Journal of the Korean
OR/MS Society
Vol. 16, No.1, June., 1991

Partially Observable Markov Decision Process with Lagged Information over Infinite Horizon

정병호 · 김성희

Byung Ho Jeong* and Soung Hie Kim**

Abstract

This paper shows the infinite horizon model of Partially Observable Markov Decision Process with lagged information. The lagged information is uncertain delayed observation of the process under control. Even though the optimal policy of the model exists, finding the optimal policy is very time consuming. Thus, the aim of this study is to find an $\epsilon-$optimal stationary policy minimizing the expected discounted total cost of the model.

$\epsilon-$optimal policy is found by using a modified version of the well known policy iteration algorithm. The modification focuses to the value determination routine of the algorithm. Some properties of the approximation functions for the expected discounted cost of a stationary policy are presented. The expected discounted cost of a stationary policy is approximated based on these properties. A numerical example is also shown.

1. INTRODUCTION

Partially Observable Markov Decision Process(POMDP) is a generalization of Markov De-

* Chonbuk National University

** Korea Advanced Institute of Science and Technology

cision process(MDP) that permits state uncertainty of system, and that allows acquisition of partial information for a current state[1, 6, 8, 10]. The quality and quantity of data received from observations play a vital role in the control of the system. In the context, a finite horizon POMDP with lagged information is introduced by Kim and Jeong[4]. The paper states that the system performance is improved by using the lagged information. A proof of this statement is provided by White.[11]

This paper considers an infinite horizon model of POMDP with the lagged information. The characteristics of the problem is very similar to those of a general POMDP without the lagged information. Hence, the algorithms that have been developed for optimal control of a general POMDP over infinite horizon can be applied to the model by modifying slightly in order to deal with the change of the system dynamics. The commonly proposed methods for solving the infinite horizon POMDP are policy iteration algorithm[9] and successive approximation method[12]. The former converges faster than the later but value determination step in policy iteration algorithm has severe approximation method to the value determination step of the policy iteration algorthm. However, since the approximated expected cost of a stationary policy is piecewise linear but, in general, not concave, the algorithm presented cost of a stationary policy is piecewise linear but, in general, not concave, the algorithm presented in the finite horizon model with lagged information[4] can not be applied to policy improvement step. Thus, this study substitutes the concave function sufficiently close to the expected cost of a stationary policy for the expected cost of the stationary policy. For this work, some important properties of a stationary policy are reviewed and the algorithm is developed based on these properties.

2. MODEL DESCRIPTION

Suppose that the internal dynamics of the controlled system can be modelled by a stochastic process. Let s(t) be a random variable defined on a sample space, $\Omega=\{1, 2, \cdots, N\}$, where discrete time $t\epsilon I$, $I=1, 2, \cdots$ The stochastic process $\{s(t), t\epsilon I\}$, called core process or underlying process, is assumed to be a finite state Markov chain with stationary NxN transition probability matrix $P^k=(P^k_{ij})$, $i, j\epsilon\Omega$, $k\epsilon K$. Suppose that the decision maker can control the core

process by choosing action from a finite set, K, of available actions. Then, the core process is completely described by P^k and the initial state vector.

$\pi(0)=\{\pi_1(0), \cdots, \pi_N(0)\}$, where $\pi_i(0)=\Pr\{s(0)=i\}$

This paper assumes that true state cannot be observed completely, and that the current and previous states can be observed partially through finite state probabilistic observers of two types. The observed states of the current and previous state are denoted by θ and δ, respectively. The probabilistic characteristics of these current and lagged partial observations are represented by two stochastic matrices $B=[b_{j\theta}]$ and $L=[1_{i\delta}]$, respectively. Kim and Jeong[4] gives the information structure of this model and provides the rule of calculating the state vector $\pi(t)$ based on the previous state vector $\pi(t-1)$ and current and lagged observations, θ and δ.

$$\pi_j(t)=\frac{\sum_i \pi_i(t-1)1_{i\delta}P^k_{ij}b_{j\theta}}{\sum_{ij}\pi_i(t-1)1_{i\delta}p^k_{ij}b_{j\theta}}$$

This transformation function can be written in terms of vectors and matrices as follows:

$$\{\theta, \delta \mid \pi, k\}=\pi L_\delta P^k B_\theta \underline{1}$$

and

$$T(\pi \mid k, \theta, \delta)=\frac{\pi L_\delta P^k B_\theta}{\{\theta, \delta \mid \pi, k\}} \qquad (1)$$

where $\underline{1}=[1, 1, \cdots, 1]^T$, $B_\theta=\mathrm{diag}[b_{j\theta}]$, and $L_\delta=\mathrm{diag}[1_{i\delta}]$.

The superscript T denotes transpose of a vector. $\{\theta, \delta \mid \pi, k\}$ represents the probability that the current and lagged observed states after one transition are θ and δ, respectively, when state vector is π and action k is chosen. These functions define the dynamics of the state of knowledge of the core process. Suppose that the sole source of information from the process is the sequence of two observations, and that the state vector $\pi(t)$ is computed after two observations are revealed. The state vector is a sufficient statistics of the state of knowledge.[4]

Denote the immediate cost operating the process at the process at the state i under action k by q^k_i. When the state vector is π and action k is chosen, the expected immediate cost $\pi_q{}^k=\sum_i \pi_i q^k_i$ is incurred to operate the process. Let β be a discount factor, $0\leq\beta<1$. Denote $\psi_t(\pi)$ as

a control function that represents the action to be chosen when the state vector at time t is π (t). Then, the expected discounted cost control problem for fixed $\pi(0)$ can be written as,

$$\underset{\psi 0,\ \psi 1,\ldots}{\text{MIN}}\ E[\Sigma_{t=0}^{\infty}\beta^{t}\pi(t)q^{\psi}t^{(\pi)}] \tag{2}$$

where the sequence of control functions $(\psi_1, \psi_2, \cdots)$ must be selected to minimize the expression. Such a sequence of control functions is defined as a policy. If a policy consists of a singel control function to be used at each time period, then the policy is termed stationary. A stationary policy is denoted by $\psi^{\infty}=(\psi,$
$\psi, \cdots)$. Let $C(\pi \mid \psi)$ be the expected discounted cost of a stationary policy, ψ^*, at an initial state vector π. Then, it is well known that $C(\pi \mid \psi)$ is the unique bounded solution to the following equation[2,9].

$$C(\pi \mid \psi)=\pi q^{\psi}+\beta\Sigma_{\theta,\ \delta}\{\theta, \delta \mid \pi, \psi\}C\{T(\pi \mid \theta, \delta, \psi) \mid \psi\} \tag{3}$$

The aim of this paper is to find a stationary policy minimizing $C(\pi \mid \psi)$ in Eq.(3). That is, the optimal cost function $C^*(\pi)$ has the following equation.

$$C^*(\pi)=\inf_{\psi} C(\pi \mid \psi)$$

Then, the problem (2) is within the scope of Blackwell's formulation[2]. It folows that the minimum expected discounted cost as a function of the initial state vector π, $C^*(\pi)$, exists and that it satisfies

$$C^*(\pi)=\text{MIN}_k[\pi q^k+\beta\sum_{\theta,\ \delta}\{\theta, \delta \mid \pi, k\}C^*(T(\pi \mid \theta, \delta, k))] \tag{4}$$

Furthermore, there exists a stationary policy that achieves this minimum cost. The stationary policy is denoted by $(\psi^*)^{\infty}$, where $\psi^*(\pi)$ is the minimizing alternative in (3). Thus, Eq. (3) can be written as

$$C^*(\pi)=\pi q^{\psi^*}+\beta\Sigma_{\theta,\ \delta}\{\theta, \delta \mid \pi, \psi^*\}C^*(T(\pi \mid \theta, \delta, \psi^*)) \tag{5}$$

where the specific dependence of $\psi^*(\pi)$ on π has been suppressed.

However, since it is far from trivial to find the solution of (3), a method to approximate expected discounted cost for a stationary policy is needed. The following section presents some

properties of $C(\pi \mid \psi)$ in (3) and an approximation method of $C(\pi \mid \psi)$ based on these properties. In the latter part of this paper, for notational simplicity, an approximation of $C(\pi \mid \psi))$ is denoted by $f^n(\pi)$, where n is the iteration number and ψ is suppressed.

3. THE APPROXIMATION OF $C(\pi \mid \psi)$

Properties to be investigated in this section will be used in the development of the algorithm to find $\epsilon-$optimal plicy and $\epsilon-$optimal cost function. Define the following operators$U_k f$, U_ψ f and $U_* f$ on any real

valued bounded function f.

$$(U_k f)(\pi)=\pi q^k+\beta\sum_{\theta,\delta}\{\theta,\delta \mid \pi, k\}f(T(\pi \mid \theta, \delta, k)) \tag{6}$$

$$(U_\psi f)(\pi)=\pi q^k+\beta\sum_{\theta,\delta}\{\theta,\delta \mid \pi, \psi\}f(T(\pi \mid \theta, \delta, \psi)) \tag{7}$$

$$(U_* f)(\pi)=\mathrm{MIN}_k[\pi q^k+\beta\sum_{\theta,\delta}\{\theta,\delta \mid \pi,k\}f(T(\pi \mid \theta,\delta,k))] \tag{8}$$

Then, since the above operators satisfy the monotone contraction assumption, they are monotone contraction mapping[2,3]. To verify the piecewise linearity of $U_k f$ and $U_* f$, an important characteristics of the transformation function (1) is shown in the following Lemma.

Lemma 1. $T(\pi \mid \theta,\delta,k)$ preserves straight lines. That is, if $0\le\lambda<1$, $\bar{\lambda}=1-\lambda$, and π^1, $\pi^2\epsilon\prod$, then $\lambda\pi^1+\bar{\lambda}\pi^2$ is a line segment in $\prod$ with each point π^1, π^2. If the transformation of the line for fixed θ, δ, and k is considered, then

$$T(\lambda\pi^1+\bar{\lambda}\pi^2 \mid \theta,\delta,k)=\mu T(\pi^1 \mid \theta,\delta,k)+\bar{\mu}T(\pi^2 \mid \theta,\delta,k)$$

where as λ ranges between 0 and 1, μ ranges between 0 and 1. Hence the image of the line segment under $T(\cdot \mid \theta, \delta, k)$ is a line seqment. Specifically,

$$\mu=\frac{\lambda\{\theta,\delta \mid \pi^1, k\}}{\lambda\{\theta,\delta \mid \pi^1,k\}+\bar{\lambda}\{\theta,\delta \mid \pi^2,k\}},\quad \bar{\mu}=1-\mu$$

Proof) From Eq.(1),

$$T(\pi \mid \theta,\delta,k)=\frac{\pi L_\delta P^k B_\theta}{\{\theta,\delta \mid \pi,k\}}=\frac{\pi L_\delta P^k B_\theta}{\pi L_\delta P^k B_\theta \underline{1}}$$

Let $Q=L_\delta P^k B_\theta$

$$T(\lambda\pi^1+\bar{\lambda}\pi^2 \mid \theta,\delta,k)=\frac{(\lambda\pi^1+\bar{\lambda}\pi^2)Q}{(\lambda\pi^1+\bar{\lambda}\pi^2)Q\underline{1}}$$

$$= \frac{\lambda\pi^1 Q}{(\lambda\pi^1+\bar{\lambda}\pi^2)Q\underline{1}} + \frac{\bar{\lambda}\pi^2 Q}{(\lambda\pi^1+\bar{\lambda}\pi^2)Q\underline{1}}$$

$$= \frac{\lambda\{\theta,\delta \mid \pi^1,k\}}{(\lambda\pi^1+\bar{\lambda}\pi^2)Q\underline{1}} T(\pi^1 \mid \theta,\delta,k) + \frac{\bar{\lambda}\{\theta,\delta \mid \pi^2,k\}}{(\lambda\pi^1+\bar{\lambda}\pi^2)Q\underline{1}} T(\pi^2 \mid \theta,\delta,k)$$

Let $\mu = \dfrac{\lambda\{\theta,\delta \mid \pi^1,k\}}{\lambda\{\theta,\delta \mid \pi^1,k\}+\bar{\lambda}\{\theta,\delta \mid \pi^2,k\}}$ and $\bar{\mu}=1-\mu$

Then, $T(\lambda\pi^1+\bar{\lambda}\pi^2 \mid \theta,\delta,k) = \mu T(\pi^1 \mid \theta,\delta,k) + \bar{\mu} T(\pi^2 \mid \theta,\delta,k)$

Theorem 1. Suppose that $f(\pi)$ is a piecewise linear concave function on $\prod$. Then, the operator $(U_k f)(\pi)$, is also a piecewise linear concave function on π for each k.

Proof) Since the sum of piecewise linear concave functions is also a piecewise linear concave function, and πq^k is obviously piecewise linear concave function for each k, it is sufficient to show that $\{\theta,\delta \mid \pi,k\} f(T(\pi \mid \theta,\delta,k))$, for arbitrary but fixed θ, and δ, is piecewise linear concave. Let $g(\pi) = \{\theta,\delta \mid \pi,k\} f(T(\pi \mid \theta,\delta,k))$ and $\pi' = \lambda\pi^1 + \bar{\lambda}\,\pi^2$ for any λ, defined in Lemma 1. Then, using the assumed concavity of $f(\pi)$,

$$\begin{aligned} g(\pi') &= \{\theta,\delta \mid \pi',k\} f(T(\pi' \mid \theta,\delta,k)) \\ &\geq \{\theta,\delta \mid \pi',k\}[\mu f(T(\pi^1 \mid \theta,\delta,k)) + \bar{\mu} f(T(\pi^2 \mid \theta,\delta,k))] \\ &= [\lambda\{\theta,\delta \mid \pi^1,k\} + \bar{\lambda}\{\theta,\delta \mid \pi^2,k\}][\mu f(T(\pi^1 \mid \theta,\delta,k) + \bar{\mu} f(T(\pi^2 \mid \theta,\delta,k))] \\ &= \lambda\{\theta,\delta \mid \pi^1,k\} f(T(\pi^1 \mid \theta,\delta,k) + \bar{\lambda}\{\theta,\delta \mid \pi^2,k\} f(T(\pi^2 \mid \theta,\delta,k)) \\ &= \lambda g(\pi^1) + \bar{\lambda} g(\pi^2) \end{aligned}$$

That is, $g(\pi)$ is a concave function of π for each combination of θ and δ. Since $(U_k f)(\pi)$ is sum of concave functions, it is also concave. Next, the piecewise linearity is shown. Since $f(\pi)$ is piecewise linear over $\prod$, there exists α−vectors which are piecewise constant on $\prod$. Thus, Eq.(6) can be written as follows.

$$\begin{aligned} (U_k f)(\pi) &= \pi q^k + \beta \textstyle\sum_{\theta,\delta} \pi L_\delta P^k B_\theta \alpha(T(\pi \mid \theta,\delta,k)) \\ &= \pi[q^k + \beta \textstyle\sum_{\theta,\delta} \pi L_\delta P^k B_\theta \alpha(T(\pi \mid \theta,\delta,k))] \\ &= \pi\gamma(\pi) \end{aligned}$$

where $\gamma(\pi)$ is piecewise constant over $\prod$. Thus, $(U_k f)(\pi)$ is a piecewise linear function. Therefore, $(U_k f)(\pi)$ is a piecewise linear concave function on $\prod$ for each k.

A collection $P=\{E_1, E_2, \cdots, E_m\}$ of subsets of $\prod$ is a partition of $\prod$ if $E_1, \cdots, E_m$ are mutually exclusive and collectively exhaustive subsets of $\prod$. if each cell of a partition is a convex

set, then the partition is called simple. Also, a stationary policy ψ^∞ is called simple with respect to a simple partition P if $\psi(\pi)=k_l$ for $\pi\epsilon E_l$, $=1, 2, \cdots, m$. Using that $(U_kf)(\pi)$ is a piecewise linear and concave function on $\prod$, some important properties of the operatios $U^\psi f$ and $U_* f$ can be verified in the following theorem.

Theorem 2. Suppose that $f(\pi)$ is piecewise linear. Then,

i) $(U^\psi f)(\pi)$ is piecewise linear whenever ψ is a simple policy.

ii) $(U_* f)(\pi)$ is piecewise linear concave and there exists a simple policy ψ such that

$U^\psi f = U_* f$.

Proof) i) Suppose that ψ is a simple policy with respect to a simple partition $\{E_l\}$. Let E_l be an arbitrary but fixed cell from the partition and suppose that $\psi(\pi)=k$ for $\pi\epsilon E_l$. Then,

$(U^\psi f)(\pi)=(U_k f)(\pi)$ for $\pi\epsilon E_l$

From Theorem 1, $U_k f$ is piecewise linear for each $k\epsilon K$. Hence $U^\psi f$ is piecewise linear on each cell E_l, and is consequently piecewise linear on $\prod$.

ii) By the theorem 1, $(U_k f)(\pi)$ is piecewise linear concave on $\prod$. By definition, $(U_* f)(\pi)$ is minimization of $U_k f$ over k. Therefore, since the minimization of piecewise linear concave functions is also piecewise linear and concave function, $(U_* f)(\pi)$ is piecewise linear concave on $\prod$. The fact that $(U_* f)(\pi)$ is piecewise linear implies that a simple policy can be characterized by a simple partition of $\prod$ and action for each cell.

Theorem 2 plays the fundamental role of the algorithm to be presented in the next section. That is, the part i) of theorem 2 implies that the expected discounted cost of a stationary policy ψ can be approximately obtained by applying the operator $U^\psi f$ repeatedly, and that the expected cost is piecewise linear on $\prod$. Furthermore, the part ii), in fact, indicates the policy improvement step can be performed by using the algorithm for finite horizon POMDP with lagged information[4]. Since the part i), however, does not guarantee the concavity of the expected discounted cost of a stationary policy ψ, the algorithm can not be directly applied to policy improvement. Thus, the concave hull, difined as $\bar{f}(\pi)=\min_l[\pi\alpha_l]$ by

Sondik[9], of a piecewise linear function, $f(\pi)=\pi\alpha_l$ for $\pi\epsilon E_l$, will be used in the policy improvement step.

Corollary. Suppose that $f^0(\pi)$ is piecewise linear on $\prod$ and $f^0(\pi)$ is the concave hull of $f^0(\pi)$. Then,

i) $f^n(\pi)=(U_\psi f^{n-1})(\pi)$, $n=1, 2, \cdots$, is piecewise linear for any simple policy ψ.

ii) $f^n(\pi)=(U_* \bar{f}^{n-1})(\pi)$, $n=1, 2, \cdots$, is piecewise linear concave function on $\prod$ and there exists a simple poplicy, ψ_n, satisfying $(U_{\psi_n} \bar{f}^{n-1})(\pi)=(U_* \bar{f}^{n-1})(\pi)$

The above corollary can be easily proved by using theorem 2 repeatedly. Furthermore, since $U_\psi f$ and $U_* f$ are monotone contraction mappings, $f^n(\pi)$ is monotone decreasing sequence. That is, by using the operator $(U_\psi f)$ iteratively, f^n converges in norm to the fixed point C^*, i.e., $U_\psi C^*=C^*$. In the next section, an algorithm suing $U_\psi f$ and $U^* \bar{f}$ is suggested.

4. ALGORITHM

In the previous section, it is shown that $U_\psi f$ is piecewise linear on $\prod$ for any simple policy whenever $f(\pi)$ is piecewise linear function on $\prod$. Thus, the expected discounted cost of any stationary policy, ψ, can be approximated by iterative use of the operator $U_\psi f$,(i.e., successive approximation method). In the case of implementation of the operator $U_\psi f$, the cost function 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. The simple partition and the corresponding piecewise linear function are updated by the transformation function. Jeong [5] describes the procedure for implementing the operator $U_\psi f$ in detail.

After predetermined iterative use of $U_\psi f$, and approximation of the expected cost of a stationary policy ψ is obtained as a piecewise linear function. The concave hull of the function is used in policy improveement. The use of the concave hull of $f(\pi)$ $\bar{f}$, leads to improved policies if the approximation is sufficiently close to $f(\pi)$. It can be easily seen that $f'(\pi) \leq \bar{f}(\pi) \leq f(\pi)$ where $f'(\pi)$ represents the expected cost of the improved policy in policy improvement step[6]. In the reminder of this section, the algorithm is presented in general terms and the proof concerning convergence is given.

An algorithm for finding an ϵ—optimal policy starts with a simple policy ψ_1 satisfying $f^1 \geq U_{\psi_1} f^1$. An iteration of the algorithm is described as follows:

At the start of the n^{th} iteration, we have a simple policy ψ_n and a piecewise linear function $\bar{f}^n$ satisfying

$\bar{f}^n \geq U_{\psi_n}\bar{f}^n$, $n=0, 1, 2, \cdots\cdots$

i) Compute $U^h_{\psi_n}\bar{f}^n$, where the integer h is the number of repetition of U_{ψ_n} which are to be performed.

ii) Set $f^{n+1}=U^h_{\psi_n}\bar{f}^n$ and compute the concave hull $\bar{f}^{n+1}$ of f^{n+1}.

iii) Find an improved policy ψ_{n+1} such that $U_{\psi_{n+1}}\bar{f}^{n+1}=U_*\bar{f}^{n+1}$

iv) If $\| U_*\bar{f}^{n+1}\| \leq(1-\beta)\epsilon$, then stop with $\epsilon-$optimal policy, ψ_n. Otherwise, increase n by 1 and go to step i).

In the algorithm, $\| \cdot \|$ is the supremum norm. If h=1 in step i), the algorithm becomes the successive approximation algorithm. That is, the algorithm is a modification of the policy iteration algorithm that applies the successive approximation method to the value determination step. As an example of the initial simple policy and piecewise linear function f satisfying $f \geq U_\psi f$, $\psi(\pi)=k$ for all $\pi\in\prod$, and $f^1(\pi)=M/(1-\beta)$ for each $\pi\in\prod$, is to be considered, where $M=\max_{i,k}Q^k_i$

The following theorem shows that if the algorithm terminates then it will provide an $\epsilon-$optimal cost function and an $\epsilon-$optimal policy.

Theorem 3. If $\| U_*\bar{f}^n-\bar{f}^n\| \leq(1-\beta)\epsilon$, then

$\| \bar{f}^n-C^*\| \leq\epsilon$, i.e., $\bar{f}^n$ is $\epsilon-$optimal.

Proof) Note that $U_*C^*=C^*$ and U_* is a contraction mapping.

$$\| \bar{f}^n-C^*\| \leq \| \bar{f}^n-U_*\bar{f}^n\| + \| U_*\bar{f}^n-U_*C^*\|$$

$$\leq \| \bar{f}^n-U_*\bar{f}^n\| +\beta\| \bar{f}^n-C^*\|$$

$$(1-\beta)\| \bar{f}^n-C^*\| \leq \| \bar{f}^n-U_*\bar{f}^n\| \leq(1-\beta)\epsilon$$

Therefore, $\| \bar{f}^n-C^*\| \leq\epsilon$

That is, since the operators, $U_\psi f$ and $U_* f$, are monotone contraction mappings, the algorithm terminates with $\epsilon-$optimal stationary policy and $\epsilon-$optimal cost function in the finite number of iterations.

5. EXAMPLE

In this section, a problem with two states and two alternatives is solved by using the developed algorithm. Each alternative changes the sate transition matrix P and the expected immediate cost vector. The parameters of this example are given in Table 1.

Insert Table 1

Table 1. Parameters for the example

action	p^k		q^k			
1	0.8	0.2	−3.0	B	0.6	0.4
	0.0	1.0	5.0		0.5	0.5
2	0.95	0.05	4.5	L	0.9	0.1
	0.85	0.15	1.0		0.1	0.9

The discount factor is given by $\beta=0.8$. The operator U^{ψ} is repeated 10 times in each iteration, i.e., $h=10$. Let the initial stationary policy be $\psi_1(\pi)=1$ for all $\pi\epsilon\prod$. The initial cost function is given by $f^1(\pi)=M/(1-\beta)=5/0.2=25$.

In the first iteration, an approximation of the expected discounted cost of ψ_1 is $\bar{f}^2(\pi)=\pi\alpha_1(\pi)$ for all $\pi\epsilon\prod$ where $\alpha_1(-10.71,25.0)$. The improved policy found in the first iteration is $\psi_2(\pi)=2$ if $0\leq\pi_1\leq0.88$, $\psi_2(\pi)=1$ if $0.88<\pi_1\leq1$. Since $\| U_*\bar{f}^1-\bar{f}^1 \| =2.35\geq(1-\beta)\epsilon=0.04$, the algorithm goes to the second iteration with $\psi_2(\pi)$ and $\bar{f}^2(\pi)$. The improved policy in the second iteration is $\psi_3(\pi)=2$ if $\pi_1\leq0.61$ and $\psi_3(\pi)=1$ if $\pi_1>0.61$. And as $\| U_*\bar{f}^3-\bar{f}^3 \| =1.74$, the given error bound is not satisfied.

The same procedures are repeated until the given error bound is satisfied and the details are omitted. Since the iteration 7 satisfies with the error bound, the policy improved in iteration 6 is an $\epsilon-$optimal stationary policy. The $\epsilon-$optimal policy is $\psi^*(\pi)=2$ if $\pi_1\leq0.38$, otherwise $\psi^*(\pi)=1$. The expected discounted cost function of the $\epsilon-$optimal policy consists of the following three $\alpha-$vectors; $\alpha_1=(5.78,-2.97)$, $\alpha_2=(-13.23,3.71)$, i.e., $C^*(\pi)=\pi\alpha_1(\pi)$ if $\pi\leq0.26$, otherwise $C^*(\pi)=\pi\alpha_2(\pi)$.

Insert Figure 1

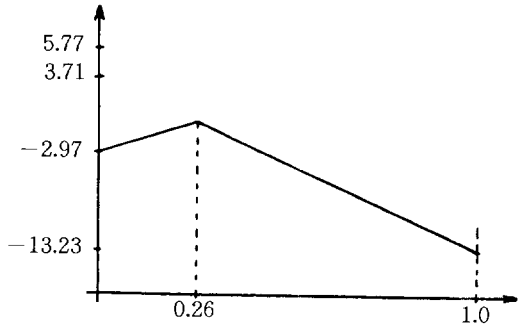

Figure 1. Cost function of the ε-optimal stationary policy for the example.

To use this control, the controller must update π_1 after each time period using the observed states of two types. When the action is chosen using $\psi^*(\pi)$, the expected discounted cost is as shown if Figure 1, i.e., $C^*(\pi)=\min_i \pi a_i$

6. CONCLUSION

This paper has extended the finite horizon POMDP model with lagged information to the infinite horizon case with discount factor. This infinite model deals with discount factors which lie in the range $0\leq\beta<1$ since total expected discounted cost of all poplicy can be infinite if $\beta=1$.

The paper presents some useful properties of a stationary policy and modifies the well-known policy iteration algorithm by applying the successive approximation method to the value determination step. That is, this study applies the modified version of the policy iteration algorithm to the infinite horizon POMDP model with lagged information.

References

1. Albright, S.C., "Structural Results for Partially Observable Markov Decision Processes,"

Oper. Res., Vol. 27, 1041−1053, 1979

2. Blackwell, D., "Discounted Dynamic Programming," Ann. Math. Stat., Vol.36, 226−235, 1965.
3. Denardo, E.V., "Contraction Mappings in the Theory Underlying Dynamic Programming," SIAM Rev., Vol.9, 165−177, 1967.
4. Kim, S.H. and Jeong, B.H., "A Partially Observable Markov Decision Process with Lagged Information," J. Opl. Res. Soc., Vol.38, 439−446, 1987.
5. Jeong, B.H., Use of Lagged Information in partially Observable Markov Decision Process, Unpublished Dissertation, Korea Advanced Institute of Science and Technology, 1989.
6. Platzman, L.K., "Optimal Infinite−Horizon Undiscounted Control of Finite Probabilistic Systems," SIAM J. Con. & Opt., Vol.18, 362−380, 1980.
7. Sawaki, K., Piecewise Linear Markov Decision Process with an aplication into Patially Observable Models, in "Recent Developments in Markov Decision Prcoesses,"(R. Hartley et al, Eds.), Acdemic Press, New York, 1980.
8. _____, "Transformation of Partially Observable Markov Decision Processes into Piecewise Linear Ones," J. Math. Anal. & Appl., Vol.91, 112−118, 1983.
9. Sondik, E.J., "The Optimal Control of Partially Observable Markov Process over the Infinite Horizon; Discounted Costs," Oper. Res., Vol.26, 348−358, 1978.
10. Tijms, H.C. and F.A. van der Duyn Schouten, "A Markov Decision Algorithm for Optimal Inspections and Revisions in a Maintenance System with Partial Information," Euro. J. Oper. Res., Vol.21, 245−253, 1984.
11. White, C.C., "Note on 'A Partially Observable Markov Decision Process with Lagged Information'," J. Opl. Res. Soc., Vol.39, 217−220, 1988.
12. White, D.J., "Finite State Approximations for Denumerable State Infinite Horizon Contracted Markov Decision Processes: The Policy Space Method," J. Math. Anal. & Appl., Vol.72, 512−523, 1979.