Journal of the Operations Research
Society of Japan
Vol.21, No.1, March, 1978

OPTIMAL CONTROL FOR PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES OVER AN INFINITE HORIZON

Katsushige Sawaki
Nanzan University
and
Akira Ichikawa
University of Warwick

(Received August 14, 1976; Revised September 16, 1977)

ABSTRACT

In this paper we consider an optimal control problem for partially observable Markov decision processes with finite states, signals and actions over an infinite horizon. It is shown that there are ϵ-optimal piecewise-linear value functions and piecewise-constant policies which are simple. Simple means that there are only finitely many pieces, each of which is defined on a convex polyhedral set. An algorithm based on the method of successive approximation is developed to compute ϵ-optimal policy and ϵ-optimal cost.

Furthermore, a special class of stationary policies, called finitely transient, will be considered. It will be shown that such policies have attractive properties which enable us to convert a partially observable Markov decision chain into a usual finite state Markov one.

I. Introduction

The partially observable Markov process, introduced by Dynkin [5], consists of two stochastic processes, the core process $\{X_n, n=1, 2, \ldots\}$, which cannot directly be observed, and the signal process $\{S_n, n=1, 2, \ldots\}$ which

becomes known at each decision epoch n = 1, 2, ... The core process is a Markov chain and the signal process is probabilistically related to the core process by the conditional probability $\gamma_{i\theta}$ of observing a signal θ given that the core process is in state i. Dynkin shows that the state occupancy probability represents a sufficient statistic for the complete past history. Åström [1] also considered a similar model with finite states and finite actions over a finite horizon, using the method of successive approximation to find ε-optimal cost vectors, however, it is only applicable to problems in two dimensions. Smallwood and Sondik [8] have independently obtained similar results. Later, Sondik [9] extended this model to the inifinite horizon and introduced the class of finitely tranient policies. White [10] has considered a partially observable semi-Markov process with a finite horizon where the controller knows the times of the core process transition. Sawaragi and Yoshikawa [7] also studied the partially observable control problem with countable states states, uncountable action sets and infinite horizon, where they have explicitly showed that such partially observable models can be transformed into an ordinary complete observable one.

In this paper, under the setting of [8], we shall consider an optimal control problem with discounted cost over an infinite horizon. We introduce three concepts of simple partitions, simple policies, and piecewise linear functions. Using only these concepts we present an algorithm to find an approximation to the optimal cost function. We also show that we can construct an ε-optimal simple stationary policy. We are guaranteed to obtain an ε-approximation of the optimal cost function in finite steps, and each step we only need to find a finite number of vectors by linear programming. Also, an application to a machine maintenance model will be discussed.

Furthermore, in this paper a special class called finite transient, of stationary policies will be considered. We shall show that such policies have very attractive properties and are useful for approximating an optimal policy. If policies are finitely transient, partially observable Markov decision processes can be reduced without loss of generality into finite state Markov decision processes with complete observation.

Sondik [9] has originally introduced the concept of finite transientness of policies for the model with finite sets of states, signals and actions over infinite horizon. However, many parts of his paper are unclear. These will be revised and clarified by giving a different definition of finitely transient policies. The same notation and symbols as in Sondik's paper are adopted here except where confusion occurs.

2. Statement of the Problem

Consider a Markov decision process (called the core process) with state set $\Omega = \{1, 2, \ldots, N\}$, with finite action set A, with probability transition matrices $\{P^a, a \in A\}$, and with immediate cost vectors $\{q^a, a \in A\}$. Let X_n be the state at the n-th transition. Assume that the process $\{X_n, n = 0, 1, 2, \ldots\}$ cannot be observed, but at each transition a signal is transmitted to the decision maker. The set of possible signals $S = \{1, 2, \ldots, \Theta\}$ is assumed to be finite. For each n, given that $X_n = j$ and that action a is to be implemented, the signal θ_n is independent of the history of the signals and actions $\{\theta_0, a_0, \theta_1, a_1, \ldots, \theta_{n-1}, a_{n-1}\}$ prior to the n-th transition and has conditional probability denoted by $\gamma^a_{j\theta} = P[\theta_n = \theta | X_n = j, a]$. At time $n = 0, 1, 2, \ldots$, let $\pi = (\pi_i)$ be the state probability (N-vector). For a transition probability $p^a = (P^a_{ij})$ and an information structure $\Gamma^a = \mathrm{diag}(\gamma^a_{j\theta})$ put $Q^a_\theta = p^a \Gamma^a_\theta$.

If the current state information vector is π, a signal θ is observed and action a has been chosen, then the next state information is given by

$$T(\pi | \theta, a) = \frac{\pi Q^a_\theta}{\{\theta | \pi, a\}} \tag{1}$$

where

$$\{\theta | \pi, a\} = \pi Q^a_\theta \underline{1} \quad \text{with} \quad \underline{1} = (1, \ldots, 1)^T.$$

Let

$$\Pi = \{\pi \in R^N : \Sigma_{i=1}^N \pi_i = 1, \ \pi_i > 0 \ \forall i\}$$

We define Δ as the family of mappings $\delta : T \times \Pi \to A$ where $T = [0, \infty)$. Each element of Δ is called a policy. Given an initial distribution $\pi(0)$ and a policy δ, the subsequent information vectors $\pi(n)$ form a Markov process. Our discounted control problem for an initial distribution $(0) = \pi$ is described by

$$\min_{\delta \in \Delta} E_\theta [\Sigma_{n=0}^{\infty} \beta^n \pi(n) q^{\delta(n, \pi(n))}],$$

where E is the expectation with respect to the signal, β, $0 \leq \beta < 1$, is the discount factor and the cost at time n is given by the inner product πq^a with action a. Let $C(\pi | \delta)$ be a *cost of a stationary policy* δ at an initial value π. Then it is well known (see [2], [3]) that $C(\pi | \delta)$ satisfies

$$C(\pi | \delta) = \pi q^\delta + \beta \Sigma_\theta \{\theta | \pi, \delta\} C(T(\pi | \theta, \delta) | \delta) . \tag{2}$$

Let $C^*(\pi)$ be the *optimal cost*, then the following is true (see [2], [4]).

Theorem 1. There exists an optimal stationary policy δ^* with $C(\pi|\delta^*) = C^*(\pi)$. Also, $C^*(\pi)$ satisfies

$$C^*(\pi) = \min_{a\in A}\{\pi q^a + \beta \sum_{\theta\in S} \{\theta|\pi, a\}C^*(T(\pi|\theta, a))\} \tag{3}$$

for any $\pi \in \Pi$.

An ε-*optimal cost function* C is one satisfying

$$\|C^* - C\| = \sup_{\pi\in\Pi} \|C^*(\pi) - C(\pi|\cdot)\| \leq \varepsilon . \tag{4}$$

A policy δ such that $C = C(\cdot|\delta)$ satisfying (4) is an ε-*optimal policy*.

For finding an ε-*optimal policy and its cost function* we define simple partitions, simple policies and piecewise (abbreviated, hereafter, by p.w.) linear functions.

Definition 1. A partition $\{V_i\}_{i=1}^m$ of Π is called *simple* if each V_i is a convex polyhedral set, where a convex polyhedral set is the solution set of a finite system of linear inequalities, i.e.,

$$V_i = \{\pi\in\Pi : v_{ij}\pi < 0,\ j = 1, 2, \ldots, n_i\}$$

where $v_{ij} \in R^N$ and $v_{ij}\pi$ is the inner product of v_{ij} and π.

Remark 1: Inequalities of the form $v\pi < 0$ contains those of the form $v\pi < \alpha$, α scalar. In fact $v\pi < \alpha$ is equivalent to $(v - \alpha\underline{1})\pi < 0$.

Lemma 1. Let $P_1 = \{V_i\}$ and $P_2 = \{W_j\}$ be two simple partitions of Π. Then, the product partition $P_1 \cdot P_2 = \{V_i \cap W_j\}$ is again simple.

Proof: Here we omit $V_i \cap W_j$ if $V_i \cap W_j = \phi$. The sets $V_i \cap W_i$ are disjoint and are convex polyhedral sets. Hence $P_1 \cdot P_2$ is simple.

Definition 2. A stationary policy δ is called *simple* with respect to a simple partition $\{V_i\}$ if $\delta(\pi) = a_i$ on V_i, $i = 1, 2, \ldots, m$.

Definition 3. A real valued function f on Π is called *piecewise linear* if $f(\pi) = f_i\pi$ on V_i, $i = 1, 2, \ldots, m$, where $\{V_i\}$ is a simple partition and $f_i \in R^N$.

Example: Define an information structure as a mapping from the set of

states (unobservable) of the core process to the set of distinctive signals θ. The decision maker chooses an information structure from the set of available structures and decides upon an action for the system.

Let $a = (a_1, a_2)$ be the pair of actions, a_1 for the system control and a_2 for information acquisition. More precisely, we have

$$P_{ij}^{a}(\theta) = P_{iJ}^{a_1} \gamma_{j\theta}^{a_2}$$

$$\pi q^{a} = \sum_{i=1} \pi_i \sum_{j=1} P_{ij}^{a_1} \sum_{\theta=1}^{\Theta} \gamma_{j\theta}^{a_2} q(i, j, \theta, a_1, a_2)$$

where $q(i, j, \theta, a_1, a_2)$ is the immediate cost of the core process when a state of the core process moves from i to j and a signal θ observed under actions a_1 for the system and a_2 for the information structure, and $\pi = (\pi_1, \ldots, \pi_N)$ is the probability vector with an interpretation π_i is the probability that the core process is in state i.

Consider a machine maintenance and repair model (e.g., Smallwood and Sondik [8]) as an application of partially observable models. But this model is a modification of Smallwood and Sondik's. The machine consists of two internal components. The states of the core process $X_n = i$, $i = 1, 2, 3$, have the following interpretation. If $i = 1$, then both components are broken down, if $= 2$ either one is broken down and if $i = 3$ both of them are working. Assume that the machine produces M finished products at each period and the machine cannot be inspected. The actions a_1 for the machine control are to repair and not to repair the machine. The actions a_2 for information acquisition are the numbers of a sample to choose out of the M finished products. The signals θ are the number of defective products in the sample, which forms the signal process $\{\theta_n, n = 1, 2, \ldots\}$. The core process $\{X_n, n = 1, 2, \ldots\}$ is the unknown states of the components of the machine. Let $\pi_i = P\{X_n = i\}$, $i = 1, 2, 3$ and put $\pi = (\pi_1, \pi_2, \pi_3)$. Then, the process $\{(X_n, \theta_n), n = 1, 2, \ldots\}$ becomes a partially observable machine maintenance and repair model with actions $a = (a_1, a_2)$ and immediate cost πq^{a}.

3. Finitely Transient Policies

In this section a special class of simple stationary policies, called finitely transient, will be studied. The class of such policies has very attractive properties even though all stationary policies do not belong to such a class.

Define, for a simple policy δ,

$$D^k = \bigcup_{\theta} \{\pi : T\{\pi|\theta, \delta\} \in D^{k-1}\}, \quad k = 1, 2, \ldots, \tag{5}$$

where $D^0 = \bigcup_i D_i^0 = \bigcup_{i,j} \{\pi \in \Pi : v_{ij}\pi = 0\}$ which forms the boundary set of the partition $\{V_i\}$ corresponding to a simple policy δ. Let $V^k = \{V_j^k\}_{j=1}^{m'}$ be the collections of sets whose boundaries are $\bigcup_{L=0}^{k} D^L$ and then V^k is a refinement of V^{k-1}, $k \geq 1$, where $V^0 = \{V_j\}$.

Definition 4. A simple policy δ is called *finitely transient* if there is an integer $k < \infty$ such that

$$T(V_j^k|\theta, \delta) \subset V_{\nu(j,\theta)}^k \quad \text{for all } \theta$$

where $T(V|\theta, \delta) = \{T(\pi|\theta, \delta) : \pi \in V\}$ and $\nu(j, \theta)$ is the index of the set containing $T\{\pi|\theta, \delta\}$ for $\pi \in V_j^k$. Let k_δ be the smallest such integer.

Lemma 2. $D^k = \phi$ for all $k \geq k_\delta$ if and only if δ is finitely transient with the index k.

Proof: Suppose that δ is finitely transient with the index k_δ, that is,

$$T(V_j^{k_\delta}|\theta, \delta) \subset V_{\nu(j,\theta)}^{k_\delta} \quad \text{for all } \theta.$$

$D^{k_\delta} = \phi$ because $T(V_j^{k_\delta}|\theta, \delta)$ is the set of all possible state information at the k_δ-th period and V_i^k is open in Π for all, i, k. Let $\angle_\delta$ be the set function defined as $\angle_\delta(B) = \bigcup_{\theta}\{\pi : T(\pi|\theta, \delta) \in B\}$

$$\begin{aligned} D^k &= \bigcup_{\theta}\{\pi : T(\pi|\theta, \delta) \in D^{k-1}\} \\ &= \angle_\delta(D^{k-1}) \\ &= \angle_\delta^k(D^0) \end{aligned}$$

If $D^k = \angle_\delta^k(D^0) = \phi$, then

$$D^{k+1} = \angle_\delta(\phi) = \phi$$

Hence, by induction $D^k = \phi$ for all $k \geq k_\delta$.

Conversely, suppose that $D^k = \phi$ for all $k \geq k$ and that

$$T(V_j^k|\theta,\ \delta) \not\subset V_{\nu(j,\theta)}^k \quad \text{for some } \theta\ .$$

So, there exist $\pi^1, \pi^2 \in V_j^k$ such that for some θ $T(\pi^1|\theta,\ \delta)$ and $T(\pi^2|\theta,\ \delta)$ do not belong to the same set $V_\nu^k(j, \theta)$.

Then, there is a constant λ, $0 < \lambda < 1$, such that $\lambda' T(\pi^1|\theta,\ \delta) + (1-\lambda')$ $T(\pi^2|\theta,\ \delta) \in D^k$ and λ' is given by $\lambda' = \lambda\pi^i Q_\theta^a / \pi Q_\theta^a$. $\lambda' T(\pi^1|\theta,\ \delta) + (1-\lambda')T(\pi^2|\theta,\ \delta)$ $= T(\lambda\pi^1 + (1-\lambda)\pi^2|\theta,\ \delta)$. By letting $\pi = \lambda\pi^1 + (1-\lambda)\pi^2$, $T(\pi|\theta,\ \delta) \in D^k$ which is a contradction.

Lemma 3. Let $Q_{\theta_1}^{a_1} Q_{\theta_2}^{a_2} \cdots Q_{\theta_k}^{a_k} = (Q_\vartheta^a)^k$ and $\underline{0}$ be a zero row vector. A simple policy δ is finitely transient if there exists an integer $k < \infty$ such that

$$V_{ij}(Q_\theta^a)^k > \underline{0} \quad \text{or} \quad < \underline{0} \quad \text{for all} \quad \theta,\ a,\ i,\ j\ .$$

Proof:
$$D^k = \bigcup_\theta \{\pi : T(\pi|\theta,\ \delta) \in D^{k-1}\}$$

$$= \bigcup_\theta \bigcup_{i,\ j} \{\pi : v_{ij}(Q_\theta^a)^k \pi = 0\}$$

Since $\pi_i \geq 0$ and $\Sigma_i \pi_i = 1$, $D^k = \phi$ if $v_{ij}(Q_\theta^a)^k > \underline{0}$ or $\underline{0}$ for all θ, a, i, j. By Lemma 2, this completes the proof.

Remark 1. In Lemmas 2 and 3, the assumption concerning δ being simple is crucial. A counter example is presented as follows: suppoese that there are only two states $N = 2$ and $\pi_2 = 1 - \pi_2 \geq 0$.
Define,

$$\delta(\pi_1) = \begin{cases} a_1 & \text{if } \pi_1 \text{ is rational} \\ a_2 & \text{otherwise} \end{cases}$$

which is stationary but not simple. Then D^δ is the uncountable discontinuous set which never becomes empty. Therefore, a finitely many partition $\{V\}_i$ does not exist.

Theorem 2. Let δ be a simple policy. Then, the following are equivalent.

(i) δ is finitely transient with the index k_δ.
(ii) $C(\pi|\delta)$ is piecewise linear.

Proof of [(i) → (ii)]: Suppose that we have a finitely many partition $V^k = \{V_j^k\}$ for $k \geq k_\delta$. Let $\bar{C}(\pi|\delta) = \pi\alpha_j$, $\pi \in V_j^k$ and $\alpha_j = q^{a_j} + \beta\Sigma Q_\theta^{a_j} \alpha_{\nu(j,\theta)}$

$$\bar{C}(\pi|\delta) = \pi\alpha_j\ ,\ \pi \in V_j^k$$

$$= \pi(q^{a_j} + \beta\Sigma_\theta Q_\theta^{a_j}\alpha_{\nu(j,\,\theta)})$$

$$= \pi q^\delta + \beta\Sigma_\theta\{\pi|\theta,\,\delta\}\frac{\pi Q_\theta^\delta}{\{\pi|\theta,\,\delta\}}\alpha_{\nu(j,\,\theta)} \quad \text{for all } \pi \in V_j^k$$

$$= \pi q^\delta + \beta\Sigma_\theta\{\pi|\theta,\,\delta\}T(\pi|\theta,\,\delta)\alpha_{\nu(j,\,\theta)} \quad \text{for } \delta \text{ finitely transient}$$

$$= \pi q^\delta + \beta\Sigma_\theta\{\pi|\theta,\,\delta\}\bar{C}(T(\pi|\theta,\,\delta)|\delta)$$

$$\equiv (U_\delta\bar{C})(\pi)$$

Since $C(\cdot|\delta)$ is the unique solution of U_δ, $C(\cdot|\delta) = \bar{C}(\cdot|\delta)$.

Proof of [(ii) → (i)] : From piecewise linearity of $C(\cdot|\delta)$, we have $C(\pi|\delta) = \pi\alpha_j$ for $\pi \in V_j^k$ with the partition $\{V_i^k\}$ for $k \geq k$ and $\delta(\pi) = a_j$, $\pi \in V_j^k$.

So $C(T(\pi|\theta,\,\delta)|\delta) = T(\pi|\theta,\,\delta)\alpha_{\nu(j,\,\theta)}$ for $\pi \in V_j^k$.

Then, we must have $T(\pi|\theta,\,\delta) \in V^k_{\nu(j,\,\theta)}$ for all $\pi \in V_j^k$ and all θ. So

$$T\{V_j^k|\theta,\,\delta\} \subset V^k_{\nu(j,\,\theta)} \quad \text{for all } \theta.$$

Corollary: If a policy δ is finitely transient with the simple partition $\{V_j\}$, then its cost $C(\pi|\delta)$ can be computed by solving the following equations;

$$C(\pi|\delta) = \pi\alpha_j \quad \text{for } \pi \in V_j,\ j = 1,\ 2,\ \ldots,\ m \tag{6}$$

and

$$\alpha_j = q^{a_j} + \beta\Sigma_\theta Q_\theta^{a_j}\alpha_{\nu(j,\,\theta)},\quad j = 1,\ 2,\ \ldots,\ m. \tag{7}$$

The proof immediately follows from Theorem 1. Note that the set of equations (7) has a unique bounded solution (see Appendix) and that m need not be equal to the number of actions.

4. Properties of U_a and U_*

This section is a study of the properties of U_a and U_*. Most of these properties will be used later in the development of the algorithm to find ε-optimal approximations to C^* and δ^*.

Let F be the space of real valued functions on Π with sup norm. Then F is a Banach space (B-space). Let Π be equipped with the Euclidean norm, and let C be the subset of continuous functions in F. Then C is a closed linear subspace (hence is itself a B-space) of F. Define operators U_a, U_* on F by

$$(U_a f)(\pi) = \pi q^a + \beta \sum_{\theta \in S} \{\theta | \pi, a\} f(T(\pi | \theta, a)), \quad f \in F,$$

$$(U_* f)(\pi) = \min_{a \in A} \{\pi q^a + \beta \sum_{\theta \in S} \{\theta | \pi, a\} f(T(\pi | \theta, a)\}.$$

Lemma 4. (i) U_a, U_* are contraction mappings with contraction coefficient β.

(ii) U_a, U_* are monotone, i.e., if $f, g \in F$ with $f \leq g$, then $U_* f \leq U_* g$ and $U_a f \leq U_a g$.

(iii) They map C into itself, thus fixed points of these operators are continuous functions.

Proof: The properties (i), (ii) are standard. (See [2], [6]). (iii) U_a clearly maps C into self. $(U_* f)(\pi)$ is the minimum of finite number of continuous, hence it is also continuous, provided f is continuous.

From Lemma 4 we get some information on C* and δ^*.

Lemma 5. The fixed point of U_* exists and is the optimal cost function C*, which is continuous.

Before stating our main results, we need two lemmas.

Lemma 6. Let f be a piecewise linear function w.r.t.$\{V_i\}$ on Π. Define a stationary policy δ_f by $U_* f$, namely, $\delta_f(\pi) = a_i$ if a_i minimizes $(U_a f)(\pi)$. Then δ_f is simple.

Proof: Let $\{V_i\}$ be the simple partition for f. Define

$$V_i(a, \theta) = \{\pi \in \Pi : T(\pi | \theta, a) \in V_i\}.$$

Then for each a, θ, $\{V_i(a, \theta)\}$ is a simple partition. In fact $V_i(a, \theta)$ is given by

$$\frac{\pi Q_\theta^a V_{ij}}{\{\theta | \pi, a\}} < 0, \quad j = 1, 2, \ldots, n_i,$$

or equivalently,

$$\pi Q_\theta^a V_{ij} < 0, \quad j = 1, 2, \ldots, n_i,$$

where v_{ij} characterizes V_i. Let $\{V^a_{i,\theta}\}$ be a simple partition defined by $\bigcap_{i,\theta} V_i(a, \theta)$ (see Lemma 1), then $U_a f$ is linear on each $V^a_{i,\theta}$

More precisely,

$$(U_a f)(\pi) = \pi q^a + \beta \sum_{\theta \in S} \pi Q^a_\theta \sum_i \chi_{V^a_{i,\theta}}(\pi) f_i,$$

where $\chi_{V^a_{i,\theta}}(\pi) = \begin{cases} 1 & \text{if } \pi \in V^a_{i\theta}, \\ 0 & \text{otherwise.} \end{cases}$ and f_i is a vector defining f.

Since δ is defined by minimizing finite number of piecewise linear functions, it is simple.

Lemma 7. (i) If f is piecewise linear, then $U_* f$ is piecewise linear.

(ii) If f is concave, then $U_* f$ is also concave.

Proof: $U_a f$ has the same property as f's. By the definition of $U_* f$, the desired results are obtained.

Theorem 3. Let $f_0 \in F$, and define

$$f_n(\pi) = (U_* f_{n-1})(\pi).$$

Let δ_n be the decision rule at stage n defined by $U_* f_{n-1}$.

(i) f_n converges to C^*.

(ii) If f_0 is piecewise linear, then so is f_n for any n. Furthermore, δ_n is simple.

(iii) If f_0 is concave, then f_n is concave.

(iv) If $f_1 \leq f_0$, then $f_n \downarrow C^*$. If $f_1 \geq f_0$, then $f_n \uparrow C^*$.

Proof: The assertions follow from Lemmas 4, 5, 6 and 7.

Remark 2. If we take $f_0(\pi) = C(\pi|\delta)$ for some stationary policy δ, then $f_n \downarrow C^*$. In particular, if we take $\delta(\pi) = a$ for all π, thus $C(\pi|\delta) = f_0(\pi) = \pi(I - \beta P^a)^{-1} q^a$, then f_n is continuous concave and piecewise linear and $f_n \downarrow C^*$. Hence C^* is continuous and concave.

Remark 3. Let $f_0(\pi) = \min_{a \in A} \pi q^a$, then f_0 is piecewise linear, concave and continuous. Hence (ii) and (iii) hold. Since f_n corresponds to the optimal cost for the n-period problem with discounting, this case is essentially equivalent to the results in [9]. If we further assume $q^a \geq 0$ for any $a \in A$, then $f_n \uparrow C^*$.

Next we shall discuss the rate of convergence.

Lemma 8. Let $f \in F$. If $||f - U_* f|| \leq (1 - \beta)\varepsilon$, then $||C^* - f|| \leq \varepsilon$.

Proof: $||C^* - f|| \leq ||U_*C^* - U_*f|| + ||U_*f - f||$

$$\leq \beta||C^* - f|| + ||U_*f - f||.$$

After arranging, the result is obtained.

Theorem 4. If $\beta^n||f_0 - U_*f_0|| \leq (1-\beta)\varepsilon$, then $||C^* - f_n|| \leq \varepsilon$.

Proof: Since we have

$$||f_n - U_*fn|| = ||U_*f_{n-1} - U_*^2f_{n-1}||$$

$$\leq \beta||f_{n-1} - U_*f_{n-1}||$$

$$\vdots$$

$$\leq \beta^n||f_0 - U_*f_0||,$$

the theorem follows directly from Lemma 8.

Remark 4. If we calculate $||f_0 - U_*f_0||$, then Theorem 4 tells us when to stop. Furthermore, at each step n we know from $||f_n - U_*f_n||$ how many steps (at most) we have to go after the step n.

5. Algorithm

Since Π is uncountable, it is far from trivial to calculate $C(\pi|\delta)$ which may not be a piecewise linear funciton of π, except the case that δ is finitely transient. In this section we shall approximate $C(\pi|\delta)$ by using the method of successive approximation.

The method of successive approximation is a well known and popular method for solving equations. The method is to start with a cost function f_0, and to iterate U_*, constructing a sequence of cost functions $f_n = U_*f_{n-1}$, $n = 1, 2, \ldots$. By Lemma 4, U_* is a contraction mapping with fixed point C^* and by Theorem 3, $\{f_n\}$ converges to C^*. By Theorem 4, n can be chosen sufficiently large, so that f_n is an ε-optimal cost function. In fact by taking logarithms of the expression in Theorem 4,

$$n > \log\left[\frac{(1-\beta)\varepsilon}{||f_0 - f_1||}\right]/\log\beta$$

is adequate.

The next theorem provides a means of constructing an ε-optimal policy from an ε'-optimal cost function and specifies the relationship between ε and ε'. The algorithm will first construct an ε'-optimal cost function. From this cost function, an ε-optimal policy is constructed.

Let f_0 be piecewise linear, and let δ_n be defined by $U_* f_{n-1}$, i.e., $\delta_n(\pi) = a_1$ if a_1 minimizes $(U_a f_{n-1})(\pi)$. Then δ_n is simple, and satisfies $U_* f_{n-1} = U_{\delta_n} f_{n-1}$, where U_δ for a stationary policy δ is defined by

$$(U_\delta f)(\pi) = \pi q^{\delta(\pi)} + \beta \sum_{\theta \in S} \{\theta | \pi, \delta(\pi)\} f(T(\pi | \theta, \delta(\pi))) .$$

Theorem 5. If $||C^* - f_{n-1}|| \leq \frac{1-\beta}{2\beta}\varepsilon$, then $||C^* - C(\cdot|\delta_n)|| \leq \varepsilon$.

Proof: It is easy to show that U_δ for any stationary policy δ is a contraction mapping and that the fixed point is $C(\cdot|\delta)$, i.e., $C(\pi|\delta) = U_\delta C(\cdot|\delta)(\pi)$. Consider

$$\begin{aligned} ||C^* - C(\cdot|\delta_n)|| &= ||U_{\delta_n} C(\cdot|\delta_n) - U_* C^*|| \\ &\leq ||U_{\delta_n} C(\cdot|\delta_n) - U_{\delta_n} C^*|| + ||U_{\delta_n} C^* - U_{\delta_n} f_{n-1}|| \\ &\quad + ||U_* f_{n-1} - U_* C^*|| \\ &\leq \beta ||C(\cdot|\delta_n) - C^*|| + \beta ||C^* - f_{n-1}|| + \beta ||f_{n-1} - C^*||. \end{aligned}$$

Here we used the equality $U_* f_{n-1} = U_{\delta_n} f_{n-1}$. Rearranging the above inequality we obtain

$$(1-\beta)||C(\cdot|\delta_n) - C^*|| \leq 2\beta ||C^* - f_{n-1}|| \leq (1-\beta)\varepsilon .$$

Hence $||C(\cdot|\delta_n) - C^*|| \leq \varepsilon$.

If the state space is uncountable, or even countably infinite, then this procedure is difficult to implement on a computer. However, since the partially observable Markov decision process has the structure of piecewise linearity and f_0 is p.w. linear, then each f_n is p.w. linear and each δ_n constructed as in the previous theorem is simple (by Lemma 6). In this case, the cost functions and policies can be specified by a finite number of items - the inequalities describing each cell of a simple partition and the corresponding action or linear function.

Algorithm to Find an ε-optimal Simple Policy:

(i) Start with any p.w. linear function f_0

(ii) Compute $f_1 = U_* f_0$.

(iii) Choose an integer n such that

$$\beta^n ||f_0 - f_1|| \leq (1 - \beta)\varepsilon',$$

where $\varepsilon' = (1-\beta)\varepsilon/2\beta$. I.e., choose $\hat{n}$ larger than

$$\log \left[\frac{(1 - \beta)^2 \varepsilon}{2\beta ||f_0 - f_1||}\right] / \log \beta.$$

(iv) Compute $f_n = U_* f_{n-1}$ successively until $n = \hat{n}$.

(v) Consequently, we obtain $f_{\hat{n}}$ such that

$$||C^* - f_{\hat{n}}|| \leq \varepsilon'.$$

(vi) Construct a policy δ satisfying

$$U_\delta f_{\hat{n}} = U_* f_{\hat{n}}.$$

Then δ is ε-optimal.

Remark 5. The algorithm can be started with $f_0 \equiv 0$.

Remark 6. The termination criterion, $n = \hat{n}$, in the algorithm has the advantage that $||f_0 - f_1||$ is computed only once. However, it has the disadvantage that $\hat{n}$ will probably be larger than necessary, causing unnecessary iterations.

An alternative would be to compute $||f_n - f_{n-1}||$ at each iteration and stop whenever $||f_n - f_{n-1}|| \leq (1-\beta)\varepsilon'/\beta$. Theorem 2 guarantees that f_n is an ε'-optimal cost function. However, the computations of $||f_n - f_{n-1}||$ will, in general, be expensive.

The best procedure is undoubtedly to check $||f_n - f_{n-1}||$ at some, but not all, iterations. For example, $\hat{n}$ might be computed based on $||f_0 - f_1||$. Then at some iteration n near $\frac{\hat{n}}{2}$, recompute n based on $||f_n - f_{n-1}||$.

Acknowledgment

The authors wish to thank Professors S. Brumelle and U. Haussmann, University of British Columbia, for their useful comments and discussions. Also worthy of acknowlwdgment is the very helpful review of one of the referees. We wish to thank him or her for the many comments. Also, special thanks go to Professor Y. Iihara, Nanzan University, for his encouragement and helpful suggestions.

References

[1] K.J. Åström, Optimal Control of Markov Process with Incomplete State Information, *J. Math. Anal. App.*, 10, 174 - 205 (1965).

[2] D. Blackwell, Discounted Dynamic Programming, *Ann. Math. Sta.*, 36, 226 - 235 (1965).

[3] E.V. Denardo, Contraction Mapping in the Theory Underlying Dynamic Progrmming, *SIAM Review*, 9, 165 - 177 (1967).

[4] M.H. Davis and P. Varaiya, Dynamic Programming Conditions for Partially Observable Stochastic System, *SIAM J. Control* 11, 226 - 261 (1973).

[5] E.B. Dynkin, Controlled Random Sequences, *Theory of Probability and Its Applications*, X, 1 - 14 (1965).

[6] S.M. Ross, *Applied Probability Models with Optimization Applications*, Holden Day, 1970.

[7] Y. Sawaragi and T. Yoshikawa, Discrete-Time Markovian Decision Process with Incomplete State Observation, *Ann. of Math. Stat.*, 41, 78 - 86 (1970).

[8] R.D. Smallwood and E.J. Sondik, Optimal Control of Partially Observable Processes over the Finite Horizon, *Opns. Research* 21, 1071 - 1088 (1973)

[9] E.J. Sondik, The Optimal Control of Partially Observable Markov Processes over the Infinite Horizon: Discounted Costs, Department of Engineering-Economic Systems, Stanford University, California, May 1975.

[10] C.C White, Procedures for the Solution of a Finite Horizon, Partially Observed, Semi-Markov Optimization Problem, *Oper. Research*, 24, 348 - 358 (1976)

Katsushige SAWAKI
Faculty of Business Administration
Nanzan University
18, Yamazato-cho
Showa-ku, Nagoya, 466
Japan

Appendix

Lemma: The set of linear equations given by

$$\alpha_j = q^{a_j} + \beta \sum_{\theta} Q_{\theta}^{a_j} \alpha_{\nu(j,\,\theta)}, \quad j = 1, 2, \ldots, m,$$

has a unique bounded solution.

Proof: Let $a_j = j$ for each j.
We may set $Q_{\theta}^{j}\alpha_{\nu(j,\,\theta)} = \tilde{Q}_{i}^{j}\alpha_i$ if $\nu(j, \theta) = i$, $i = 1, 2, \ldots, m$.
Then we have

$$\alpha_j = q^j + \beta\tilde{Q}_1^j\alpha_1 + \ldots + \beta\tilde{Q}_m^j\alpha_m.$$

Let

$$\alpha = \begin{pmatrix} \alpha_1 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_m \end{pmatrix}, \qquad q = \begin{pmatrix} q^1 \\ \cdot \\ \cdot \\ \cdot \\ q^m \end{pmatrix} \qquad \text{and} \qquad \tilde{Q} = \begin{pmatrix} \tilde{Q}_1^1 & \tilde{Q}_2^1 & \cdots\cdots & \tilde{Q}_m^1 \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \tilde{Q}_1^m & \tilde{Q}_2^m & \cdots\cdots & \tilde{Q}_m^m \end{pmatrix}.$$

Hence, we obtain

$$\alpha = q + \beta\tilde{Q}\alpha, \quad \text{that is,} \quad \alpha = \left(I - \beta\tilde{Q}\right)^{-1} q$$

where since $||\beta\tilde{Q}|| < 1$ for $0 \leq \beta < 1$ with the sup norm, there exists $\left(I - \beta\tilde{Q}\right)^{-1}$.