Technical Notes and Correspondence

Partially Observable Markov Decision Processes With Reward Information: Basic Ideas and Models

Xi-Ren Cao and Xianping Guo

***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 on 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 POMDPs depending on whether the reward function is known and whether the reward at each step is observable. This exploratory work may attract attention to these interesting problems.**

***Index Terms*—Partially observable Markov decision process (POMDP), reward-information policy.**

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 note, we argue that the reward history also contains information for system states, and we provide some studies based on this fact. (Although such reward information has been used in a special case of the bandit problem [1], [8] for instance, our POMDP formulation is for the general case with the Markov model.)

We discuss the discrete time model. An MDP concerns with a state space X and an action space A. At time step t, $t = 0, 1, \ldots$, 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 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 $r(x', a, x, w)$, with w being a random noise representing the uncertainty of rewards. Precisely, we denote the reward accumulated in period $[t, t+1)$ as

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

Manuscript received March 7, 2005; revised October 23, 2006 and November 7, 2006. Recommended by Associate Editor Y. Wardi. This work was supported by the Hong Kong UGC, NSFC, and RFDP.

X.-R. Cao is with the Hong Kong University of Science and Technology, Hong Kong (e-mail: eecao@ust.hk).

X. Guo is with Zhongshan University, Guangzhou 510275, China (e-mail: mcsgxp@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TAC.2007.894520

where $\{w_t\}$ is a reward-disturbance process and is assumed to have a distribution law $\mu_w(du|x, a)$. For simplicity, we assume here that the initial distributions p_0 and μ_0 for initial state x_0 and initial reward z_0 respectively, are known. A detailed model will be discussed in Section II.

Let $\mathbf{a} = \{a_0, a_1, \ldots\}$ be the sequence of actions a_t taken at $t = 0, 1, \ldots$, respectively. With the transition law, this sequence of actions and the initial distribution p_0 determine a unique probability measure with its corresponding expectation denoted simply by E and the state trajectories denoted as $x_t(\mathbf{a}, p_0)$, $t = 0, 1, \ldots$. For simplicity, we will omit the symbol $\mathbf{a}$ and p_0 in the expression of state x_t. Therefore, for any action sequence $\mathbf{a}$ and an initial distribution p_0, we can define the discounted- and average-performance criteria as

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

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. When the reward-disturbance w_t is mutually independent and identically distributed and also independent to all the other random variables in the system, we can define

$$\bar{r}(x_t, a_t) = \int r(x_t, a_t, x, u) P(dx|x_t, a_t) \mu_w(du|x_t, a_t). \tag{4}$$

In this case, the performance criteria (2) and (3) take the simplified form with $E[r(x_t, a_t, x_{t+1}, w_t)]$ replaced by $E[\bar{r}(x_t, a_t)]$.

The goal of the optimal control problem is to find a sequence $\mathbf{a}$ that maximizes the performance (2) or (3) by using the *information available* to us. Such problems are often called the *POMDPs*. The main contribution of this note is based on a simple fact: when the system state is not completely observable, the observed reward history certainly contains information on the unknown state. With this observation, we can characterize the POMDP-based learning and optimization problem into four categories.

POMDPs based on only observation history $\{y_t\}$ have been widely studied; see [2], [3], [7], [9], [10], [13], and [14], for instance. The common approach in the analysis of a POMDP is to first construct a completely observable MDP (i.e., a standard 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. 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 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 policies.

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) 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 $t = 0, 1, \ldots$, 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 only know the function $r(x', a, x, w)$, 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 $t = 0, 1, \ldots$, 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 approximations of it 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* (RI-policy); see Definition 2.2 for details.

The information provided by the observations of the rewards z_t, $t = 0, 1, \ldots$, is already discussed in bandit problems in [1] and [8] for instance. However, we feel that none has clearly classified the problems into the four categories as we do in this note in the general framework, and the bandit problem is not a standard POMDP.

In this note, 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. After that, we examine in some details the case where both the function $r(x', a, x, w)$ is known and the reward z_t, $t = 0, 1, \ldots$, are observable. The existence and algorithms of optimal-RI policies for the discounted- or average-reward criteria are mainly concerned, and 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:

$$\{X, Y, A, P(dx|x', a), Q(dy|x', a, x) \quad Q_0(dy|x), p_0, r(x', a, x, w), \mu_0\} \quad (5)$$

where

i) X, *the state–space*, is a Borel space;
ii) Y, *the observation space*, is also a Borel space;
iii) A, *the control set*, is a Borel space as well;
iv) $P(dx|x', a)$, *the state transition law*, is a stochastic kernel on X given $X \times A$;
v) $Q(dy|x', a, x)$, *the observable kernel*, is a stochastic kernel on Y given $X \times A \times X$;
vi) $Q_0(dy|x)$, *the initial observable kernel*, is a stochastic kernel on Y given X;
vii) p_0, *the initial distribution*, is the (*a priori*) initial distribution on X;
viii) $r(x', a, x, w)$, *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 $\mu_w(\cdot|x', a, x)$ that may depend on (x', a, x);
ix) μ_0, *the initial distribution for the system's initial reward* z_0, is a distribution on the set Z.

Definition 2.1: The model (5) with properties i)–ix), 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 priori* distribution p_0 and an initial reward z_0 with the distribution μ_0; in addition, an initial observation y_0 is generated according to the kernel $Q_0(\cdot|x_0)$. If at time step t (≥ 0) the system is at state 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 μ_0 on the performance criteria is straightforward, we will omit the notation μ_0 even when the quantities indeed depend on it.

At time $t \geq 0$, all the information about the past can be represented by the observation-reward-action history denoted as $h_t := (p_0, y_0, z_0, y_1, z_1, a_1, \ldots, y_{t-1}, z_{t-1}, a_{t-1}, y_t, z_t)$. Thus, the action a_t at time $t \geq 0$ can be taken according to a given kernel $\pi_t(\cdot|h_t)$ on A.

Definition 2.2: The sequence of stochastic kernels $\{\pi_t, t = 0, 1, \ldots, \}$ is called a reward-information (RI) policy.

For a given RI-policy π, because the action sequence $\mathbf{a}$ is determined by π, the discounted- and average-performance criteria are defined as (2) and (3) with $\mathbf{a}$ replaced by π. The goal of POMDPs is to find a π that maximizes/minimizes the performance (2) or (3) over all RI-policies.

1) Example 1: A stochastic control problem is typically modeled as

$$\begin{aligned} &\text{a)} \quad x_{t+1} = F(x_t, a_t, \xi_t),\ t = 0, 1, \ldots \\ &\text{b)} \quad y_{t+1} = G(x_t, a_t, x_{t+1}, \eta_{t+1}),\ t = 0, 1, \ldots \\ &\text{c)} \quad y_0 = G_0(x_0, \eta_0) \end{aligned} \quad (6)$$

where x_t, a_t, and y_t are, respectively, the state, the control, and the observation at time t; $\{\xi_t\}$ is a state-disturbance process, and $\{\eta_t\}$ an observation (or measurement) noise. We assume that the initial probability distribution of x_0 is p_0. Equation (6) is typically called a *partially observable* system.

The system (6) with the reward structure (1) fits the general setting of POMDPs. Let x_t, y_t, a_t take values in Borel space 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 $\mu_w(\cdot|x', a, x)$ as in viii) of (5), respectively. We also denote by $\mu_{\eta_0}(\cdot)$ the distribution of η_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 function of any set B, and by $\mathcal{B}(S)$ the Borel σ-algebra of any Borel space S. Then the state transition law $P(\cdot|x, a)$ is given by

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

and the initial observation kernel $Q(\cdot|x', a, x)$ 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 $C \in \mathcal{B}(Y)$. If $x_0 = x$, then

$$Q_0(C'|x) = \int_{S_o} I_C[G(x, s)]\mu_{\eta_0}(ds), C' \in \mathcal{B}(Y)$$

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

$$R(D|x', a, x) := \int_U I_D[r(x', a, x, s)]\mu_w(ds|x', a, x) \tag{7}$$

for all Borel set $D \in \mathcal{B}(Z)$. 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 $z_t, t = 0, 1, \ldots$, 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 MDPs (e.g., [2], [4]–[6], and [11]), 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 online (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.

2) 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 later 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_t is used. We refer to this case as POMDPs with *full reward information* (POMDPs-FRI). Since both observation histories y_t and z_t provide information for the distribution of state x_t, the optimal performance of Case (a) (POMDP-FRI) should be no worse than that of Case (b) (POMDPs-PRI).

3) 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 a 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

$$J = E\left\{\int_{t_0}^{t_f} [x^T, u^T] \begin{bmatrix} A(t) & N(t) \\ N^T(t) & B(t) \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} dt\right\} \tag{8}$$

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 $J = E\{\int_{t_0}^{t_f} z(t)dt\}$. 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 the problem formulation is the same as our model (5).

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

4) 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 online (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 a_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 a possibly better policy than Case (b). This is a difficult problem and will be left for further research.

5) 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 (8), 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 experiment, 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 previous cases, we give an example.

6) 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 denoted as $r(x_t, x_{t+1})$ with $r(L, L) = r(R, R) = 1$, 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).

7) 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 can hear a beep. Suppose that the *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:

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

With this, the state probability distribution after hearing or not hearing a beep can be easily updated.

8) 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, if there is no periodicity, 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 to minimize

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

with $p_l + p_r = 1$.

9) 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.

10) 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 actions. We can search for the best choice in this space of 2^N elements using various approaches such as the ordinal optimization [12], [15] and genetic algorithms etc.

As we can see, Case b is the standard POMDP problem and has been widely studied, Case a can be solved by the standard POMDP methods once the conditional distribution of the system state is updated by the additional information from the rewards. Case c is a difficult problem involving the estimation of the reward function r, Case d contains little information and may resort to searching.

III. Conclusion

Our main observation is that the reward history in a POMDP contains information on 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.

POMDPs-FRI (reward function known and z_t observable) can be converted to the standard MDPs. 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 note 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 η_{t+1} is unknown, and in the LQG problem, the variance of the Gaussian noise $v(t)$ is unknown. Thus, we can formulate another class of POMDPs.

The study on the four problems of POMDPs is in progress.

Acknowledgment

The authors would like to thank the Associate Editor and the anonymous referees for many fine comments and suggestions that have improved this note.

References

[1] V. Anantharam, P. Varaiya, and J. Walrand, "Asymtotically efficient allocation rules for the mutiarmed bandit problem with multiple plays — Part II: Markovian reward," *IEEE Trans. Autom. Control*, vol. AC-32, no. 11, pp. 977–982, Nov. 1987.

[2] A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Markus, "Discrete-time controlled Markov processes with average cost criterion: A survey," *SIAM J. Control Optim.*, vol. 31, pp. 282–344, 1993.

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

[4] X.-R. Cao and X. P. Guo, "A unified approach to Markov decision problems and performance sensitivity analysis with discounted and average criteria: Multichain cases," *Automatica*, vol. 40, pp. 1749–1759, 2004.

[5] X. P. Guo and X.-R Cao, "Optimal control of ergodic continuous-time Markov chains with average sample-path rewards," *SIAM J. Control Optim.*, vol. 44, pp. 29–48, 2005.

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

[7] O. Hernández-Lerma and R. Romera, "Limiting discounted-cost control of partially observable stochastic systems," *SIAM J. Control Optim.*, vol. 40, pp. 348–369, 2001.

[8] S. R. Kulkarni and G. Lugosi, "Finite-time lower bounds for the two-armed bandit problem," *IEEE Trans. Autom. Control*, vol. 45, no. 4, pp. 711–714, Apr. 2000.

[9] W. S. Lovejoy, "A survey of algorithmic results for partially observable Markov decision processes," *Ann. Oper. Res.*, vol. 35, pp. 47–66, 1991.

[10] M. Ohnish, H. Kawai, and H. Mine, "An optimal inspection and replacement policy under incomplete state information," *Eur. J. Oper. Res.*, vol. 27, pp. 117–128, 1986.

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

[12] C. Song, X. H. Guan, and Y. C. Ho, "Constrained ordinal optimization – A feasibility based approach," *Discrete Event Dyna. Syst.: Theory Appl.*, vol. 16, pp. 279–299, 2006.

[13] C. C. White, "Bonds on optimal cost for a replacement problem with partial observation," *Naval Res. Logist. Quart.*, vol. 26, pp. 415–422, 1979.
[14] S. Yoshikazu and Y. Tsuneo, "Discrete-time Markovian decision processes with incomplete state observation," *Ann. Math. Statist.*, vol. 41, pp. 78–86, 1970.
[15] Q. C. Zhao, Y. C. Ho, and Q. S. Jia, "Vector ordinal optimization," *J. Optim. Theory Appl.*, vol. 125, pp. 259–274, 2005.

Distributed Geodesic Control Laws for Flocking of Nonholonomic Agents

Nima Moshtagh and Ali Jadbabaie

***Abstract*—We study the problem of flocking and velocity alignment in a group of kinematic nonholonomic agents in 2 and 3 dimensions. By analyzing the velocity vectors of agents on a circle (for planar motion) or sphere (for 3-D motion), we develop a geodesic control law that minimizes a misalignment potential and results in velocity alignment and flocking. The proposed control laws are distributed and will provably result in flocking when the underlying proximity graph which represents the neighborhood relation among agents is connected. We further show that flocking is possible even when the topology of the proximity graph changes over time, so long as a weaker notion of joint connectivity is preserved.**

***Index Terms*—Cooperative control, distributed coordination, flocking, multiagent systems.**

I. Introduction

Cooperative control of multiple autonomous agents has become a very active part of control theory research. The main underlying theme of this line of research is to analyze and/or synthesize spatially distributed control architectures that can be used for motion coordination of large groups of autonomous vehicles. Each vehicle is assumed to be capable of local sensing and communication, and there is often a global objective, such as swarming, or reaching a stable formation, etc. A nonexhaustive list of relevant research in control theory and robotics includes [1], [3], [5], [8]–[10], [12], [13], [19].

On the other hand, such problems of distributed coordination have also been studied in areas as diverse as statistical physics and dynamical systems (in the context of synchronization of oscillators and alignment of self propelled particles [18]), in biology, and ecology, and in computer graphics in the context of artificial life and simulation of social aggregation phenomena, and in distributed computation [17], in the context of reaching consensus in parallel and distributed processing.

Most of the above cited research on distributed control of multivehicle systems has been focused on fully actuated systems [16], or planar under-actuated systems [8]. Our goal here is to develop motion coordination algorithms that can be used for distributed control of a group of nonholonomic vehicles in 2 and 3 dimensions. Using results of Bullo *et al.* [2] we develop *geodesic control laws* that result in flocking and velocity alignment for nonholonomic agents in 3 dimensions.

In order to introduce the idea of a geodesic control law to the reader, we start with the special case of planar motion in Section III. We will show that the planar version of such a control law (where the velocity vector is restricted to stay on a circle) is exactly the well-known Kuramoto model of coupled nonlinear oscillators [6], [14]. Such a control law is a gradient controller that minimizes a potential function which represents the aggregate "misalignment energy" between all agents. In Section V we return to the general case of 3-D motion and we develop control laws that result in stable coordination and velocity alignment of a group of agents with a fixed connectivity graph. In Section VI, we show that flocking is possible even when the topology of the proximity graph changes over time. Finally, in Section VII, we provide simulations that show the effectiveness of the designed controllers. But, let us review the concepts of graph theory that we use in this note for stability analysis.

II. Graph Theory Preliminaries

In this section, we introduce some standard graph theoretic notation and terminology. An (undirected) graph $\mathbb{G}$ consists of a vertex set, $\mathcal{V}$, and an edge set $\mathcal{E}$, where an edge is an unordered pair of distinct vertices in $\mathbb{G}$. If $x, y \in \mathcal{V}$, and $(x, y) \in \mathcal{E}$, then x and y are said to be adjacent, or neighbors and we denote this by writing $x \sim y$. The number of neighbors of each vertex is its valence. A path of length r from vertex x to vertex y is a sequence of $r+1$ distinct vertices starting with x and ending with y such that consecutive vertices are adjacent. If there is a path between any two vertices of a graph $\mathbb{G}$, then $\mathbb{G}$ is said to be connected. If there is such a path on a directed graph ignoring the direction of the edges, then the graph is weakly connected.

The adjacency matrix $A(\mathbb{G}) = [a_{ij}]$ of an (undirected) graph $\mathbb{G}$ is a symmetric matrix with rows and columns indexed by the vertices of $\mathbb{G}$, such that $a_{ij} = 1$ if vertex i and vertex j are neighbors and $a_{ij} = 0$, otherwise. The valence matrix $D(\mathbb{G})$ of a graph $\mathbb{G}$ is a diagonal matrix with rows and columns indexed by $\mathcal{V}$, in which the (i,i)-entry is the valence of vertex i. The (un)directed graph of a (symmetric) matrix is a graph whose adjacency matrix is constructed by replacing all nonzero entries of the matrix with 1. Matrix A has *property SC* if and only if $|A|$ is the adjacency matrix of a strictly connected graph.

The symmetric singular matrix defined as:

$$L(\mathbb{G}) = D(\mathbb{G}) - A(\mathbb{G})$$

is called the Laplacian of $\mathbb{G}$. The Laplacian matrix captures many topological properties of the graph. The Laplacian L is a positive semidefinite M-matrix (a matrix whose off-diagonal entries are all nonpositive) and the algebraic multiplicity of its zero eigenvalue (i.e., the dimension of its kernel) is equal to the number of connected components in the graph. The n-dimensional eigenvector associated with the zero eigenvalue is the vector of ones, $\mathbf{1}$.

Given an orientation of the edges of a graph, we can define the *incidence matrix* of the graph to be a matrix B with rows indexed by vertices and columns indexed by edges with entries of 1 representing the source of a directed edge and -1 representing the sink. The Laplacian matrix of a graph can also be represented in terms of its incidence matrix as $L = BB^T$ independent of the orientation of the edges.

Manuscript received September 16, 2005; revised March 10, 2006. Recommended by Associate Editor F. Bullo. This work was supported in part by the following grants: ARO-MURI Grant W911NF-05-1-0381, ONR YIP N00014-04-1-046 and ONR N00014-06-1-0436, and NSF-ECS-0347285.

The authors are with the department of Electrical and Systems Engineering and GRASP Laboratory, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: nima@grasp.upenn.edu; jadbabai@grasp.upenn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2007.894528