Journal of Informatics and Mathematical Sciences
Volume 1 (2009), Number 1, pp. 33–42
© RGN Publications

An Educational Management Problem with Continuous Signal Space

John E. Goulionis and V. K. Benos

Abstract. *Partially Observable Markov Decision Process* (POMDPs) have been suggested as a suitable model to formalizing the planning of educational management. In this paper, we discuss a specialization of POMDPs that is tailored to a frequently re-occurring type of educational problem, with five states (bad, moderate, good, very good, excellent), two teaching methods a traditional based to National program and a new education method based to the British system. We extend the model of POMDPs with finite discrete signal space to a more natural model where the signal space is continuous instead of finite. We consider the significant and realistic problem with probability density functions for the signals to be uniformly distributed. We prove the piecewise affinity of the infinite horizon optimal utility function associated with this problem. To solve this problem we use a procedure that take advantage of special problem structure, and we provide optimal policies to stochastic and dynamic decisions naturally arise in finding the optimal educational method.

1. Introduction

Educational models are meant to contribute to all aspects of student's development. A model is validated if its effectiveness in contributing to student's development has been scientifically confirmed. The introduction and continuation of British Education System is a burning issue in Greece. The merits and demerits of this system are being discussed continuously in public forums. A uniform education system is demanded to be implemented in Greece. In such circumstances there was a need of a study to compare the British education system and National program with respect to curriculum, teaching methodology and evaluation mechanism of these programs.

As our educational systems increases in size and complexity, and as they become increasingly dependent upon the devices and techniques of the new educational technology, a systematic quantitative approach to the design and operation of these

2000 *Mathematics Subject Classification.*90-04; 90C39; 90C15; 65K05.
*Key words and phrases.*OR in education; Markov processes; Bayesian statistics.

educational systems is a vital necessity. For this reason we begin the process for the simple yet very important system composed of a class of students.

Partially Observable Markov Decision Process POMDP is a general sequential decision-making model, where the effects of actions are nondeterministic and only partial information about world states is available. Recently POMDPs have been suggested as providing a suitable, integrated approach to educational management problems Sondik (1971), Goulionis (2005). In these models it is assumed that uncertainty exists in the transitions of the system itself and in our knowledge of which state the system truly occupies. Therefore, the objective is to find an optimal policy based on the observations of the system and the previous decision rules applied.

Value iteration is a popular algorithm for finding solutions to POMDPs, Bertsekas (1997). It can find an optimal value function for a finite horizon POMDP and can also find an arbitrarily good estimate of the optimal value function for an infinite-horizon POMDP. This method is inefficient for two reasons. First, a (DP) update is expensive due to the need of accounting for all belief states in a continuous belief space. Second, this method needs to contact a large number of (DP) updates before its convergence Papadimitriou *et al.* (1987).

The piecewise affinity of the optimal function for our model could have also been obtained using the value iteration method of Littman (1996). We use the piecewise affinity of the optimal cost function $V_{\beta}(\pi)$ to obtain analytic expression to compute the optimal reward/cost and policy, (depending explicitly and only on the basic data of the problem). In particular, we illustrated the fact that the rich mathematical structure of the Markov models sometimes enables explicit results (e.g., closed form formulas), that give considerable insight about the behavior of the physical systems being modeled.

In section 2, the model is described in detail some assumptions are provided, and we discuss a specialization of POMDPs associates with an educational system problem. We obtain a set of appropriate model parameters and investigate the structural properties of the case with five-states.

The rest of the paper is organized as follows.

In section 3 we extend the usual model of POMDP. We consider that signal distributions have density functions whereas in the section 2 we assume that they are discrete. This is a realistic problem for educational models, because examinations are the integral part of Education Systems. Effectiveness and authenticity of the education system cannot be ascertained without tests. By the tests we have the signals and the partial information about the state of the class, but the results of the tests usually give continuous distribution. For this problem we compute the optimal policy in the two action educational model. We develop a solution procedure utilizing the properties of the utility function. In section 4 we give a numerical example.

2. A POMDP teaching model

In the modelling of physical systems the concept of the state of the physical system has proved to be a very valuable tool for the characterization of system performance Sondik (1978), Goulionis *et al.* (2007). This idea may also be an important aid to the description of the learning process.

Thus in the class of the students the internal state of a class is measured of the internal state of the students that constitute the class. We shall use the internal state of a student as a representation of his learning characteristics. The internal state of a student depends on different factors, as hereditary roots, familial and social environments, personal model of thinking, preexisted knowledge, sentimental reasons and generally psychological factors. Therefore, the internal state of a class depends on many factors and for this reason is unknown. However we can take a sense of this internal state by some observations, for example (score in a test, participation in the learning process, the language of a body etc) see Goulionis (2005).

A *Partially Observable Markov Decision Process* (POMDP), is a collection $(S, A, P, \Theta, R, c, \beta)$. The POMDP consists of a core process x_t, an observation process z_t, and a decision process α_t.

The core process $\{x_t,\ t = 0, 1, 2, \ldots\}$ is a discrete-time Markov process.

(1) The deterioration levels of the class are classified into a finite number of states. The state numbers are ordered to reflect the degree of the deterioration. The states (Weak, average, good, very good, excellent) are coded with the numbers 5, 4, 3, 2, 1 respectively; $S = \{1, 2, 3, 4, 5\}$ the set of states.

(2) At any given time period, the teacher selects one of the actions of the set A. He has a finite number of actions (educational-methods) available in order to control the situation of the class. The first method is Cheap. The second method is luxurious. The actions at each time are coded with 1,2 respectively; $A \equiv \{1, 2\}$. The teacher is unable to observe the state of the class directly and must make his/her decisions sequentially based upon partial information time.

(3) State transitions occur according to a Markov Chain whose transition probabilities are determined by the choice of the material to be presented to the class. To accomplish the effect of the teaching method upon the internal state knowledge of a class by transitions from one state to another state, we have a transition probability matrix $P^{\alpha} = (p_{ij}^{a})$. The state process evolves according to the transition probabilities define by

$$p_{ij}(a) = P\{x_{t+1} = j | x_t = i, a_t = a\},$$

where $i, j = 1, 2, \ldots, 5$, $a \in A$.

(4) An observer does not directly observe the core process. He sees instead one of the outputs which is probabilistically related to z_t. At each time period, the state of the class is monitored incompletely by some monitoring mechanism.

The outcome of the monitoring is classified into finite levels $\Theta = \{1, 2, \dots, 5\}$. The signals for our model represent outcomes of the tests. For simplicity we consider five types of observations are coded in $1, 2, \dots, 5$. We consider that the observation of type 1 ($\theta = 1$), is favorable for the state 1 (the best state of a class), while the observation of type 5 ($\theta = 5$), is favorable for the state 5, (the worst state of a class). For example we consider that we have an observation of type 1, 2, 3, 4, 5 if we have success in the test over 90%, 70-90%, 50-70%, 30-50% and less than 30% respectively. Thus the sets of signals is $\Theta = \{1, 2, 3, 4, 5\}$.

(5) The observation process is related to the state and the control processes by means of the conditional probabilities $r_{x_t, y_{t+1}}(u_t)$ defined by

$$r_{i\theta}(a) = P\{y_{t+1} = \theta | x_t = i, a_t = a\}, \quad i = 1, 2, \dots, 5.$$

(6) The cost structure considered here is as follows:
$C^a(i)$, where $c(i, a)$ is the scalar valued cost accrued, when the current state is $i \in S$ and action is $\alpha \in A$.

(7) $\beta \in (0, 1)$ is a discount factor.

Although the state of the core process is not known with certainty, it is possible to calculate the probability that the core process is in a given state. In particular we define:

$$\pi_i(t) = \Pr\{x_t = i | z_0, \dots, Z_t, \alpha_0, \dots, z_{t-1}\}.$$

The vector $\pi(t) = (\pi_1(t), \pi_2(t), \dots, \pi_N(t))$ is called information vector, and the space of all such vectors, Π, is called the information space. We have: $\sum_{i=1}^{N} \pi_i(t) = 1$ and $\pi_i \geq 0$. It is well known that $\pi(t)$ is a sufficient statistic Bertsekas (1997). More precisely, $\pi(t)$ summarizes all of the necessary information of the history of the process for choosing an action at time t.

If the information vector at time t is π and an alternative α is selected, and if an output θ results, then the new information $\pi(t+1)$ is given by $T(\pi|\theta, \alpha)$. By Bayes' rule.

$$T(\pi(t)|\theta, \alpha) = \pi(t+1) = \frac{\pi \cdot P^\alpha \cdot R^\alpha_\theta}{\{\theta|\pi, \alpha\}}. \tag{2.1}$$

$\{\theta|\pi, \alpha\} = \pi \cdot P^\alpha \cdot R^\alpha_\theta \cdot \underline{1}$ is the probability of receiving observation θ at stage $t+1$, given that $\pi(t)$ and α is the action selected at stage t. Assuming $\underline{1} = \text{col}\{1, \dots, 1\}$.

$$T_i(\pi(t)/\theta, \alpha) = \frac{r^\alpha_{i\theta} \cdot \sum_{j=1}^{j=N} \pi_j(t) \cdot p^\alpha_{ji}}{\sum_{j=1}^{j=N} r^\alpha_{i\theta} \sum_{j=1}^{N} \pi_j(t) \cdot p^\alpha_{ji}}. \tag{2.2}$$

R^α_θ be the diagonal matrix having $r^\alpha_{i\theta}$ as its j-th diagonal term and zeros for all off-diagonal terms.

The objective of a POMDP is to find an optimal policy among the admissible policies such that it minimizes a given performance index, typically the total expected discounted cost to be accrued over the infinite horizon, conditioned on the a priori $\pi(0)$. These costs are defined in terms of the state x_t for each admissible strategy, δ, and information vector $\pi(0)$ of the initial state by:

Discounted-cost (DC)

In terms of the information vector $\pi(t)$ we have that:

$$J_\beta(\delta,\pi(0)) : \lim_{n\to\infty} E^{\delta}_{\pi(0)}\left[\sum_{t=0}^{n}\beta^t \pi(t)c^{a(t)}\right]. \tag{2.3}$$

we define:

$$V_\beta(\pi) \equiv \inf_{\bar{\delta}} J_\beta(\bar{\delta},\pi). \tag{2.4}$$

Then, $V_\beta(\pi)$ is the total expected discounted cost accrued when an optimal policy is selected, given that the initial information vector is π, and future costs are discounted at rate β. It is well known Goulionis (2007) that $V_\beta(\pi)$ is the unique solution of the functional equation:

$$V_\beta(\pi) = \max_{\alpha}\left\{\pi\cdot c^\alpha + \beta\cdot\sum_{\theta}\{\theta/\pi,\alpha\}\cdot V_\beta(T(\pi/\theta,\alpha))\right\} \tag{2.5}$$

In this point some notation and operators useful for later sections are introduced.

Notations and operators

A policy is a function which maps the state space into the action space; i.e., if δ is a policy, then $\delta : \Pi \to A$ where $\delta(\pi)$ is the action taken in state $\pi \in \Pi$. Let the policy space Δ be the set of all stationary policies. Let B be the set of all bounded real valued functions on Π. In this paper, the norm $\|\cdot\|$ is the supreme norm; for example, if $\upsilon \in B$, then $\|\upsilon\| = \sup\{\|\upsilon(\pi)| : \pi \in \Pi\}$.

When computing optimal policies in the infinite horizon case, we need only consider stationary policies Sondik (1978). A stationary policy is denoted by $(\delta)^\infty = (\delta,\delta,\ldots)$.

For convenience define the local income function h which assigns a real number to each triple (π,α,υ) with $\pi \in \Pi$, $a \in A$, and $\upsilon \in B(\Pi)$.

$$h : \Pi \times A \times B(\Pi) \to \mathbb{R}$$

$$h(\pi,\alpha,\upsilon) := \pi\cdot c^\alpha + \beta\sum_{\theta}\{\theta|\pi,a\}\cdot\upsilon(T(\pi,a,\theta)) \tag{2.6}$$

$$[H_\delta(\upsilon)](\pi) = h(\pi,\delta(\pi),\upsilon). \tag{2.7}$$

Finally an optimal operator is defined as:

$$H\upsilon := \max_{\delta\in\Delta}[H_\delta\upsilon]. \tag{2.8}$$

The operators have the following properties: *Boundness*; *monotonicity* and the *contraction property*, Lovejoy (1991).

This contraction property can be stated thus: For some fixed β, $0 \leq \beta < 1$, then $||H_\delta \upsilon - H_\delta u|| \leq \beta \cdot ||\upsilon - u||$ for all u, $\upsilon \in B(\Pi)$ and $\delta \in \Delta$.

The metric ρ on $B(\Pi)$ is defined by $\rho(v, \upsilon) = \sup\{|v(\pi) - \upsilon(\pi)| : \pi \in \Pi\}$. H_δ, and H are isotone contractions on $B(\Pi)$ with unique fixed points, say V_δ and V^*, respectively Lovejoy (1991) and if we start with any $V_0 \in B(\Pi)$ and recursively define:

$$V_{\delta,t} = H_\delta V_{\delta,t-1}$$

and

$$V_t = HV_{t-1}$$

then as t goes to infinity V_t and $V_{\delta,t}$ will converge in metric ρ on $B(\Pi)$ to V^* and V_δ, respectively, and $V^* = \sup\{V_\delta : \text{all}\, \delta\}$, $V^* = V^\delta$ if and only if

$$V^* = HV^* = H_\delta V^* . \tag{2.9}$$

Therefore the optimal value function or its approximation can be computed using *dynamic programming techniques*. The sequence of estimates converges to the unique fixed-point solution which the direct consequence of Banach's theorem for contraction mappings, Monahan (1982).

Sondik (1978) show that for any finite t, the optimal value function V_t^* is piecewise linear and convex, i.e., $V_t^*(\pi) = \max\{\pi \cdot \gamma : \gamma \in \Gamma_t\}$ for some finite set Γ_t of vectors in R^n. Using the representation for V_t^* in the dynamic programming recursion $V_t^* = HV_{t-1}^*$, they show that this latter mapping can be represented as:

$$V_t^* = HV_{t-1}^*(\pi) = \max\left\{\pi \cdot \left[q^\alpha + \beta \cdot P^a \sum_\theta R_\theta^\alpha \cdot \gamma^{l(\pi,\alpha,\theta)}\right] : \alpha \in A\right\}, \tag{2.10}$$

where $l(\pi, \alpha, \theta)$ is the index of the $\gamma \in \Gamma_t$ that maximizes $\pi \cdot P^\alpha \cdot R_\theta^\alpha \cdot \gamma$. Thus, given Γ_t and any $\pi \in \Pi(s)$, one would find $l(\pi, \alpha, \theta)$ for each $a \in A$ and $\theta \in \Theta$, and then find the optimal action and $V_t^*(\pi)$ from (2.10). The inner bracketed term in (2.10) evaluated at the optimal action is a gradient vector for V_t^* at π; call this $\gamma(\pi)$.

The simplest way to define the control function $\delta^* : \Pi \rightarrow A$ from the ith approximation of the value function V_t is via greedy one-step lookahead:

$$\delta^*(\pi) : arg \max_a\{\pi.q^\alpha + \beta. \sum_{\theta=1}^{M}\{\theta/\pi, \alpha\} \cdot V_i(T(\pi, \theta, \alpha))\}. \tag{2.11}$$

The accuracy of the approximate solution (t-th value function) with regard to V^* can be expressed in terms of the Bellman error ϵ, Littman (1996).

3. Educational models with uniformly distributed signal processes

In this section we suppose that the basic assumptions are the same. The only difference is the nature of the signal processes. We assume that the signal distributions have density functions whereas in the section 2 we assume that they are discrete. This is a realistic problem for educational models. The parameters of these probability density functions depend on the system state as well as the action taken in the previous decision epoch. More precisely, for each educational system state $i \in S$, each action $a \in A$, for each time $t = 0, 1, 2, \ldots$, there is a probability density function $f^a_{i,t}$ on the set of signals. For the infinite horizon problem, we assume that the probability density function is time invariant and the dependency on t is suppressed from the notation. Then the conditional density function on the set of the signal can then be computed as

$$\bar{f}(\theta|\pi, \alpha) = \sum_{K=1}^{N} \sum_{j=1}^{N} \pi_j \cdot P^{\alpha}_{j,\kappa} \cdot f^{\alpha}_{\kappa}(\theta)$$

or, in matrix form

$$\bar{f}(\theta|\pi, \alpha) = \pi \cdot P^{\alpha} \bar{R}^{\alpha}_{\theta} \cdot 1, \tag{3.1}$$

where $\bar{R}^{\alpha}_{\theta}$ is a diagonal matrix with $f^{\alpha}_{i}(\theta)$ as its diagonal elements and 1 is a N-dimensional column vector with all elements being 1. Analogous to the definition of $T(\pi, \alpha, \theta)$, define $\bar{T}(\pi, \alpha, \theta)$ as the probability distribution of the system state at the next time epoch, given that the probability distribution of the current system state is π, the action applied is a, and the signal obtained in the next time epoch is θ; i.e,

$$\begin{aligned} \bar{T}_i(\pi, \alpha, \theta) &= \Pr(x_t + 1 = i|\pi, Y_t + 1 = \alpha, Z_t + 1 = \theta) \\ &= \frac{\sum_{j=1}^{N} \pi_j \cdot P^{\alpha}_{j,i} \cdot f^{\alpha}_{i}(\theta)}{\sum_{\kappa=1}^{N} \sum_{j=1}^{N} \pi_j \cdot P^{\alpha}_{j,\kappa} \cdot f^{\alpha}_{\kappa}(\theta)} \end{aligned} \tag{3.2}$$

or, in matrix form,

$$\bar{T}(\pi, \alpha, \theta) = \frac{\pi \cdot P^{\alpha} \cdot \bar{R}^{\alpha}_{\theta}}{\bar{f}(\theta|\pi, \alpha)} = \frac{\pi \cdot P^{\alpha} \cdot \bar{R}^{\alpha}_{\theta}}{\pi \cdot P^{\alpha} \cdot \bar{R}^{\alpha}_{\theta} \cdot 1}. \tag{3.3}$$

Hence

$$\begin{aligned} v_t(\pi) &= \max_{\alpha \in A} E\{q^{\alpha}(x_t) + \beta \cdot v_{t+1}(\bar{T}(\pi, \alpha, \theta))|\pi, \alpha\} && (3.4) \\ &= \max_{\alpha \in A} \left\{ \sum_{i=1}^{N} \pi_i \cdot q^{\alpha}(i) + \beta \cdot \int_{\theta \in \Theta} \tilde{f}(\theta|\pi, \alpha) \cdot v_{t+1}(\bar{T}(\pi, \alpha, \theta)) \cdot d\theta \right\}, && (3.5) \\ v_t(\pi) &= \max_{\alpha \in A} \left\{ \sum_{i=1}^{N} \pi_i \cdot q^{\alpha}(i) + \beta \cdot \int_{\theta \in \Theta} \pi \cdot P^{\alpha} \cdot \bar{R}^{\alpha}_{\theta} \cdot \gamma^{l(\pi, \alpha, \theta)} \cdot d\theta \right\}. && (3.6) \end{aligned}$$

A uniform distribution is commonly used to model a process without much available information. The algorithm developed for the discrete signal space can then be applied to solve this type of problem and are more efficient than the method which discussed in the previous section. Let Θ be the signal space. Also, at decision epoch t, let $\Theta^{a}_{t,i}$ be the signal space for the process given that the state of a class is i and that the decision taken at previous decision epoch is a. The probability density function for the signal in this signal space is uniformly distributed. A trivial case occurs if the state can be deduced for sure from the observed signal; i.e.,

$\Theta(t,i,\alpha)\cap\Theta(t,j,\alpha)=\varnothing$ for all pairs of states i and j. This can clearly be formulated as a completely observable MDP. Of course the above technique fails if the supports of the signal distribution overlap. However, if the signal distributions are uniform, then the problem can be reformulated as a POMDP with finite signal space. $\Theta'(t,i,\alpha)=\Theta-\Theta(t,i,\alpha)$ and $\bar{\Theta}(t,i,\alpha)=\{\theta(t,i,\alpha),\Theta'(t,i,\alpha)\}$. Then $\bar{\Theta}(t,i,\alpha)$ is a partition of the signal space Θ. Let $\Theta_t=\{\widehat{\Theta}_{t,1},\widehat{\Theta}_{t,2},\ldots,\widehat{\Theta}_{t,K})$ be the product partition of $\bar{\Theta}(t,i,\alpha)$ for all system states i and actions a. Since there are only a finite number, 5, of system states, there are only a finite number of elements in Θ_t. The key to converting a uniformly distributed signal problems to a discrete signal problems is that the only information provided by the signal is the cell of the partition in which it occurs. Each element in Θ_t can be viewed as a signal in a finite signal space problem. We can extend this simple case. By following the same procedures as discussed above it can be shown that educational models formulated with POMDPs whose signal processes are step functions can also be formulated as POMDPs with finite signals.

Lemma 3.1. *For every $i\in S$, $f_i^{\alpha}(\cdot)$ is constant on every element of Θ_t.*

Proof. Let θ_1, θ_2 be any two arbitrary signals in any $\widehat{\Theta}_{i,j}\in\Theta_t$. By the method discussed above for generating the elements in Θ_t, either $\widehat{\Theta}_{t,j}\bigcap\Theta(t,i,\alpha)=\varnothing$ or $\widehat{\Theta}_{t,j}\subseteq\Theta(t,i,\alpha)$ for all system states i. If $\widehat{\Theta}_{t,j}\cap\Theta(t,i,\alpha)=\varnothing$, then $f_i^{\alpha}(\theta_1)=f_i^{\alpha}(\theta_2)=0$. If $\widehat{\Theta}_{t,j}\subseteq\Theta(t,i,\alpha)$, then by uniform assumption, $f_i^{\alpha}(\theta_1)=f_i^{\alpha}(\theta_2)$. □

Theorem 3.2. *$T(\pi,a,\cdot)$ is constant on every element of Θ_t.*

Proof. It is obvious using Lemma 3.1. □

We give now the following definition.

Definition 3.3. $\Theta_{\pi,\alpha,\bar{\gamma}}=\{\theta\in\Theta:\pi\cdot P^{\alpha}\cdot\bar{R}_{\theta}^{\alpha}\cdot\bar{\gamma}\geq\pi\cdot P^{\alpha}\cdot\bar{R}_{\theta}^{\alpha}\cdot\gamma,\ \forall\ \gamma\in\Gamma\}$. Then

$$
\begin{aligned}
Hv(\pi) &= \max_{\alpha\in A}\left\{\sum_{i=1}^{N}\pi_i\cdot q^{\alpha}(i)+\beta\sum_{\Theta_{\pi,\alpha,\bar{\gamma}}}\pi P^{\alpha}\cdot\int_{\theta\in\Theta_{\pi,\alpha,\bar{\gamma}}}\bar{R}_{\theta}^{\alpha}\cdot\bar{\gamma}\cdot d\theta\right\}\\
&= \max_{\alpha\in A}\left\{\sum_{i=1}^{N}\pi_i\cdot q^{\alpha}(i)+\beta\cdot\sum_{\Theta_{\pi,\alpha,\bar{\gamma}}}\pi\cdot P^{\alpha}\cdot\xi^{\alpha}(\pi,\bar{\gamma})\right\}\\
&= \max_{\alpha\in A}\left\{\pi\cdot\left[q^{\alpha}+\beta\cdot\sum_{\Theta_{\pi,\alpha,\bar{\gamma}}}P^{\alpha}\cdot\xi^{\alpha}(\pi,\bar{\gamma})\right]\right\},\\
\xi^{\alpha}(\pi,\bar{\gamma}) &= \begin{pmatrix}\int_{\theta\in\Theta_{\pi,\alpha,\bar{\gamma}}}f_1^{\alpha}(\theta)\cdot\bar{\gamma}_1\cdot d\theta\\ \int_{\theta\in\Theta_{\pi,\alpha,\bar{\gamma}}}f_2^{\alpha}(\theta)\cdot\bar{\gamma}_2\cdot d\theta\\ \vdots\\ \int_{\theta\in\Theta_{\pi,\alpha,\bar{\gamma}}}f_N^{\alpha}(\theta)\cdot\bar{\gamma}_N\cdot d\theta\end{pmatrix},
\end{aligned}
$$

where $\bar{T}(\pi,\alpha,\theta)\cdot\gamma^{l(\pi,\alpha,\theta)}\geq\bar{T}(\pi,\alpha,\theta)\cdot\gamma^{k}$ for all γ^{k} supporting v_{t+1}.

4. **Numerical example**

Let us illustrate the method for computing with the following example. We have a class The states (bad, moderate, good, very good, excellent) are coded with the numbers 5, 4, 3, 2, 1 respectively; $S = \{1, 2, 3, 4, 5\}$ the set of states. We have two different teaching methods.

$$p^1 = \begin{bmatrix} 0.7 & 0.2 & 0.1 & 0 & 0 \\ 0.4 & 0.4 & 0.1 & 0.1 & 0 \\ 0.3 & 0.3 & 0.4 & 0.1 & 0 \\ 0 & 0 & 0.1 & 0.1 & 0.8 \end{bmatrix}, \quad q^1 = \begin{bmatrix} 10 \\ 7 \\ 5 \\ -6 \\ -7 \\ -8 \end{bmatrix},$$

$$p^2 = \begin{bmatrix} 0.8 & 0.1 & 0.1 & 0 & 0 \\ 0.7 & 0.3 & 0 & 0 & 0 \\ 0.1 & 0.5 & 0.4 & 0 & 0 \\ 0 & 0.3 & 0.7 & 0 & 0 \\ 0 & 0.1 & 0.1 & 0.8 & 0.1 \end{bmatrix}, \quad q^2 = \begin{bmatrix} 12 \\ 9 \\ 6 \\ -7 \\ -8 \\ -9 \end{bmatrix}.$$

The signal processes have the following density functions for $\theta > 0$. (other 0)

$$\begin{array}{ll} f_1^1 = e^{-\theta}, & f_1^2 = 20.e^{-\theta}, \\ f_2^1 = 8.e^{-\theta}, & f_2^2 = 25.e^{-\theta}, \\ f_3^1 = 8.e^{-\theta}, & f_3^2 = 26.e^{-\theta}, \\ f_4^1 = 9.e^{-\theta}, & f_4^2 = 28.e^{-\theta}, \\ f_5^1 = 10.e^{-\theta}, & f_5^2 = 30.e^{-\theta}. \end{array}$$

Let us assume that $\beta = 0.9$, $\pi = (0, 0, 1, 0, 0)$. Then Hv at $\pi = (0, 0, 1, 0, 0)$ is 5.70 and the linear support corresponding to this state is $[-3.57, 4.42, 4.58, 5.49, 6.65]$.

References

[1] D. Bertsekas (1997), *Dynamic-Programming*, PrenticeHall, Englewood Cliffs, New Jersey.

[2] J.E. Goulionis (2005), A model of learning using POMDPs. *Mathematical Inspection* $(M\alpha\theta\eta\mu\alpha\tau k\acute{\eta}\ \epsilon\pi\iota\theta\epsilon\acute{\omega}\rho\eta\sigma\eta)$ **63**, 49–68.

[3] J. E. Goulionis and V. K. Benos (2007), Optimal control limit strategies, using the partially observable Markov decision processes, *Advances and Applications in Statistics* **7**(3), 357–388.

[4] J. E. Goulionis (2007), Structural properties for a two state POMDP with an average cost criterion, *Journal of Statistics and Management Systems* **10** (5), 715–733.

[5] M. L. Littman (1996), *Algorithms for Sequential Decision Making*, Ph.D. thesis, Department of Computer Science, Brown University,

[6] W. S. Lovejoy (1991), Computationally feasible bounds for POMDP, *Operation Research* **39**(1),(162–176).

[7] G. E. Monahan (1982), A survey of partially observable Markov decision processes, *Management Science* **28**,1–16.

[8] G. H. Papadimitriou and J. N. Tsitsiklis (1987), The complexity of Markov decision processes, *Mathematics of Operations Research* **12**, 441–450.

[9] S. J. Edward (1978), The optimal control of partially observable Markov decision processes over the infinite horizon, *Discounted Costs Operation research* **26**, 282–304.

John E. Goulionis, *Department of Statistics and Insurance Science, University of Piraeus, 80 Karaoli & Dimitriou Street, 18534 Piraeus, Greece*.
E-mail: `jgouli@unipi.gr, sondi@otenet.gr`

V. K. Benos, *Department of Statistics and Insurance Science, University of Piraeus, 80 Karaoli & Dimitriou Street, 18534 Piraeus, Greece*.
E-mail: `vbenos@unipi.gr`

Received May 13, 2009
Revised June 16, 2009
Accepted June 27, 2009