43rd IEEE Conference on Decision and Control
December 14-17, 2004
Atlantis, Paradise Island, Bahamas

# Partially Observable Markov Decision Processes With Reward Information

Xi-Ren Cao[1] and Xianping Guo[2]
[1] Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology, Hong Kong
[2] The School of Mathematics and Computational Science
Zhongshan University, Guangzhou, 510275, P. R. China

*Abstract*— **In a partially observable Markov decision process (POMDP), if the reward can be observed at each step, then the observed reward history contains information for the unknown state. This information, in addition to the information contained in the observation history, can be used to update the state probability distribution. The policy thus obtained is called a *reward-information policy* (RI-policy); an optimal RI policy performs no worse than any normal optimal policy depending only on the observation history. The above observation leads to four different problem-formulations for partially observable Markov decision processes (POMDPs) depending on whether the reward function is known and whether the reward at each step is observable.**

Keywords: Reward-information policies

## I. Introduction

Markov decision processes (MDPs) are widely used in many important engineering, economic, and social problems. Partially observable Markov decision processes (POMDPs) are extensions of MDPs in which the system states are not completely observable. The solutions to POMDPs are based on the state probability distributions which can be estimated by using the information obtained via observations. In this paper, we argue that the reward history also contains information for system states, and we provide some studies based on this fact.

We discuss the discrete-time model. An MDP concerns with a state space $X$ and an action space $A$. At time step $t$, $t = 0, 1, \cdots$, the state is denoted as $x_t$ and the action, $a_t$. When an action $a \in A$ is taken at state $x' \in X$, the state transition law is denoted as $P(dx|x', a)$, for $x \in X$. In a POMDP, at any time step $t$, the state $x_t$ is not directly observable; instead, an observation $y_t$ can be made; and $y_t$ depends on $x_{t-1}, x_t$, and $a_{t-1}$ and obeys a probability law $Q(dy_t|x_{t-1}, a_{t-1}, x_t)$ on the observation space $Y$. In particular, it is natural to assume that the the initial observation $y_0$ depends only on $x_0$ and also obeys a probability law $Q_0(dy_0|x_0)$ on $Y$.

In addition, there is a reward (or cost) function $r(x', a, x, w)$, which specifies a random reward (or cost) with $w$ being a random noise representing the uncertainty. Precisely, we denote the reward accumulated in period $[t, t+1)$ as

$$z_{t+1} = r(x_t, a_t, x_{t+1}, w_t),\ t = 0, 1, \cdots, \tag{1}$$

where $\{w_t\}$ is a reward-disturbance process. For simplicity, we assume here that the initial distributions $p_0$ and $\mu_0$ for initial state $x_0$ and initial value $z_0$ respectively, are known. A detailed model will be discussed in Section 2.

Let $\mathbf{a} = \{a_0, a_1, \cdots\}$ be a sequence of actions taken at $t = 0, 1, \cdots$, respectively. With the transition laws, this sequence of actions and the initial state $x_0$ determine a unique state trajectory denoted as $x_t(\mathbf{a}, x_0)$, $t = 0, 1, \cdots$. For simplicity, we will omit the symbol $\mathbf{a}$ and $x_0$ in the expression of state $x_t$. Therefore, for any action sequence $\mathbf{a}$, we can define the discounted- and average-performance criteria as

$$V_\beta(p_0, \mathbf{a}) := \sum_{t=0}^{\infty} \beta^t E[r(x_t, a_t, x_{t+1}, w_t)],\ 0 < \beta < 1, \tag{2}$$

and

$$J(p_0, \mathbf{a}) := \limsup_{N\to\infty} \frac{\sum_{t=0}^{N} E[r(x_t, a_t, x_{t+1}, w_t)]}{N+1}, \tag{3}$$

respectively, where “E” denotes the expectation corresponding to all the randomness involved in the system. When the reward-disturbance $w_t$ is mutually independent and independent to all the other random variables in the system,

Supported by a grant from Hong Kong UGC. The research of the second author has also been supported by the Natural Science Foundations of China and Guangdong Province and by Zhongshan University Advanced Research Center, China. E-mail addresses: eecao@ust.hk and mcsgxp@zsu.edu.cn

we can define

$$\bar{r}(x_t, a_t) = E[r(x_t, a_t, x_{t+1}, w_t)].$$

The expectation is taken with respect to the distribution of $w_t$ and the state transition law $P(dx|x', a)$, which yields the distribution of $x_{t+1}$ given $x_t$ and $a_t$. In this case, the performance criteria (2) and (3) takes the simplified form:

$$V_\beta(p_0, \mathbf{a}) := \sum_{t=0}^{\infty} \beta^t E[\bar{r}(x_t, a_t)], \qquad 0 < \beta < 1, \quad (4)$$

and

$$J(p_0, \mathbf{a}) := \limsup_{N\to\infty} \frac{\sum_{t=0}^{N} E[\bar{r}(x_t, a_t)]}{N+1}, \qquad (5)$$

respectively.

The optimal control problems is to find a sequence $\mathbf{a}$ that maximize the performance (2) or (3) by using the *information available* to us. Such problems are often called the *partially observable Markov decision processes* (POMDPs). The main contribution of this paper is based on a simple fact: when the system state is not completely observable, the observed reward history certainly contains information about the unknown state.

POMDPs based on observation history $\{y_t\}$ only have been widely studied; see [1], [2], [3], [4], [5], [6], [9], [12], [13], [14], [15], [18], [20], [22], [24], [25], [26] for instance. The common approach in the analysis of a POMDP is to first construct a completely observable Markov decision process (i.e., a standard Markov decision process (MDP)) that is equivalent to the POMDP in the sense that not only they have equal optimal values but also their corresponding policies have equal performance. (A policy is a strategy that assigns an action to the system at any time $t$ based on the information available up to $t$.) The state of the equivalent MDP at time $t$ is the conditional distribution of the state of the POMDP given the information available up to time $t$. The existence of the optimal Markov policies for POMDPs, etc, can be easily derived by using the equivalence and the well-developed theory for MDPs. Thus, solutions to POMDP depend on those to MDPs. The reward history can certainly improve the conditional distribution and therefore can improve the policy.

However, the structure of the information contained in the reward history is different from that in the observation history. This can be explained by a comparison with the situation in MDPs. There are two main approaches to MDPs. One is the analytical approach based on the Bellman equation (the optimality equation), in which the reward function $\bar{r}(x', a)$ in (4) or (5) is assumed to be known. This approach belongs largely to the area of operations research. The other was developed in the artificial intelligence community, which takes a learning point-of-view. In this approach, rewards $\bar{z}_t := \bar{r}(x_t, a_t)$ (we will simply denote it as $z_t$ for simplicity) at all times $t = 0, 1, \cdots$ are observed from the system directly. The optimal policy is determined by analyzing these data. In MDPs, because the state $x_t$ is completely observable, knowing the function $\bar{r}$ is equivalent to observing $z_t$. That is, the problem formulations for both approaches are essentially the same for MDPs.

In POMDPs, however, knowing the reward function $r(x', a, x, w)$ (or $\bar{r}(x', a)$) is not the same as observing the value of $z_t = r(x_t, a_t, x_{t+1}, w_t)$ (or $\bar{r}(x_t, a_t)$), because $x_t$ is not observable. Thus, the information available to us for the analytical approach (assuming $r(x', a, x, w)$ is known) and the learning-based approach (assuming $z_t$ is observable) are different. Specifically, if we know only $r(x', a, x, w)$, then we do not know the exact value of $z_t$. On the other hand, if we are able to observe $z_t$ for all $t = 0, 1, \cdots$, we may obtain some more information on the system states; and if, furthermore, we know the function $r(x', a, x, w)$ then we can update the probability distribution of $x_t$ using the fundamental probability theory. Even if we do not know $r(x', a, x, w)$, we may derive it's approximations with statistic inference methods. Thus, there are four different problem formulations for POMDPs, depending on whether the reward function is known and whether the reward at each step is observable, each contains different information about the system state. In all these cases, the optimal policy depends not only on the histories of the observation process and the actions taken at each step, but also on the history of the rewards that are observed. Such a policy will be called a *reward-information policy*.

Applying the same idea to the observation $y_t$, we can formulate another class of POMDP problems where $y_t$, $t = 0, 1, \cdots$, are observable, but the probabilities laws $Q_0(dy_0|x_0)$ and $Q(dy_t|x_{t-1}, a_{t-1}, x_t)$ are unknown or only partially known.

To the best of our knowledge, the information structure regarding the observations of the rewards, $z_t$, $t = 0, 1, \cdots$, has not been well explored in literature. In this paper, we first propose four different problem formulations for POMDPs, as explained in the above discussion. Then we discuss the differences among them as well as the approaches to these problems. We hope our exploratory work can attract research attention to these interesting problems.

## II. Problem Formulations for POMDPs

In general, a POMDP consists of the following elements:

$$\{X, Y, A, P(dx|x', a), Q(dy|x', a, x), \\ Q_0(dy|x), p_0, r(x', a, x, w), \mu_0\}, \tag{6}$$

where:

(a) $X$, *the state space*, is a Borel space;

(b) $Y$, *the observation space*, is also a Borel space;

(c) $A$, *the control set*, is a Borel space too;

(d) $P(dx|x', a)$, *the state transition law*, is a stochastic kernel on $X$ given $X \times A$;

(e) $Q(dy|x', a, x)$, *the observable kernel*, is a stochastic kernel on $Y$ given $X \times A \times X$;

(f) $Q_0(dy|x)$, *the initial observable kernel*, is a stochastic kernel on $Y$ given $X$;

(g) $p_0$, *the initial distribution*, is the (a priori) initial distribution on $X$;

(h) $r(x', a, x, w)$, *the reward function*, is a measurable function on $X \times A \times X \times U$, and takes values in a Borel set $Z$ in the space of all real numbers; $w$ is a disturbance variable with a distribution $\mu_r(\cdot|x', a, x)$ that may depend on $(x', a, x)$;

(i) $\mu_0$, *the initial distribution for the system's initial wealth variable* $z_0$, is a distribution on the set $Z$.

**Definition 2.1.** The model (6) with the above properties (a)-(i), is called a partially observable Markov decision process (POMDP).

A POMDP evolves as follows. At the initial decision step $t = 0$, the system has an initial (unobservable) state $x_0$ with a prior distribution $p_0$ and an initial wealth $z_0$ with the distribution $\mu_0$; in addition, an initial observation $y_0$ is generated according to the kernel $Q_0(\cdot|x_0)$. If at time step $t$ ($\geq 0$) the state of the system is $x_t$ and a control $a_t \in A$ is applied, then the system moves to state $x_{t+1}$ at step $t + 1$ according to the transition law $P(dx_{t+1}|x_t, a_t)$; an observation $y_{t+1}$ is generated by the observation kernel $Q(dy_{t+1}|x_t, a_t, x_{t+1})$, and a reward $z_{t+1} = r(x_t, a_t, x_{t+1}, w_t)$ accumulated in the time period $[t, t + 1)$ is received at time step $t + 1$. (In this definition, $z_{t+1}$, instead of $z_t$, is used; this satisfies causality, i.e., the reward is received after action $a_t$ is taken). Since the effect of $\mu_0$ on the performance criteria is straightforward, we will omit the notation $\mu_0$ even when the quantity indeed depends on it.

For a given sequence of actions $\mathbf{a} = \{a_0, a_1, \cdots\}$, the discounted- and average-performance criteria are defined as (2) and (3). The goal of POMDPs is to find a sequence $\mathbf{a}$ that maximizes one of the performance (2) or (3) by using the information available to us.

**Example 1.** A stochastic control problem is typically modeled as

$$\begin{aligned} &(a) \qquad x_{t+1} = F(x_t, a_t, \xi_t), \ t = 0, 1, \cdots, \\ &(b) \qquad y_{t+1} = G(x_t, a_t, x_{t+1}, \eta_{t+1}), \ t = 0, 1, \cdots, \\ &(c) \qquad y_0 = G_0(x_0, \eta_0), \end{aligned} \tag{7}$$

where $x_t$, $a_t$, and $y_t$ are, respectively, the state, the control, and the observation at time $t$; $\{\xi_t\}$ is the state-disturbance process, and $\{\eta_t\}$ the observation (or measurement) noise. We assume that the initial probability distribution of $x_0$ is $p_0$. (7) is typically called a *partially observable* system.

The system (7) with the reward structure (1) fits the general setting of POMDPs. Let $x_t, y_t, a_t$ take values in Borel spaces $X, Y$ and $A$, respectively. Suppose that $\{\xi_t\}, \{\eta_{t+1}\}$ and $\{w_t\}$ are sequences of independent and identically distributed (in time) random variables with values in Borel spaces $S_s, S_o$ and $U$, respectively, and we assume that they may depend on states and actions. Thus, their distributions are denoted by $\mu_\xi(\cdot|x, a)$ (with $x_t = x$ and $a_t = a$), $\mu_\eta(\cdot|x', a, x)$ (with $x_t = x', x_{t+1} = x$ and $a_t = a$), and $\mu_r(\cdot|x', a, x)$ (with $x_t = x', x_{t+1} = x$ and $a_t = a$), respectively. We also denote by $\mu_{\eta_0}(\cdot)$ the distribution of $\eta_0$ taking values in $S_o$. Let $F$, $G$ and $G_0$ be given measurable functions, and $x_0$ be independent of $\{\xi_t\}, \{\eta_{t+1}\}$ and $\{w_t\}$.

We denote by $I_B[\cdot]$ the indicator functions of any set $B$. Then the state transition law $P(\cdot|x, a)$ is given by

$$P(B|x, a) = \int_{S_s} I_B[F(x, a, u)]\mu_\xi(du|x, a)$$

for every Borel set $B$ in $X$. Similarly, if $x_t = x', a_t = a$ and $x_{t+1} = x$, the observation kernel $Q(\cdot|x', a, x)$ is given by

$$Q(C|x', a, x) = \int_{S_o} I_C[G(x', a, x, v)]\mu_\eta(dv|x', a, x)$$

for all Borel set $C$ in $Y$. If $x_0 = x$, then

$$Q_0(C'|x) = \int_{S_o} I_C[G_0(x, s)]\mu_{\eta_0}(ds)$$

for all Borel set $C'$ in $Y$; whereas, if $x_t = x', x_{t+1} = x$ and $a_t = a$, then the observation value $z_{t+1}$ is obtained by the *reward-observation kernel* $R(\cdot|x', a, x)$ on $Z$ given $X \times A \times X$, defined by

$$R(D|x', a, x) := \int_U I_D[r(x', a, x, s)]\mu_r(ds|x', a, x), \tag{8}$$

for all Borel set $D$ in $Z$. Thus, the above discussion regarding the reward information applies to this control problem. It is easy to see that the same is true for time variant systems with $F$ and $G$ replaced by $F_t$ and $G_t$ depending on $t$, respectively.□

However, the information available to us are different for POMDPs depending on whether the reward function $r(x', a, x, w)$ is known and whether the reward at each step $z_t$, $t = 0, 1, \cdots$, can be observed. This leads to four different problem formulations for POMDPs specified as follows:

(a) The function $r(x', a, x, w)$ is known, and the reward $z_t$ can be observed at each step $t$;

(b) The function $r(x', a, x, w)$ is known, but the reward $z_t$ cannot be observed at each step $t$;

(c) The function $r(x', a, x, w)$ is unknown, but the reward $z_t$ can be observed at each step $t$;

(d) The function $r(x', a, x, w)$ is unknown, and the reward $z_t$ cannot be observed at each step $t$.

In the standard Markov decision processes (e.g. [1], [7], [8], [17]), the reward function $\bar{r}(x, a)$ does not involve randomness. With analytical approaches, it is natural to assume that the reward function is known. However, with on-line (or sample path based) approaches such as reinforcement learning, it is convenient to assume that the reward at each step $z_t$ can be exactly observed, which is used to update the estimate of the value function. Because the state is completely observable, knowing the function $\bar{r}(x, a)$ is the same as knowing the reward $z_t$. Therefore, the assumptions in both cases are equivalent. In the case of POMDPs, these assumptions have different implications and we will discuss the four cases listed above separately.

**Case (a).** ($r(x', a, x, w)$ known, $z_t$ observable) We emphasize that there is a fundamental difference between Cases a and b discussed below in POMDP problems. If $z_t = r(x', a, x, w)$ is observable, then the value of $z_t$ certainly provides information to state $x$ via $r(x', a, x, w)$. Therefore, once $z_t$ is obtained, we can update the conditional distribution of the state, which should be more accurate than only the observation $y$ is used. We refer to this case as POMDPs with *full reward information* (POMDPs-FRI). This case will be discussed in details in Sections 3 and 4. We will show that a POMDP-FRI can be converted to an MDP problem, both observation histories $y_t$ and $z_t$ provide information for the distribution of state $x_t$. Therefore, the optimal performance of Case a (POMDP-FRI) should be no worse than that of Case b below (POMDPs-PRI).

**Case (b).** ($r(x', a, x, w)$ known, $z_t$ not observable) This is the standard formulation for most analytical approaches. We use the classical LQG problem in stochastic control as an example to illustrate the idea. The system is described by a linear stochastic differential equation,

$$\frac{dx}{dt} = F(t)x + G(t)u + w(t),$$

where $x$ is the $m$-dimensional state vector, $u$ is the control action, and $w(t)$ is a Gaussian white noise. The measurement is an $n$-dimensional vector

$$y(t) = H(t)x(t) + v(t),$$

with $v(t)$ being a Gaussian white noise. The performance to be maximized is

$$J = E\left\{\frac{1}{2}\int_{t_0}^{t_f} [x^T, u^T] \begin{bmatrix} A(t) & N(t) \\ N^T(t) & B(t) \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} dt\right\}, \tag{9}$$

where $t_f$ is a termination time. If we write

$$z(t) = [x^T, u^T] \begin{bmatrix} A(t) & N(t) \\ N^T(t) & B(t) \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix},$$

then $J = E\{\frac{1}{2}\int_{t_0}^{t_f} z(t)dt\}$. Apparently, we assume that the form of $z(t)$, i.e., $A(t)$, $B(t)$ and $N(t)$ are known, but we do not assume that the value of $z(t)$ can be obtained at any time $t$. Because the state is partially observable and the reward function is known, the reward is also partially observable. We refer to this case as POMDPs with *partial reward information* (POMDPs-PRI). Although the LQG problem is defined in a continuous time domain with a finite horizon, the basic principle for problem formulation is the same as our model (6).

Case b is well studied in literature and it is well known that the problem can be converted to an MDP with all possible state distributions as its states (called belief states).

**Case (c).** ($r(x', a, x, w)$ unknown, $z_t$ observable) For many practical systems, the function $r(x', a, x, w)$ is very complicated and cannot be exactly determined; however, the instant reward $z_t$ can be observed. For instance, in communication networks, even the state of the system is hard to observe, but the instant reward (or cost), such as dropping a packet, can be observed. In addition, the on-line (or sample-path-based) optimization approaches depend on observing the current reward to adjust their estimates for the value functions (or potentials). In reinforcement learning algorithms, the essential fact is the value of the reward at each step, the form of the reward function is not needed. Therefore, Case (c) is also practically important. We refer to this case as POMDPs with *incomplete reward information* (POMDPs-IRI).

Although the form of $r(x', a, x, w)$ is unknown, with the reward observation sequence $z_t = r(x_t, a_t, x_{t+1}, w_t)$, the distribution of $w_t$, the distribution of $x_t$ obtained from the observation history $y_t$, and the action $x_t$ we can try to estimate the function $r(x', a, x, w)$ using statistic theory. Therefore, with $z_t$ observed, using the estimated function, we can apply similar approaches as Case (a) to obtain more information about $x_t$ and possibly a better policy than Case (b). This is a difficult problem and will be left for further research.

**Case (d).** ($r(x', a, x, w)$ unknown, $z_t$ not observable) Few information is available in this case. However, if we can obtain (observe) the total reward in a time period, such as the value of $J$ in the LQG problem (9), we still can get some information about how good we are doing in the entire interval $[t_0, t_f)$. Therefore, if we are allowed to repeat the operation, we will be still able to learn from the past operations. Thus, this case still presents a meaningful research (albeit hard) problem. We refer to this case as POMDPs with *no reward information* (POMDPs-NRI).

To understand more about the above cases, we give an example.

**Example 2.** A robot moves among three rooms lining up in a row. The rooms are denoted as L, M, and R, representing the left, the middle, and the right rooms, respectively. The robot can take two actions in each room. In room M, if action $A_l$ ($A_r$) is taken, the robot will move to room L with probability 0.8 (0.2) and to room R with probability 0.2 (0.8). In room L, if action $A_l$ ($A_r$) is taken, the robot will hit the left wall then stay in room L with probability 0.8 (0.2), or will move to room M with probability 0.2 (0.8). Similarly, In room R, if action $A_r$ ($A_l$) is taken, the robot will hit the right wall then stay in room R with probability 0.8 (0.2), or will move to room M with probability 0.2 (0.8).

A unit cost will be received if the robot hits a wall. The cost function is $r(L, L) = r(R, R) = 1$. and $r = 0$ for other cases. The goal is to design a policy that minimizes the long-run average cost.

The system states are L, M, and R. With the MDP model, the state is observable, and the optimal policy is obvious: Take action $A_r$ at state $L$ and action $A_l$ at state $R$. With POMDPs, the state is not observable and we need to consider four cases (assume there is no additional observation $y$).

Case a. $r$ is known and $z_t = r(x_t, x_{t+1})$ is observable. For example, we know that when the robot hits a wall we will hear a beep. Suppose that a priori probabilities of the states are $p_0(L)$, $p_0(M)$, and $p_0(R)$, respectively. If we hear a beep after action $A_l$ ($A_r$) is taken, we have the following conditional probabilities

$$p(beep|L, A_l) = 0.8, \quad p(beep|M, A_l) = 0,$$
$$p(beep|R, A_l) = 0.2.$$

With this, the state probability distribution after a beep can be easily updated.

Case b. $r$ is known but $z_t = r(x_t, x_{t+1})$ is not observable. This is a standard POMDP problem. No additional information can be obtained by rewards. The state distribution has to be estimated by observations. In this particular problem, no additional observation is available. Given any initial state probability distribution $p_0$, the system eventually will reach some steady state distribution denoted as $\pi = (\pi(L), \pi(M), \pi(R))$. Suppose that with this state distribution we take a random policy: take action $A_l$ with probability $p_l$ and take action $A_r$ with probability $p_r$. Then the transition matrix (list the states in the order of L, M, and R):

$$P = \begin{bmatrix} 0.8p_l + 0.2p_r & 0.2p_l + 0.8p_r & 0 \\ 0.8p_l + 0.2p_r & 0 & 0.2p_l + 0.8p_r \\ 0 & 0.8p_l + 0.2p_r & 0.2p_l + 0.8p_r \end{bmatrix}$$

Then we have $\pi = \pi P$. The problem becomes minimize

$$\pi(L)(0.8p_l + 0.2p_r) + \pi(R)(0.8p_r + 0.2p_l)$$

with $p_l + p_r = 1$.

Case c. $r$ is unknown and $z_t = r(x_t, x_{t+1})$ is observable. That is, we can hear a beep when a cost is incurred, but we don't know why there is a beep. In this case, we need to learn the pattern for the beeps. For example, we may find that if we take action $A_l$ twice and mean while we hear the beep twice, then it is more likely that we will hear a beep if we take $A_l$ again; and so on. This is the learning-based approach. After learning for some times, we may find the form of the function $r$ based on the patterns we learned. Then the problem becomes Case a.

Case d. $r$ is unknown and $z_t = r(x_t, x_{t+1})$ is not observable. If the total cost in a finite period of $N$ steps can be obtained and the experience is repeatable, we can still do something. There are $2^N$ possible ways to choose the actions. We can search for the best choice in this space of $2^N$ elements using various approaches such as the generic algorithms etc. □

As we can see, Case b is the standard POMDP problem and has been widely studied, Case c is a difficult problem involving the estimation of the reward function $r$, Case d contains little information and may resort to searching. The rest of the paper mainly focuses on Case a, POMDPs-FRI.

## III. Conclusion

Our main observation is that the reward history in a POMDP contains information about the distribution of the unknown state. This leads to four different problem-formulations for POMDPs depending on whether the reward function $r(x', a, x, w)$ is known and whether the reward at each step $z_t$ is observable. The policy depending on both the observation and reward histories is called a reward-information (RI) policy. An optimal RI policy performs no worse than any normal optimal policy.

POMDPs-FRI (reward function known and $z_t$ observable) can be converted to the standard MDPs, and the optimality conditions for both the discounted- and average-performance criteria are obtained. For POMDPs-PRI (reward function unknown and $z_t$ observable), one approach is to approximately estimate the function $\bar{r}(x', a) = E[r(x', a, x, w)]$ and then apply the solution to POMDPs-FRI. This certainly requires further research. In most reinforcement learning algorithms, it is assumed that $z_t$ can be observed; these problems therefore belong to POMDPs-PRI or POMDPs-FRI. POMDPs-IRI (reward function known and $z_t$ unobservable) is a typical problem in control theory (e.g., the LQG problem). Finally, POMDPs-NRI (reward function unknown and $z_t$ unobservable) only make sense when the process repeats and the total reward is known. The study in this paper demonstrates the fundamental difference between the analytical approaches (no observation on reward is made) and the learning based approaches in the POMDP framework.

Finally, we note that the same idea applies to the observation $y_t$. That is, we may assume that we can observe $y_t$ but its distributions $Q_0(dy_0|x_0)$ and $Q(dy_t|x_{t-1}, a_{t-1}, x_t)$ are unknown or only partially known. For example, in Example 1, the function $G$ is unknown or the distribution of $\eta_{t+1}$ is unknown, and in the LQG problem, the variance of the Gaussian noise $v(t)$ is unknown, etc. Thus, we can formulate another class of POMDPs which is similar to Case c for the RI policies.

## References

[1] Arapostathis, A., Borkar, V. S., Fernandez-Gaucherand, E., Ghosh, M. K. and Markus, S. I., *Discrete-time controlled Markov processes with average cost criterion: a survey*, SIAM J. Control Optima, 31 (1993), 282-344.

[2] Andriyanov, V. A., Kogan, I. A. and Umnov, G., A., *Optimal control of partially observable discrete Markov process*, Automat. Remote Control, 4 (1980), 555-561.

[3] Borkar, V. S., *Average cost dynamic programming equations for controlled Markov chains with partial observations*, SIAM J. Control Optim., 39 (2001), 673-681.

[4] Borkar, V. S., *Dynamic programming for ergodic control with partial observations*, Stoch. Proc. Appl., 103 (2003), 293-310.

[5] Fernandez-Gaucherand, E., Arapostathis, A. and Markus, S. I., *On the average cost-average optimality equation and the structure of optimal policies for partially observable Markov decision processes*, Ann. Oper. Res., 15 (1991), 425-432.

[6] Hernández-Lerma, O., *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.

[7] Hernández-Lerma, O. and Lasserre, J.B., *Discrete-time Markov control processes: basic optimality criteria*, Springer, New York, 1996.

[8] Hernández-Lerma, O. and Lasserre, J.B., *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

[9] Hernández-Lerma, O. and Romera, R., *Limiting discounted-cost control of partially observable stochastic systems*, SIAM J. Control Optim., 40 (2001), 348-369.

[10] Ho, Y. C. and Cao, X. R., *Perturbation Analysis of Discrete-Event Dynamic Systems*, Boston, Kluwer, 1991.

[11] Kemeny, J.G. and Snell, J.L., *Finite Markov chains.* New York: Van Nostrand, 1990.

[12] Lovejoy, W. S., *Some monotonicity results for partially observable Markov decision processes*, Oper. Res., 35 (1987), 736-743.

[13] Lovejoy, W. S., *A survey of algorithmic results for partially observable Markov decision processes*, Ann. Oper. Res., 35 (1991), 47-66.

[14] Ohnish, M., Kawai, H. and Mine, H., *An optimal inspection and replacement policy under incomplete state information*, European J. Oper. Res., 27 (1986), 117-128.

[15] Ohnish, M., Kawai, H. and Mine, H., *An optimal inspection and replacement policy under incomplete state information: average cost criterion*, In Stochastic Models in Reliability Theory (Osaki, S. and Hatoyama, Y., eds), Lect. Notes Econ., Math. Systems, Vol. 235, Springer-Verlag, Berlin, 1984, 187-197.

[16] Platzman, L. K., *Optimal infinite horizon undiscounted control of finite probability systems*, SIAM J. Control Opti., 18 (1980), 362-380.

[17] Puterman, M.L., *Markov Decision Processes: Discrete Stochastic dynamic Programming.* New York: Wiley, 1994.

[18] Rhenius, D., *Incomplete information in Markovian decision models*, Ann. Statis., 2 (1974), 1327-1334.

[19] Ross, S. M., *Quality control under Markovian deterioration*, Management Sci., 17 (1971), 587-596.

[20] Sondic, E. J., *The optimal control of partially observable Markov decision processes*, Ph.D. thesis, Electrical Engineering Dept., Stanford University, Stanford, CA, 1971.

[21] Ross, S. M., *Arbitrary state Markovian decision processes*, Ann. Math. Statist., 39 (1968), 2118-2122.

[22] Wang, W., *Optimal replacement policy with unobservable states,* J. Appl. Prob., 14 (1977), 340-348.

[23] Wang, W., *Computing optimal quality control policies — two actions,* J. Appl. Prob., 14 (1977), 340-348.

[24] White, C. C., *Bonds on optimal cost for a replacement problem with partial observation*, Naval res. Logist. Quart., 26 (1979), 415-422.

[25] White C. C., *Optimal control— limit strategies for a partially observed replacement problem*, Internat. J. Systems Sci., 10 (1979), 321-331.

[26] Yoshikazu, S. and Tsuneo, Y., *Discrete-time Markovian decision processes with incomplete state observation*, Ann. Math. Statist., 41 (1970), 78-86.