AD-A032 231

Markovian Decision Processes with Limited State Observability and Unobservable Costs

Rand Corp Santa Monica Calif

Nov 72

ABSTRACT

MARKOVIAN DECISION PROCESSES WITH LIMITED STATE OBSERVABILITY AND UNOBSERVABLE COSTS

James D. Steele, Ph.D.

Consider a finite-state finite-action Markovian Decision Process for which the state space has been partitioned into subsets. The decisionmaker can only observe the subset to which the states of the process belong, and not the actual states of the process. In addition, the costs are unobservable in the sense that the total discounted cost is to be assessed at infinity. An approach to this problem, which makes use of the probability distributions over the state space, is developed.

MARKOVIAN DECISION PROCESSES WITH LIMITED STATE OBSERVABILITY AND UNOBSERVABLE COSTS

James D. Steele, Ph.D.
The Rand Corporation, Santa Monica, California

INTRODUCTION

Consider the situation in which a decisionmaker periodically observes a process, at times $t = 0,1,2,\ldots$, and at each observation classifies the process as being in one of a possible number of states. In the first section of this paper, we will require that the set of all possible states of the process be a finite set. In the later sections, we will consider situations in which the set of all possible states is uncountable. After each observation, the decisionmaker chooses an action from a set of possible actions. Throughout this paper, the set of all possible actions will be assumed to be a finite set. At this point a cost, which depends on the current state of the process and on the particular action chosen, is incurred and the next state of the process is chosen according to transition probabilities which depend on the current state and the particular action chosen. The objective of the decisionmaker is to choose actions in a manner such that some particular cost criterion is minimized. Throughout this paper, the cost criterion used will be the total expected discounted cost of operating over the infinite future. The above basically describes a Markovian Decision Process with the particular cost criterion as defined.

In the first section of this paper, we review some of the concepts and definitions associated with the finite-state finite-action Markovian Decision Process. We define the concept of a policy for taking actions for the decisionmaker and we develop the expressions for the expected discounted costs associated with the use of certain types of policies. In this section, we assume that the decisionmaker knows the current value of the state of the process at each observation point. Also in this section, we assume that the decisionmaker knows, immediately after observing the current state and taking an action, the value of the cost incurred at that point.

In the remaining sections of this paper, we consider a finite-state finite-action Markovian Decision Process in which the decisionmaker is not told the exact state of the process at the observation points. Rather the decisionmaker is only told that the current state belongs to a particular subset of possible states. We call this "limited state observability." The extreme case in which the decisionmaker is given no information about the current state (i.e. the subset to which the current state belongs is simply taken to be the entire set of all possible states of the process), is called "complete unobservability." In this case, we say the Markovian Decision Process has unobservable states. Also, in the remaining sections of this paper, we assume that the decisionmaker is not told the value of the costs incurred at any observation point. Rather we assume that the decisionmaker will be told the value of each current cost far enough into the future so that we may assume that the total cost will be assessed at infinity. In this case, we say that the Markovian

Decision Process has "unobservable costs."

In this paper, we develop a methodology for analyzing this type of a situation. We then refer the reader to Steele [4], where some mathematical results are developed for this problem. A rather complete treatment for the two-state two-action case with unobservable states and unobservable costs may be found in Steele [3].

FINITE-STATE FINITE-ACTION MARKOVIAN DECISION PROCESSES

Consider the Markovian Decision Process (MDP) defined by the following objects:

State space $S = \{1,2,3,\ldots,N\}$, for finite N,

Action space $A = \{a_1,a_2,\ldots,a_M\}$, for finite M,

Cost set $C = \{C(i,a_j) : i \in S,\ a_j \in A\}$, where all costs are taken to be finite,

Transition probabilities = $\{q_{ij}(a_K) : i,j \in S,\ a_K \in A\}$,

Discount factor α, such that $0 \leq \alpha < 1$.

At times, $t = 0,1,2,\ldots$, a decisionmaker observes the current state, $X_t \in S$, of the process. After observing the current state, the decisionmaker then chooses an action $a_t \in A$ and incurs a cost $C(X_t,a_t) \in C$. The next state of the process is then chosen according to the transition probabilities $q_{X_t j}(a_t)$.

A policy for the decisionmaker will be defined as any rule for taking actions at each observation point $t = 0,1,2,\ldots$. A particular policy may be such that at each observation point, t, the action taken, a_t, may depend on the entire observed sequence of states and actions from time $t = 0$ up to and including the current observation X_t. A policy will be called Markovian if at each point $t = 0,1,2,\ldots$, the action taken, a_t, depends on the current state, X_t, of the process but does not depend on the observed sequence of states and actions from time $t = 0$ up to and including time $t-1$. A particular policy may be

randomized in the sense that at each observation time $t = 0,1,2,\ldots$, the action a_t is chosen according to some random procedure. A particular policy, W, will be called deterministic if at each observation point $t = 0,1,2,\ldots$, there exists a map $f_t : S \to A$ such that the policy W chooses the current action a_t according to the rule $a_t = f_t(X_t)$. In other words, a deterministic policy may be defined in terms of a sequence of maps from S into A by

$$W = (f_0, f_1, f_2, \ldots, f_t, \ldots).$$

A particular policy, W, will be called stationary if there exists a single map $f : S \to A$ such that at each observation point $t = 0,1,2,\ldots$, the policy W chooses the current action a_t according to the rule $a_t = f(X_t)$. A stationary policy W therefore may be defined as $W = (f, f, f, f, \ldots)$. In this paper, we will simply consider a stationary policy W and its associated map f as being the same. Therefore, we say that a stationary policy for (MDP) is a map $f : S \to A$.

For any policy W, we define the total expected discounted cost of starting in state i at $t = 0$, and using the policy W over the infinite future, we

$$V_W(i) = E_W \left[\sum_{t=0}^{\infty} \alpha^t C(X_t, a_t) \mid X_0 = i \right],$$

where E_W is used to indicate the dependence of the conditional expectation on the policy W. If W is the stationary policy defined in terms of the map f, then we note that

$$V_f(i) = C(i, f[i]) + \alpha \sum_{j=1}^{N} q_{ij}(f[i]) V_f(j)$$

Other criteria for minimization may be defined for the (MDP). However, in this paper we will only consider the case where the decisionmaker attempts to minimize the total expected discounted cost. Howard [2] analyzed (MDP's) having finite-state and finite-action spaces and proved that an optimal stationary policy (i.e. a stationary policy which minimizes the total expected discounted cost) always exists. The Howard Policy Improvement Routine is a method by which an optimal stationary policy for (MDP) may be found.

LIMITED STATE OBSERVABILITY AND UNOBSERVABLE COSTS

Consider the Markovian Decision Process (MDP) as previously defined, having state space $S = \{1,2,\ldots,N\}$ for finite N. Let h be any map defined on S and having image in the set of positive integers $\{1,2,\ldots,N\}$. That is to say, that $h : S \to h(S) = \{\lambda_1,\lambda_2,\ldots\lambda_u\}$ where $u \leq N$ and $\lambda_j \subset S$ for $j = 1,2,\ldots,u$. In other words, each element of the set h(S) is a subset of S, or h partitions S into subsets.

We now consider the situation where the decisionmaker cannot observe the current state, $X_t \in S$, at each observation point $t=0,1,2,\ldots$; rather the decisionmaker can only observe the subset, $Y_t \in h(S)$, to which the current state X_t belongs. We say that the Markovian Decision Process, and hence the decisionmaker, has limited state observability. The extreme case where $u=1$, i.e. $h(S) = \{\lambda_1\} = \{S\}$, represents the case of unobservable states. The extreme case where $u=N$, i.e. $h(S) = \{\lambda_1,\lambda_2,\ldots,\lambda_N\}$, represents the case of complete state observability which of course is simply the case summarized earlier in this paper. Let $\tilde{S}$ be the set of all probability distributions over S, i.e.

$$\tilde{S} = \left\{\overline{P} = (P_1,P_2,\ldots P_N) \in E_N : 0 \leq P_i \leq 1, \ \sum_{i=1}^{N} P_i = 1, \ i=1,2,\ldots N\right\},$$

where E_N is N-dimensional Euclidean space, and we let P_i be the probability of being in state i. For each j such that $1 \leq j \leq N$, we let $\overline{e}_j = (0,0,\ldots,\underbrace{1}_{j^{th}\text{ place}},0,0,\ldots 0) \in E_N$ be the probability vector having 0's in every place except the j^{th} place. Next, we let $H(j_1,j_2,\ldots j_r)$ for $1 \leq r \leq N$, and $1 \leq j_i \leq N$, be the convex hull of the set

$\{\bar{e}_{j_1}, \bar{e}_{j_2}, \ldots, \bar{e}_{j_r}\}$. We note that

$$H(j_1, j_2, \ldots j_r) \subset \overset{\gamma}{S}$$

$H(j_1, j_2, \ldots j_r)$ = {all probability vector's in E_N having o's in every place except the j_1^{th} place, the j_2^{th} place,..., the j_r^{th} place}

If $\lambda_K \epsilon h(S)$ is defined by $\lambda_K = \{\lambda_{K1}, \lambda_{K2}, \ldots, \lambda_{K\ell_K}\} \subset S$, for $1 \leq \ell_K \leq N$, then we note that $H(\lambda_{K1}, \lambda_{K2}, \ldots, \lambda_{K\ell_K})$, which for convenience we will write as $H(\lambda_K)$, may be considered as being the set of all probability distributions over λ_K. We may now prove the following theorem.

THEOREM: Suppose that at any time $t = 0,1,2,\ldots$, we observe $Y_t = \lambda_i$, for some $\lambda_i \epsilon h(S)$, and we are told that the current distribution over the state space $\overset{\gamma}{S}$ is $\bar{P}_t$. Suppose then that we take the action $a_t \epsilon A$, then both the new observation Y_{t+1} and the new distribution $\bar{P}_{t+1}$ depend only on the current distribution $\bar{P}_t$ and the current action a_t.

Proof: We know that $Y_{t+1} \epsilon h(S)$, i.e. $Y_{t+1} = \lambda_K$ for some K such that $1 \leq K \leq u$. The conditional probability of λ_K, given λ_i, $\bar{P}_t$, and a_t, is given by

$$\Pr\{Y_{t+1} = \lambda_K | Y_t = \lambda_i, \bar{P}_t = \bar{P}, a_t = a\} =$$

$$= \sum_{n \epsilon \lambda_K} \sum_{m \epsilon \lambda_i} q_{mn}(a) P_m, \text{ where we have}$$

written $\bar{P}_t = \bar{P} = (P_1, P_2, \ldots, P_N)$.

This proves that Y_{t+1} depends only on $\bar{P}_t$ and a_t.

Next, we note that since we must observe $Y_{t+i} = \lambda_K$ for some $K = 1,2,\ldots,u$, then it follows that we must have $\overline{P}_{t+1} \in H(\lambda_K)$ for some $K = 1,2,\ldots,u$. If we do observe $Y_{t+1} = \lambda_K$, for some K, then we have

$\overline{P}_{t+1} = \overline{P}' = (P_1', P_2', \ldots, P_N')$, where the components are given by

$$P_j' = \frac{\sum_{\ell \in \lambda_i} q_{\ell j}(a)\, P_\ell}{\sum_{n \in \lambda_K} \sum_{m \in \lambda_i} q_{mn}(a) P_m}, \quad \text{for } j \in \lambda_K$$

and

$$P_j' = 0, \quad \text{for } j \notin \lambda_K$$

We shall write

$$q(\overline{P}', \lambda_K | \overline{P}, \lambda_i, a) = \sum_{n \in \lambda_K} \sum_{m \in \lambda_i} q_{mn}(a)\, P_m, \text{ and}$$

note that

$$q(\overline{P}', \lambda_K | \overline{P}, \lambda_i, a) = \Pr\{\lambda_K | \lambda_i, \overline{P}, a\}.$$

Also, we find that

$$\sum_{\lambda_K \in h(S)} q(\overline{P}', \lambda_K | \overline{P}, \lambda_i, a) = 1.$$

Therefore, we have that, for those λ_K, such that $q(\overline{P}', \lambda_K | \overline{P}, \lambda_i, a) \neq 0$, the $\overline{P}'$ will be given, with probability $q(\overline{P}', \lambda_K | \overline{P}, \lambda_i, a)$, by

$\overline{P}' = (P_1', \ldots, P_N')$ where

$$P_j' = q(\overline{P}', \lambda_K | \overline{P}, \lambda_i, a)^{-1} \sum_{\ell \in \lambda_i} q_{\ell j}(a) P_\ell, \text{ for } j \in \lambda_K$$

and

$$P_j' = 0, \text{ for } j \notin \lambda_K.$$

This, together with the fact that

$$\sum_{\lambda_K \in h(S)} q(\overline{P}', \lambda_K | \overline{P}, \lambda_i, a) = 1,$$

shows that $\overline{P}_{t+1}$ depends only on $\overline{P}_t$ and a_t. Q.E.D.

Next, we write $Q(a) = [q_{ij}(a)]$ for the matrix of transition probabilities associated with the action a. We let $Q(a)_{\lambda_K}$ represent the matrix derived from $Q(a)$ by replacing the columns of $Q(a)$ that are not associated with elements of λ_K, with columns of zeros. We let $\overline{1} \in E_N$ represent the vector having all components equal to one. We may now write

$$q(\overline{P}', \lambda_K | \overline{P}, \lambda_i, a) = \langle \overline{P} Q(a)_{\lambda_K}, \overline{1} \rangle, \text{ as an inner product in } E_N.$$

We also have $\overline{P}' \in H(\lambda_K)$ given by

$$\overline{P}' = \langle \overline{P} Q(a)_{\lambda_K}, \overline{1} \rangle^{-1} \overline{P} Q(a)_{\lambda_K}, \text{ with}$$

probability $\langle \overline{P} Q(a)_{\lambda_K}, \overline{1} \rangle$, when this probability is not equal to zero.

We now assume that the decisionmaker cannot observe the current cost $C(X_t, a_t)$, at any observation point $t=0,1,2,\ldots$. We note that if the current distribution, $\overline{P}_t$, is known, then we may compute the current value of the expected cost $\langle \overline{P}_t, \overline{C}(a_t) \rangle$, where $\overline{C}(a_t) =$

$\left(C(1,a_t),\ C(2,a_t),\ldots,C(N,a_t)\right)$ is the vector of costs associated with the action a_t.

We may now define the new Markovian Decision Process ($\hat{MDP}$) by the following objects.

State space $\hat{S}$ = {all probability distributions over S}, with supplementary information given by observations in h(S).

Action space $\hat{A} = A = \{a_1, a_2, \ldots, a_M\}$

Transition probabilities $\{q(\bar{P}', \lambda_K | \bar{P}, \lambda_i, a) : \bar{P}', \bar{P} \in \hat{S},\ \lambda_K, \lambda_i \in h(S),\ a \in A\}$

Cost set $\{\langle \bar{P}, \bar{C}(a) \rangle : \bar{P} \in \hat{S},\ a \in A\}$

Discount factor α, such that $0 < \alpha < 1$.

We note that ($\hat{MDP}$) has uncountable state-space and finite-action space.

The class of Markovian Decision Processes to which ($\hat{MDP}$) belongs has been analyzed by Blackwell [1]. His analysis shows that an optimal stationary policy (i.e. a map $f : \hat{S} \to A$, which minimizes the total expected discounted cost) for this class of problems always exists, and that the Howard Policy Improvement Routine may be extended to this class of problems. However, in the finite-state finite-action class of problems, the set of stationary policies is finite, and therefore the Howard Policy Improvement Routine will produce an optimal policy in a finite number of steps. In the uncountable-state finite-

action class of problems, the set of stationary policies is uncountable, and therefore the Howard Policy Improvement Routine will not in general produce an optimal policy in a finite number of steps.

A preliminary analysis of ($\tilde{MDP}$) is presented in Steele [4]. The analysis is developed for $h(S) = \{\lambda_1\} = \{S\}$. However, the results apply to the cases in which h is arbitrary.

BIBLIOGRAPHY

[1] Blackwell, D. (1965), *Discounted Dynamic Programming*, Annals of Mathematical Statistics 36, pp. 226-235.

[2] Howard, R.A. (1960), *Dynamic Programming and Markov Processes*, Cambridge, Massachusetts, Technology Press and New York, Wiley.

[3] Steele, J.D., *On Two-state Two-action Markovian Decision Processes With Unobservable States and Unobservable Costs*, Unpublished Ph.D. dissertation, Department of Mathematics, University of California at Los Angeles, 1972.

[4] Steele, J.D., *A Model for the Analysis of Markovian Decision Processes with Unobservable States and Unobservable Costs*, The Rand Corporation, P-4917, October 1972.