DOCUMENT RESUME

ED 124 161 IR 003 551

AUTHOR Wollmer, Richard D.
TITLE Partially Observable Markov Decision Processes Over an Infinite Planning Horizon with Discounting. Technical Report No. 77.
INSTITUTION University of Southern California, Los Angeles. Behavioral Technology Labs.
SPONS AGENCY Advanced Research Projects Agency (DOD), Washington, D.C.; Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
PUB DATE Mar 76
CONTRACT N0014-75-C-0838
NOTE 25p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Computer Assisted Instruction; *Decision Making; Instructional Systems; *Linear Programing; Mathematical Applications; *Mathematical Models; Operations Research; Probability Theory; *Systems Approach
IDENTIFIERS Markov Models

ABSTRACT

The true state of the system described here is characterized by a probability vector. At each stage of the system an action must be chosen from a finite set of actions. Each possible action yields an expected reward, transforms the system to a new state in accordance with a Markov transition matrix, and yields an observable outcome. The problem of finding the total maximum discounted reward as a function of the probability state vector may be formulated as a linear program with an infinite number of constraints. The reward function may be expressed as a partial N-dimensional Maclaurin series. The coefficients in this series are also determined as an optimal solution to a linear program with an infinite number of constraints. A sequence of related finitely constrained linear programs is solved which then generates a sequence of solutions that converge to a local minimum for the infinitely constrained program. This model is applicable to computer assisted instruction systems as well as to other situations. (Author/CH)

DEPARTMENT OF PSYCHOLOGY

UNIVERSITY OF SOUTHERN CALIFORNIA

Technical Report No. 77

PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES OVER AN INFINITE PLANNING HORIZON WITH DISCOUNTING

March 1976

Richard D. Wollmer

Sponsored by

Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

and

Advanced Research Project Agency
Under Contract No. N00014-75-C-0838

ARPA TECHNICAL REPORT

March 1976

1.	ARPA Order Number	: 2284
2.	ONR NR Number	: 154-355
3.	Program Code Number	: 1 B 729
4.	Name of Contractor	: University of Southern California
5.	Effective Date of Contract	: 76 January 1
6.	Contract Expiration Date	: 76 December 31
7.	Amount of Contract	: $200,000.00
8.	Contract Number	: N00014-75-C-0838
9.	Principal Investigator	: J. W. Rigney (213) 746-2127
10.	Scientific Officer	: Marshall Farr
11.	Short Title	: Learning Strategies

This Research Was Supported

by

The Advanced Research Projects Agency

and by

The Office of Naval Research

And Was Monitored by

The Office of Naval Research

SUMMARY

This is the last in a series of technical reports concerned with mathematical approaches to instructional sequence optimization in instructional systems. The problem treated here is very closely related to that treated by Smallwood and Sondik (4). Both papers deal with Markov decision processes where the true state of the system is not known with certainty. Hence the state of the system is characterized by a probability vector. Each action yields an expected reward, transforms the system to a new state and yields an observable outcome. One wishes to determine an action for each probability state vector so as to maximize the total expected reward. Smallwood and Sondik (4) solve this problem exactly for a finite time horizon. This report treats the infinite time horizon with a discount factor, using a partial N dimensional Maclaurin series to approximate the total optimal reward as a function of the probability state vector. While this model was developed for computed aided instruction, it is applicable to other situations as well. This model also is of considerable theoretical value.

ABSTRACT

This paper describes a system that may be in any one of states $1,2,\ldots,N$. The true state of the system is not known with certainty and consequently is described by a probability vector. At each stage an action must be chosen from a finite set. Each possible action returns an expected reward, transforms the system to a new state in accordance with a Markov transition matrix, and yields an observable outcome. It is required to determine an action for each possible state vector in order to maximize the total expected reward over an infinite time horizon under a discount factor, β, where $0<\beta<1$.

The problem of finding the total maximum discounted reward as a function of the probability state vector may be formulated as a linear program with an infinite number of constraints. The reward function may be expressed as an N dimensional Maclaurin series and in this paper it is approximated by a partial series consisting of terms up to degree n. The coefficients in this series are also determined as an optimal solution to a linear program with an infinite number of constraints. A sequence of related finitely constrained linear programs are solved which generate a sequence of solutions that converge to a local minimum for the infinitely constrained program. It is an open question as to whether this local minimum is actually a global minimum. However it should be noted that the function being approximated is convex and consequently has the property that any local minimum is a global one as well.

ACKNOWLEDGEMENTS

The research discussed in this report was monitored by Dr. Marshall Farr and Dr. Joseph Young, Personnel and Training Research Programs, Office of Naval Research, and Dr. Harry F. O'Neil, Jr., Program Manager, Human Resources Research, Advanced Research Projects Agency. Their support and encouragement is gratefully acknowledged.

TABLE OF CONTENTS

PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES OVER AN INFINITE PLANNING HORIZON WITH DISCOUNTING

1. Introduction

This paper describes a system that may be in anyone of states 1, 2,...,N. The true state of the system is not known with certainty and consequently is described by a probability vector. At each stage an action must be chosen from a finite set. This action returns an expected reward, transforms the system to a new (but not necessarily different) state according to a Markov process, and yields an observable outcome. The problem addressed here is that of determining an action for each possible state vector in order to maximize the total expected reward over an infinite horizon under a discount factor, β, where $0<\beta<1$.

Smallwood and Sondik (4) have treated this problem for the finite horizon case without a discount factor and have determined that the total maximum expected reward is a piecewise linear function of the probability state vector. Their results can be trivially extended to include the discount case.

The observable state case, that is the case where the true state of the system is known with certainty has been treated extensively. For both the finite and infinite horizon under a discount factor, Howard (1) developed a policy improvement routine for determining an optimal action and the optimal cost for each state.

II. Formulation

In this formulation, the notation of Smallwood and Sondik will be used. It is assumed that this system can be modeled by an N-state discrete time Markov decision process.

The observed state of the system is characterized by a probability vector π where π_i is the probability the true state of the system is i.

At each point in time an action must be selected from a finite set. Associated with an action, a, is a probability transition matrix P^a where P^a_{ij} is the conditional probability the system will make its next transition to state j given the current state is i and action a is taken. An observed outcome follows each action with $r^a_{j\theta}$ denoting the probability of observing output θ given the new state of the system is j and action a was taken. In addition an immediate reward $w^a_{ij\theta}$ is incurred if action a is taken, output θ is observed, and the system makes the transition from state i to state j. Thus if action a is taken and output θ is observed, the new state is π' where

$$\pi'_j = \left[\Sigma_i \pi_i P^a_{ij} r^a_{j\theta}\right] \Big/ \left[\Sigma_{ij} \pi_i P^a_{ij} r^a_{j\theta}\right] \qquad (1)$$

The above transformation is summarized by

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

A policy is a rule that assigns an action to each possible state vector. It is required to find a policy that maximizes the expected discounted rewards over all periods for each possible state vector. Let $V(\pi)$ be the total discounted reward associated with such a policy.

Then $V(\pi)$ must satisfy the following recursive equation.

$$V(\pi) = \max_a \left[\sum_{i=N}^{N} \pi_i \sum_{j=1}^{N} p_{ij}^a \sum_{\theta} r_{j\theta}^a \left\{ w_{ij\theta}^a + \beta T(\pi/a,\theta) \right\} \right] \quad (3)$$

$$\text{Letting } q_i^a = \sum_{j,\theta} p_{ij}^a r_{j\theta}^a w_{ij\theta}^a \quad (4)$$

equation (3) is simplified somewhat to equation (5)

$$V(\pi) = \max_a \left[\sum \pi_i q_i^a + \beta \sum_{i,j,\theta} \pi_i p_{ij}^a r_{j\theta}^a V\left[T(\pi/a,\theta)\right] \right] \quad (5)$$

Once the function for $V(\pi)$ is known, an optimal action for π can can be determined as one which maximizes the right hand side of (5).

III. A Learning Example

As an illustration, it will be shown how the system described in the previous section may be applied to the human learning process.

Consider a course which is given in several levels of instruction. The levels are denoted 1, 2,...,N with N being the easiest and 1 the hardest. The structure of the levels is a definite hierarchy in the sense that if a student knows the material at level i he must also know the material at any level j>i. Several examples where this situation may apply follow:

The first situation is one where the material covered at one level includes all that covered at preceding levels, plus some additional material. An example of this is a program developed at Behavioral Technology Laboratories (BTL) to teach students Kirchoff's Laws. This course is comprised of eleven levels with the lowest level defining the units for voltage, current and resistance up to the highest level which deals with the application of Ohm's Law and Kirchoff's voltage and current laws in complex networks. Another program developed at BTL is a short course in trigonometry consisting of five levels. At the lowest level students are given the definitions of the six basic trigonometric ratios. Then the student is given a right triangle in which the lengths of the sides are determined by a random number generator and the student is asked to determine these ratios for one of the acute angles. Succeeding levels deal with material on relationships between these ratios and problems testing the student's knowledge of these relationships.

A second situation is one where the material and problems covered at a particular level are virtually the same as the immediately preceding level except more clues and hints are given at the preceding level. A good example of this is a version of the Kirchoff's laws program considered

earlier at BTL in which problems would be given in level as follows:

1. Problems are given in steps with cues and knowledge of results at each step.

2. Problems are given in steps with no cues or knowledge of results at each step.

3. The student solves problems in steps but he chooses the steps.

4. The student is simply given problems and asked to solve them.

A third situation is one in which a student is to be drilled in a skill in order that he be able to perform it rapidly. Thus the exercises are virtually the same at all levels but the time constraints are tighter at the higher levels. In the BTL intercept trainer for the radar intercept observer function, the student is trying to fire a missile at the nose of a target and then turn around and fire another missile at the tail of that aircraft. The first missile is a radar guided missile fired when in the forward quarter and the second a heat seeker fired when in the rear quarter of the enemy aircraft. He is given a radar reading and must correct his angle of approach so as to be on a lead collision course that will insure a high hit probability when he fires the missile. At higher levels the student is given such problems at faster aircraft speeds.

Note, however, the assumption given for this model would not be applicable for the situation where a given level did not use certain material introduced at preceding levels.

A student is in state i if he knows the material of level i but not at any level more difficult than i and in state N+1 if he does not know the material at any level.

There are N actions and action i consists of instructing the student in the material of level i and then giving the student a test

on that material. For each action there are two possible outcomes--either the student passes the test or he fails it. The objective is to develop an adaptive instructional sequence so that the student demonstrates knowledge of the material at level 1 as quickly as possible. Knowledge at level 1 is demonstrated by passing a test on the material at level 1. The reward, $w^a_{ij\theta}$, would be the negative of the expected time it would take to obtain instruction at level a and the system goes from state i to state j and θ (success or failure at a) is observed. For completeness a trap state ϕ would be needed. The student goes to state ϕ with probability one once he successfully completes the material at level 1. The only action in state ϕ is to do nothing which yields a zero reward and keeps the student in state ϕ with probability one.

Wollmer (6) treats the more restricted problem where $p^a_{ij} = 0$ unless i=j or if i=a and j=i+1. Thus if a student is in state i, he remains in state i unless he receives instruction at level i+1, in which case he either remains in state i or advances to state i+1. This would not allow the possibility of forgetting.

Other situations where partially observable Markov Decision processes occur are in machine replacement, decoding from sources transmitting over a noisy channel, medical diagnosis, and searching for a moving object.

Note, that if the assumption of a strict hierarchy in levels were dropped, the set of states would expand from N+2 to 2^N+1 including the trap state.

IV. The Maximum Reward Function

In this section it will be shown that a maximum reward function exists and that it is a convex function of the reward π.

Let $V_n(\pi)$ be the maximum reward function for the n period horizon. Then

$$V_n(\pi) = \max_a \left[\sum_i \pi_i q_i^a + \beta \sum_{i,j,\theta} p_{ij}^a r_{j\theta}^a V_{n-1}\left[T(\pi/a,\theta)\right] \right] \qquad (6)$$

Smallwood and Sondik (4) have shown that $V_n(\pi)$ is *

1. Convex
2. Piecewise Linear

It will be shown that $\lim_{n \to \infty} V_n(\pi)$ exists and is convex in π.

Define f_n so that $|V_n(\pi) - V_{n-1}(\pi)| \leq f_n$ all n and f_n is the smallest real number with this property and $V_o(\pi) = 0$. The f_n's are well defined since all $V_n(\pi)$ are bounded above and below.

Lemma 1: $f_{n+1} \leq \beta f_n$

Proof : Choose $a(\pi)$ as the action that maximizes the right hand side of (6) for $V_{n+1}(\pi)$ if $V_{n+1}(\pi) \geq V_n(\pi)$ or for $V_n(\pi)$ otherwise.

Then $|V_{n+1}(\pi) - V_n(\pi)| \leq |\beta \sum_{i,j,\theta} p_{ij}^a r_{j\theta}^a (V_n[T(\pi/a,0)]$

$$-V_{n-1}[T(\pi/a,0)]| \leq \beta f_n.$$

Corollary 1: For $n^* > n$, $|V_{n^*}(\pi) - V_n(\pi)| < \varepsilon(n)$

where $\varepsilon(n) \to 0$.

---

*While Smallwood and Sondik assume $\beta=1$, their results hold for $0<\beta\leq 1$.

Proof: From lemma 1, $f_n \leq \beta^{n-1} f_1$ and consequently

$$|V_n^*(\pi) - V_n(\pi)| \leq \sum_{i=n+1}^{n^*} f_i \leq \beta^n f_1 \sum_{i=0}^{\infty} \beta^i = f_1 \beta^n/(1-\beta)$$

Theorm 1: The function $V_n(\pi)$ is absolutely convergent.

Proof: Choose any particular $\pi=\bar{\pi}$. By Corollary 1, the $V_n^*(\pi)$ is bounded above and below and hence has an infinite covergent subsubseqûence with limit $V^*(\pi)$. Choose $\varepsilon > 0$ and n such that $\varepsilon(N) < \varepsilon$ for $N \geq n$ and $\varepsilon(n)$ is as defined in corollary 1. For any $N \geq n$ and $\bar{n} \geq n$ in the convergent subsequence $|V_N(\pi) - V_{\bar{n}}(\pi)| < \varepsilon$ and consequently $|V_N(\pi) - V^*(\pi)| < \varepsilon$. Since n is independent of π, the theorem is proven.

Thus $V(\pi) = \lim_{n \to \infty} V_n(\pi)$ is well defined.

Theorem 2: $V(\pi)$ is convex in π.

Proof: Define $f(V,\pi_1,\pi_2) = V(\frac{1}{2}\pi_1 + \frac{1}{2}\pi_2) - \frac{1}{2}V(\pi_1) - \frac{1}{2}V(\pi_2)$.

Assume $V(\pi)$ is not convex and choose π_1 and π_2 such that $f(V,\pi_1,\pi_2) = k > 0$. Choose n such that $N > n \rightarrow |V_N(\pi) - V(\pi)| < K/2$. $|f(V,\pi_1,\pi_2) - f(V_N,\pi_1,\pi_2)| < K$. Thus $f(V_N,\pi_1,\pi_2) > 0$ which is impossible since $V_N(\pi)$ is convex.

Note, that the piecewise linear property of $V_n(\pi)$ does not imply piecewise linearity of $V(\pi)$ as any continuous function may be expressed as the limit of a sequence of piecewise linear functions.

V. Linear Program Formulation

In the case of the observable finite state Markov decision processes with a discount factor, the problem of finding a maximum return for each state may be formulated as a linear program. The development of this may be found in Ross (6). In this section it is shown that a modification of this formulation extends to the problem formulated in Section II. Portions of the development which are similar to the finite state case will be outlined but without rigorous proofs.

Consider the set B of all continuous bounded functions defined on $S = \left\{\pi/\pi_i \geq 0 \text{ all } i, \sum_{i=1}^{n} \pi_i = 1\right\}$. Let the operator A be defined on this set as follows.

$$Au(\pi) = \max_a \left[\sum_i \pi_i q_i^a + \beta \sum_{i,j,\theta} \pi_i p_{ij}^a r_{j\theta}^a U\,[T(\pi)/a,\theta)]\right] \qquad (7)$$

Note that

1. $u \leq v \rightarrow Au \leq A_v$
2. $Au \in B$ all $u \in B$
3. $A: B \rightarrow B$ is a contraction mapping on B.

The Operator A is the optimal return function for the one period problem in which a terminal reward $u(\pi)$ is given for the terminal state. Since $A: B \rightarrow B$ is a contraction mapping, it has a unique fixed point, $V = Av = \lim_{n \rightarrow \infty} A^n u$ for any $u \in B$. By Equation (3), this unique fixed point must be the optimal reward function. Let us consider any u such that $Au \leq u$. Then $u \geq Au \geq A^2 u \geq \lim_{n \rightarrow \infty} A^n u = v$. Thus the optimal return function V minimizes $u(\pi)$ for each $\pi \in S$ among all functions u satisfying $Au \leq u$.

In the finite state case where the above conditions also hold, it is noted that minimizing u_i for each state i may be accomplished by

minimizing the sum of the u_i's. For this problem where such a sum would be infinite, the average value of $u(\pi)$ may be minimized. Thus, finding the function $u(\pi)$ is equivalent to solving the following infinite constrained program.

Find min z, u such that

$$Z = \int_{\sum \pi_i = 1,\ \pi_i \geq 0} \cdots \int u(\pi) d\pi_n d\pi_{n-1} d\pi_{n-2} \cdots d\pi_1 \tag{8}$$

subject to

$$\sum_i \pi_i q_i^a + \beta \sum_{i,j,\theta} \pi_i p_{ij}^a r_{j\theta}^a u[T(\pi/a,\theta)] \leq u(\pi) \text{ for} \tag{9}$$

$$\pi_i \geq 0,\ \sum_i \pi_i = 1$$

Since the function $u(\pi)$ is continuous and defined on a closed bounded set, it may be expressed in an N-dimensional Maclaurin series:

$$V(\pi) = C_o + \sum_{i_1,\ i_2,\ldots,i_n} C_{i_1,i_2,\ldots i_n} \pi_1^{i_1} \pi_2^{i_2} \cdots \pi_N^{i_N} \tag{10}$$

If $V(\pi)$ is expressed as such a series or approximated by a partial series consisting of terms up to degree n, the coefficient of $C_{i_1,i_2,\ldots i_n}$ in (8) is simply

$$\int_0^1 \pi_1^{i_1} \int_0^{1-\pi_1} \pi_2^{i_2} \cdots \int_0^{1-\pi_1-\pi_2-\cdots-\pi_{N-1}} \pi_N^{i_N} d\pi_N d\pi_{N-1} \cdots d\pi_1 \tag{11}$$

In evaluating the integral the following lemma is needed.

Lemma 2: $\int_o^a (a-x)^m x^n dx = \frac{m!\ n!}{(m+n+1)!} a^{m+n+1}$

Proof: Integrating by parts one obtains for the above integral $-\frac{1}{m+1}x^{n-1}(a-x)^{m}\Big|_{o}^{a} + \frac{n}{m+1}\int_{o}^{a}(a-x)^{m+1}x^{n-1}dx$

$= \frac{n}{m+1}\int_{o}^{a}(a-x)^{m+1}x^{n-1}dx$. Applying this relationship recursively, one obtains $\frac{n!m!}{(m+n)!}\int_{o}^{a}(a-x)^{m+n}dx = \frac{m!n!}{(m+n+1)!}a^{m+n+1}$

From this lemma, expression (11) can be evaluated.

Theorem 3: The value of expression (11) is $\prod_{j=1}^{n} i_j! \Big/ \left[\sum_{j=1}^{n}(i_j+1)\right]!$

Proof: Integrating (11) with respect to π_n gives

$$\frac{i_n!}{(i_n+1)!}\int_0^1 \pi_1^{i_1}\int_0^{1-\pi_1}\pi_2^{i_2}\cdots\int_0^{1-\sum_1^{n-2}\pi_j}(1-\sum_1^{n-1}\pi_j)^{i_n+1}\pi_{n-1}^{i_{n-1}}\,d\pi_{n-1}\cdots d\pi_1$$

Applying lemma 1 with $a=1-\sum_1^{n-2}\pi_j$ and integrating with respect to π_{n-1} yields

$$\frac{i_n!\ i_{n-1}!}{(i_n+i_{n-1}+2)!}\int_0^1 \pi_1^{i_1}\int_0^{1-\pi_1}\pi_2^{i_2}\cdots\int_0^{1-\sum_1^{n-3}\pi_j}(1-\sum_1^{n-1}\pi_j)^{i_n+i_{n-1}+2}\,d\pi_{n-2}\cdots d\pi_1$$

Continual application of lemma 2 yields $\prod_{j=1}^{n} i_j! \Big/ \left(\sum_{j=1}^{n}(i_j+1)\right)!$

Thus if $V(\pi)$ is to be approximated by an n^{th} degress polynomial function in π, then substituting the expression of theorem 3 and (1) in (8) and (9) and rearranging terms yields:

Find C_o, $C_{i_1 i_2 \ldots i_n}$, min z such that

$$z = C_o + \sum\left[\left(\prod_{j=1}^{n} i_j!\right)\Big/\left(\sum_{j=1}^{n}(i_j+1)\right)!\right] C_{i_1, i_2 \ldots i_n} \qquad (12)$$

$$(1-\beta)C_0 + \sum \left(\prod_{j=1}^{n} \pi_j^{i} - \beta\Sigma_\theta \left[k^{(\theta)}_{i_1 i_2 \ldots i_N} / d^{(\theta)}_{i_1 i_2 \ldots i_n} \right] \right) \times$$

$$C_{i_1 i_2 \ldots i_n} \geq \sum_{i=1}^{n} \pi_i q_i^a \qquad (13)$$

$$\text{where } k^{(\theta)}_{i_1 i_2 \ldots i_n} = \prod_{j=1}^{n} (\Sigma_i \pi_i p^a_{ij} r^a_{j\theta})^{i_j} \qquad (14)$$

$$d^{(\theta)}_{i_1 i_2 \ldots i_N} = (\sum_{ij} \pi_i p^a_{ij} r^a_{j\theta})^{\sum (i_j - 1)} \qquad (15)$$

for all θ, all $\pi \geq 0$ such that $\sum \pi_i = 1$

Thus the problem of solving the program (8-9) with a multinomial approximation of $u(\pi)$ becomes a linear program (12-15) with an infinite number of constraints and unrestricted variables. Note that the minimum value of $\bar{z}$ obtained in the linear program (12-15) would actually be larger than that obtained in the program (8-9).

VI. Computational Procedure

Given an optimal solution to the linear program (12-15), consider the set of constraints for which the $C_{i_1 i_2 \ldots i_N}$ are basic. If the program was solved with these constraints only, the same solution would be obtained and all other constraints would be satisfied. Thus, while the program consists of an infinite number of constraints, only a finite number need to be included provided the correct ones are chosen. This will be taken advantage of by solving the program with a finite subset of the constraints, introducing an unsatisfied constraint, then dropping any that are not binding, and continuing until an optimal solution is obtained.

Let the quantity $f(\pi,C)$ be defined as follows:

$$F(\pi,C) = (1-\beta)C_0 + \Sigma\left(\prod_{j=1}^{n} \pi_j^{i_j} - \beta\Sigma_{\theta}\left[k^{(\theta)}_{i_1 i_2 \ldots i_N} d^{(\theta)}_{i_1 i_2 \ldots i_N}\right]\right) \times$$

$$C_{i_1 i_2 \ldots i_N} - \sum_{i=1}^{n} \pi_i q_i^a \qquad (16)$$

The constraints (13) are equivalent to $F(\pi,C) \geq 0$ all π. Thus if at least one constraint is not satisfied for a given C vector, the value of π that minimizes $F(\pi,C)$ is the most unsatisfied one.

The procedure for solving the linear program (12-15) is given in algorithm 1.

Algorithm 1

1. Formulate the linear program with any finite subset of the constraints in (13).
2. Solve the linear program for C.
3. Delete any constraints for which a slack variable is basic.
4. Solve the following non-linear program.

Find $\pi \geq 0$, min Z such that

$$Z' = f(\pi, C) \tag{17}$$

$$\sum_{i=1}^{N} \pi_i = .1 \tag{18}$$

If $Z' \geq 0$, terminate as C is optimal. Otherwise introduce the constraint corresponding to the value of π that optimizes (17-18) and go back to Step 2.

A local optimum to (17-18) may be found by algorithm 2.

Algorithm 2

1. Choose an arbitrary probability vector and evalute $f(\pi, C)$.

2. Find an order pair (i,j) such that increasing π_i by ε and decreasing π_j by ε decreases $f(\pi, C)$ without violating $0 \leq \pi_i \leq 1$ and $0 \leq \pi_j \leq 1$. If no such pair can be found, terminate as π is a local optimum.

3. Increase π_i to $\bar{\pi}_i$ and decrease π_j to $\bar{\pi}_j$ such that neither the pair (i,j) or (j,i) satisfied the conditions of Step 2. Then go back to Step 2.

For finiteness, the ε of Step 2 would be chosen ahead of time.

There are several ways of performing Step 3 to find the new value of π_i and π_j. One efficient way is to first bracket π_i and π_j between π_i', π_j' and π_i'' and π_j'' and continually reduce the difference between these by a factor of one half, thus converging on a single point. Initially π_i' and π_j' would be the current values of π_i and π_j and $\pi_i'' = \pi_i + \delta$, $\pi_j'' = \pi_j - \delta$ where $\delta = \min [1-\pi_i, \pi_j]$. Then consider the pair $\bar{\pi}_i = \frac{1}{2}(\pi_i' + \pi_i'')$ and $\bar{\pi}_j = \frac{1}{2}(\pi_j' + \pi_j'')$. If $f(\bar{\pi}, C)$ is a local

minimum under the restriction that all components of π other than π_i and π_j are held constant, then $\bar{\pi}$ is the desired point. Otherwise, let $\bar{\pi}_i$ and $\bar{\pi}_j$ replace π_j' and π_j' if the direction of decrease is towards π_i'' and π_j'' but let $\bar{\pi}_i$ and $\bar{\pi}_j$ replace π_i'' and π_j'' if the direction of decrease is towards π_i' and π_j'. If neither direction yields a decrease, let $\bar{\pi}_i$ and $\bar{\pi}_j$ replace π_i' and π_j' if $f(\pi')>f(\pi'')$ but replace π_i'' and π_j'' otherwise. Step 3 would terminate when $\pi_i'' - \pi_i' < \varepsilon_1$ where $\varepsilon_1 < \varepsilon$.

Note that if the C vector approximation of $U(\pi)$ were exact, any local minimum of $f(\pi,C)$ would be a global minimum due to the convexity of $V(\pi)$. While this is not guaranteed in the approximation, one could take random samples of π in an attempt to find a vector yielding a lower value of z' than the local minimum or evaluate z' for all π vectors whose components are multiples of 1/n where n is large if the result min $z'=0$ is obtained.

When introducing an unsatisfied constraint, it is recommended that the dual simplex method be used to solve the resulting program which is already dual feasible.

The sequence of min z values generated by algorithm 1 is non-decreasing, bounded above, and hence must have a limit. It is an open question as to whether this limit is the true min z or in particular if the sequence of z' values in algorithm 2 tend to zero. Consider the sequence of linear programs solved by algorithm 1 and assume the number of equations in each equals the number of components in the C vector plus one. It has already been shown that it will not exceed this number and if it is less, additional constraints with all coefficients being zero may be added. Consider also the sequence of matrices formed by the probability vectors that generate these constraints. Since these

are bounded above, these matrices, and consequently the set of linear programs for algorithm 1 must have a convergent subsequence. Consider now the sequence of constraints generated by this sequence in algorithm 2. By the same argument this sequence must have a convergent subsequence. In this latter sequence, either $f(\pi,C)\to 0$ or else the cost coefficient in the pivot column tends to zero for if not the increase in min $\bar{z}$ would not tend to zero which is impossible since min $\bar{z}$ is bounded above.

If the sequence of $f(\pi,C)$ values generated by problem 2 did not appear to tend to zero after many iterations while the change in min $\bar{z}$ did appear to tend to zero, some possible ways out are as follows. First one may sample a large number of probability vectors and find one which would give the largest increase in $\bar{z}$ on a single pivot. Second, one may search all probability vectors that are multiples of $1/n$ where n is a large number and find the one which gives the largest increase in $\bar{z}$ for one pivot.

It should be noted that if the sequence of $\bar{z}$ values obtained in algorithm 2 do not tend to zero, then one has a situation somewhat analogous to cycling in the dual simplex method. Since cycling almost never occurs in the primal simplex method, there appears to be some basis for thinking that the sequence of $\bar{z}$ values would tend to zero the majority of times.

One could of course only consider constraints generated by probability vectors whose components are multiples of $1/n$. By imposing a lexicographic ordering, one could insure a true optimum in a finite number of steps.

VII. Bounds on Accuracy

In solving the non-linear program (17-18) in Step 4 of the algorithm to find the most unsatisfied constraint of the linear program (12-15), one may wish to terminate the program when $\bar{z} > -\delta$ rather than for $\bar{z} \geq 0$ where δ is a small positive number. If so, the value of $\bar{z}$ obtained for (12) will be less than the true minimum for $\bar{z}$ since the program has been optimized for only a subset of the constraints. However, it is easy to see from (12) and (13) that increasing C_0 by $\delta/(1-\beta)$ yields a feasible solution and increases $\bar{z}$ by that same amount. Consequently, this feasible set would come to within $\delta/(1-\beta)$ of minimizing $\bar{z}$.

The question now arises as to how close $\tilde{V}(\pi)$, the Maclaurin series approximation to $V(\pi)$, is to the true value of $V(\pi)$. To answer this consider the operator $Au(\pi)$ defined in equation (7) and define:

$$||Au - u|| = \max_{\pi} |Au - u| \tag{19}$$

Since the operator A is a constraction mapping with $|Au - Av| \leq \beta|u - v|$ it can be shown that $||A^{n+1}u - A^n u|| \leq \beta^n ||Au - u||$ and $||A^n u - u|| \leq (1-\beta^n)||A^n u - u||/(1-\beta)$ and $V(\pi) = \lim_{n \to \infty} A^n u$, it follows that

$$|V(\pi) - \tilde{V}(\pi)| \leq ||A\tilde{v} - \tilde{v}||/(1-\beta) \tag{20}$$

One could find a local maximum to $|A\tilde{v} - \tilde{v}|$ by an incremental procedure similar to that used to find the most unsatisfied constraint to introduce into the linear programming problem. Alternatively, one could enumerate (20) for all possible probability vectors whose components are multiples of 1/n.

REFERENCES

1. Howard, R. A. Dynamic Programming and Markov Processes, John Wiley and Sons, New York, 1960.

2. Manne, A. "Linear Programming and Sequential Decisions," Management Science 6, 259-267 (1960).

3. Ross, S. M. Applied Probability Models with Optimization Applications, Holden-Day, San Francisco, 1970.

4. Smallwood, R. D. & E. J. Sondik, "The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon," Operations Research 21, 1071-1087, (1973).

5. Wolfe, P., and Dantzig, G. B. "Linear Programming in a Markov Chain, Operations Research 10, 702-710 (1962).

6. Wollmer, R. D., "A Markov Decision Model for Computer-Aided Instruction," Behavioral Technology Laboratories, University of Southern California, Technical Report #72, December 1973.