements are reduced, computational effort is reduced, and arithmetic precision is easily controlled.

The concept of using iterative methods in linear programming is now applied to the problem of finding a 1-optimal policy in discrete-time Markov decisions with an infinite horizon and no discounting. Since such problems may be large and, at the very least, require the repetitive solution of linear programs, the use of iterative methods for solving linear programs appears desirable.

Non-Discounted Discrete Time Markov Decisions

At each stage a system is observed to be in one of a finite set of states S. For any state $i \in S$ observed at time t (= 0,1,2,...), a decision $a \in A_i$ is made. Each set A_i is finite. The outcome of the decision is an immediate expected reward $c_i(a)$ and a probability of moving to state j in the next stage. The conditional probability is denoted by $P_{ij}(a)$. A stationary nonrandomized policy δ is a vector valued function that specifies a decision for each state. That is, for each $i \in S$, $\delta(i) \in A_i$. Let F be the set of all nonrandomized stationary policies. Then each policy $\delta \in F$ has associated with it an N x 1 vector of expected rewards $c(\delta)$ and an N x N matrix of transition probabilities $P(\delta)$ where N is the number of states in S.

For a discounted infinite horizon problem, the vector, $\pi(\delta)$, represented the present values for each state and is given by:

$$\pi(\delta) = \sum_{t=0}^{\infty} \alpha^t P(\delta)^t c(\delta)$$

where α is the discount factor and $0 < \alpha < 1$.

A non-discounted infinite horizon problem may be solved by considering the limiting process of $\pi(\delta)$ as $\alpha \to 1^-$. Miller and Veinott [4] have given a complete characterization for finding the 1-optimal (or average overtaking optimal) policy. They have also demonstrated the existence of α-optimal policies and have derived a Laurent series expansion [5].

To find a 1-optimal policy, the policies optimizing the first term of the Laurent series expansion are found. From these, the set of policies optimizing the second term is determined. This process continues until the 1-optimal policy is determined.

In this paper, the linear programming method given by Denardo [1] for finding bias optimal policies (and 1-optimal policies by repeated applications) is modified to allow efficient iterative methods [2] in the determination of 1-optimal policies.

Throughout this paper only problems possessing irreducible state spaces for every $\delta \in F$ are considered. Using extensions given by Denardo [1] and Koehler, Whinston, and Wright [2], more complicated structures can be considered in a natural manner.

Gain Optimality

When only policies optimizing the gain or average expected reward (the first term of the Laurent series expansion) are desired, Manne's [3] linear programming formulation may be used. The formulation is:

$$(1) \qquad \text{Min} \sum_{i \in S} \sum_{a \in A_i} c_i(a) x_{ia}$$

Subject to

$$\sum_{i \in S} \sum_{a \in A_i} x_{ia} = 1$$

$$\sum_{i \in S} \sum_{a \in A_i} (\delta_{ij} - P_{ij}(a)) x_{ia} = 0 \qquad \forall i \in S$$

$$x_{ia} \geq 0 \qquad i \in S,\ a \in A_i$$

where

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Ignoring the last constraint equation (which is redundant and will be associated with state s), the formulation given in (1) can be expressed in matrix notation as:

(2) $$\text{Min} \quad c_1' x_1 + c_2' x_2$$

Subject to

$$\begin{pmatrix} e_1' & e_2' \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$x_1, x_2 \geq 0$$

where e is a vector of ones and e' its transpose. x_1 and c_1 are vectors containing, respectively, the x_{ia} and $c_i(a)$ corresponding to all $a \in A_i$, $i \in S$ except $i = s$. x_2 and c_2 are correspondingly associated with x_{sa} and $c_S(a)$ for $a \in A_S$. Similarly, e_1 and e_2 and A_{21} and A_{22} correspond to the constraints of the original problem.

The above partitioning is such that A_{22} is non-positive and A_{21} contains exactly one positive element per column. Furthermore, by the nature of the constraint coefficients, the rows of A_{21} are non-trivial. Hence A_{21} is a Leontief matrix [6].

The constraint set of equation (2) does not permit the ready application of iterative solution procedures. Although such is the case, if a nonsingular transformation T can be found such that

$$T \begin{pmatrix} 1 \\ 0 \end{pmatrix} \geq 0$$

and

$$T \cdot \begin{pmatrix} e_1' & e_2' \\ A_{21} & A_{22} \end{pmatrix}$$

is Leontief. Thus, iterative procedures for solving the linear program in equation (2) may be used.

Consider

$$T = \begin{pmatrix} I_1 & -y' \\ 0 & I_2 \end{pmatrix}$$

where I is an identity matrix and y is a vector satisfying

$$A_{21}' y \geq e_1$$

Such a vector can be found by solving

(3) $$\text{Min} \quad y'e$$

Subject to

$$A_{21}' y \geq e_1$$

This problem is readily solved using iterative procedures outlined in an earlier paper [2] since A_{21} is Leontief.

Since S is irreducible for all $\delta \in F$ and A_{21} is Leontief, then $y \geq 0$. Hence, an equivelent representation of equation (2) is:

(4) $$\text{Min} \quad c_1' x_1 + c_2' x_2$$

Subject to

$$\begin{pmatrix} e_1'-y'A_{21} & e_2'-y'A_{22} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$x_1, x_2 \geq 0$$

The compelling attribute of the above formulation is that the constraint set is Leontief and efficient iterative procedures [2] may be used to solve the problem. The optimal dual values to equation (2) are found by pre-multiplying the optimal dual values to equation (4) by $\bar{T}'$.

Bias Optimal Policies

Denardo [1] has demonstrated that a bias optimal (the second term of the Laurent series expansion) policy may be found by restricting the policy space and altering the cost structure of the gain optimal policies. After such changes, another linear program is solved to determine the bias optimal policies.

Consider the linear program given in equation (4). In compact form, equation (4) may be written as:

(5) $$\text{Min} \quad c'x$$

Subject to

$$Bx = b$$

$$x \geq 0$$

Let J be a set of column identifications and B_J a submatrix of B consisting of the columns labelled in J. Let

$$N = \underset{i \in S}{\{(a,i) : a \in A_i\}}$$

Then B_J, $J \in N$ is a feasible basis to the problem in equation (5). Actually, B_J may be expressed as

$$B_J = I - P'(\delta)$$

where J and some policy δ correspond in a natural manner.

Let the set K be a subset of N constructed using the ideas of Denardo's Problem II [1]. Furthermore, let $\bar{c}$ be the vector of altered costs according to Denardo's Problem II. Then equation (5) may be modified to:

$$\text{MIN} \quad \bar{c}'x$$

$$\text{Subject to}$$

$$Bx = b$$

$$x \geq 0$$

and only basis B_J, $J \in K \subseteq N$ are considered. Since the constraint set has not been changed, the problem is still a Leontief structured linear program and may be solved using iterative methods [2]. The restricted basis entry is the only imposition on the algorithm (which, incidentally, simplifies the computational procedure). Hence, a bias optimal policy may be found by changing the cost structure and basis entry rules while leaving the constraint set unchanged. Furthermore, iterative methods may be used for the determination of the bias optimal policy.

1 - Optimal Policy

As Denardo [1] has pointed out, by repeated application of the above procedure, a 1-optimal policy may be determined.

It is important to note that the constraint set remains Leontief once Manne's original problem has been converted to a Leontief structured linear Program. In all cases, iterative procedures for solving linear programs may be employed.