Forty-Sixth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 23-26, 2008

Identifying tractable decentralized control problems on the basis of information structure

Aditya Mahajan Ashutosh Nayyar Demosthenis Teneketzis
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109–2122, USA
Email: {adityam,anayyar}@umich.edu, teneket@eecs.umich.edu

Abstract—**Sequential decomposition of two general models of decentralized systems with non-classical information structures is presented. In model A, all agents have two observations at each step: a common observation that all agents observe and a private observation of their own. The control actions of each agent is based on all past common observations, the current private observation and the contents of its memory. At each step, each agent also updates the contents of its memory. A cost function, which depends on the state of the plant and the control actions of all agents, is given. The objective is to choose control and memory update functions for all agents to either minimize a total expected cost over a finite horizon or to minimize a discounted cost over an infinite horizon. In model B, the agents do not have any common observation, the rest is same as in model A. The key idea of our solution methodology is the following. From the point of view of a fictitious agent that observes all common observations, the system can be viewed as a centralized system with partial observations. This allows us to identify information states and obtain a sequential decomposition. When the system variables take values in finite sets, the optimality equations of the sequential decomposition are similar to those of partially observable Markov decision processes (POMDP) with finite state and action spaces. For such systems, we can use algorithms for POMDPs to compute optimal designs for models A and B.**

I. Introduction

Decentralized systems arise in different branches of engineering. Examples include the Internet, telecommunication networks, sensor networks, surveillance networks, monitoring and diagnostic systems, MANET (mobile ad-hoc networks), cognitive radio, control of UAVs (unmanned aerial vehicles), robotics, etc. Most of these applications are independent areas of research. However, from an abstract level, these applications have similar salient features and similar design difficulties. We believe that if we can capture these salient features in a simple model and understand how to resolve the conceptual difficulties for that model, then these insights would provide design guidelines for applications. In this paper, we study two general models of decentralized systems that arise in different applications and show that, when viewed appropriately, the optimal design of these decentralized systems is similar to the optimal design of partially observable Markov decision processes (POMDPs). This enables us to obtain a sequential decomposition of these models, and allows us to use the computational results of POMDPs to obtain optimal solutions for these models.

Decentralized systems consist of multiple components (or agents); each component has partial information about the state of the system but there is no centralization of information, i.e., no agent knows the information available to all other agents. In many decentralized systems, all components/agents have a common objective: optimize the performance with respect to a system-wide objective (e.g., probability of correct detection with minimum energy consumption in sensor, surveillance, and UAV networks, congestion avoidance in transportation and telecommunication networks, throughput in MANETs and telecommunication networks, etc.). The agents can coordinate their activities to achieve their objective.

The decentralization of information makes the design of decentralized systems drastically different from the design of centralized systems. The results of standard Markov decision theory [1] are only applicable to centralized systems and cannot be used directly to obtain optimal solutions of decentralized systems. To the best of our knowledge, there is only one known methodology for obtaining appropriate information states for decentralized problems—Witsenhausen's standard form [2]. However, the standard form is applicable to only finite horizon problems; we are interested in solution methodology that can be used for both finite and infinite horizon problems.

The main contributions of this paper are the following. We obtain sequential decomposition of two general models for decentralized systems for both finite and infinite horizon problems. This decomposition is based on identifying common knowledge between the information structures of the different agents of the system. We further identify classes of information structures for which our sequential decomposition is similar to the sequential decomposition of POMDPs with finite state and action spaces. Such problems can be solved efficiently using the numerical techniques for POMDPs.

The paper is organized as follows. In Section II, we present two models of decentralized systems and formulate finite and infinite horizon problems for them. In Section III, we provide sequential decomposition for these models. In Section IV we illustrate the key features of our analysis using a simple multiaccess broadcast system as an example. In Section V, we identify conditions under which the sequential decomposition is tractable. We conclude in Section VI.

978-1-4244-2926-4/08/$25.00 ©2008 IEEE

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.

II. Problem Formulation

In this section we consider two models of decentralized systems and formulate the problem of optimal design of thse models for finite and infinite horizon.

A. Model A

Consider a discrete time systems that consists of a plant and n agents/controllers. Let $X_t \in \mathcal{X}$ denote the state of the plant at time t, and U_t^k denote the control action of agent k, $k = 1, \dots, n$, at time t. The system evolves as follows

$$X_{t+1} = f_t(X_t, U_t^1, \dots, U_t^n, W_t), \tag{1}$$

where $f_t(\cdot)$ is the *plant function* and $W_t \in \mathcal{W}$ denotes the process noise.

All agents receive a common message and a private message. The common message Y_t takes values in $\mathcal{Y}$ and is generated according to:

$$Y_t = c_t(X_t, U_{t-1}^1, U_{t-1}^2, \dots, U_{t-1}^n, Q_t), \tag{2}$$

where $c_t(\cdot)$ is the *observation channel* and Q_t denotes the observation noise.

The private message to agent k, which is denoted by Z_t^k and takes values in $\mathcal{Z}^k$, is generated according to:

$$Z_t^1 = h_t^1(X_t, N_t^1), \tag{3a}$$

$$Z_t^2 = h_t^2(X_t, U_t^1, N_t^2) \tag{3b}$$

$$\dots = \dots$$

$$Z_t^n = h_t^n(X_t, U_t^1, U_t^2, \dots, U_t^{n-1}, N_t^n) \tag{3c}$$

where h_t^k is the *observation channel of agent* k and N_t^k denotes the observation noise of agent k.

Each agent has unlimited memory to store the common messages and has either perfect recall or limited memory to store the private messages. We model this by assuming that at any time t, agent k knows $Y_1, \dots, Y_t$ and its memory contents M_{t-1}^k, where M_{t-1}^k takes values in $\mathcal{M}_t^k$. If $\mathcal{M}_t^k$ equals $\mathcal{Z}_1^k \times \cdots \mathcal{Z}_{t-1}^k \times \mathcal{U}_1^k \times \cdots \times \mathcal{U}_{t-1}^k$ then agent k has perfect recall.

At each time time t, after agent k gets his private observation Z_t^k, it generates a control action U_t^k and updates its states to M_t^k according to

$$U_t^k = g_t^k(Y_1, \dots, Y_t, Z_t^k, M_{t-1}^k), \tag{4}$$

$$M_t^k = l_t^k(Y_1, \dots, Y_t, Z_t^k, M_{t-1}^k), \tag{5}$$

respectively. The functions g_t^k and l_t^k are the *control law* and the *memory update rule/function* of agent k. Let $\mathcal{Y}^t$ denote $\mathcal{Y} \times \cdots \times \mathcal{Y}$ (t-times), $\mathscr{G}_t^k$ denote the family of functions from $\mathcal{Y}^t \times \mathcal{Z}_t^k \times \mathcal{M}_{t-1}^k$ to $\mathcal{U}_t^k$ and $\mathscr{L}_t^k$ denote the family of functions from $\mathcal{Y}^t \times \mathcal{Z}_t^k \times \mathcal{M}_{t-1}^k$ to $\mathcal{M}_t^k$.

At each time an instantaneous cost $\rho_t(X_t, U_t^1, \dots, U_t^k)$ is incurred. We assume that the initial state X_1 of the plant is a random variable with PDF (probability density function) P_{X_1}. We assume that $\{W_1, t = 1, \dots\}$, $\{Q_1, t = 1, \dots\}$, and $\{N_1^k, t = 1, \dots\}$, $k = 1, \dots, n$, are mutually independent sequence of independent random variables that are also independent of X_1. The PDFs of W_t, Q_t, and N_t^k are given by P_{W_t}, P_{Q_t} and $P_{N_t^k}$, respectively. The variables X_1, $\{W_1, t = 1, \dots\}$, $\{Q_1, t = 1, \dots\}$, and $\{N_1^k, t = 1, \dots\}$, $k = 1, \dots, n$ are called *primitive random variables*.

We are interested in two optimization problems, one for finite horizon and the other for infinite horizon.

1) Finite horizon case: Consider model A that operates for a finite horizon T. The choice of control laws and memory update rules for all agents for the entire horizon is called a *design* or a *strategy*. We denote a design by $(\boldsymbol{G}, \boldsymbol{L})$, where $\boldsymbol{G} := (G^1, G^2, \dots, G^n)$, $G^k := (g_1^k, g_2^k, \dots, g_T^k)$, $k = 1, \dots, n$, and $\boldsymbol{L} := (L^1, L^2, \dots, L^n)$, $L^k := (l_1^k, l_2^k, \dots, l_T^k)$. The peformance of a design is quantified by the expected total cost under that design, which is given by

$$J_T(\boldsymbol{G}, \boldsymbol{L}) := \mathbb{E}\left\{ \sum_{t=1}^{T} \rho_t(X_t, U_t^1, \dots, U_t^n) \,\middle|\, \boldsymbol{G}, \boldsymbol{L} \right\} \tag{6}$$

We are interested in the following optimization problem.

Problem 1 (The finite horizon problem): Given a horizon T and for each time $t = 1, \dots, T$, given the plant function f_t, the observation functions c_t and h_t^k, $k = 1, \dots, n$, the cost function ρ_t, and the statistics P_{X_1}, P_{W_t}, P_{Q_t} and $P_{N_t^k}$ of the primitive random variables, determine a design $(\boldsymbol{G}^*, \boldsymbol{L}^*)$ that is optimal with respect to the performance criterion of (6), i.e.,

$$J_T(\boldsymbol{G}^*, \boldsymbol{L}^*) = J_T^* := \inf_{\boldsymbol{G}, \boldsymbol{L} \in (\mathscr{G}^T, \mathscr{L}^T)} J_T(\boldsymbol{G}, \boldsymbol{L}) \tag{7}$$

where $\mathscr{G}^T := (\mathscr{G}_1^1 \times \mathscr{G}_2^1 \times \dots \mathscr{G}_T^1) \times \cdots \times (\mathscr{G}_1^n \times \mathscr{G}_2^n \times \dots \mathscr{G}_T^n)$ and $\mathscr{L}^T := (\mathscr{L}_1^1 \times \mathscr{L}_2^1 \times \dots \mathscr{L}_T^1) \times \cdots \times (\mathscr{L}_1^n \times \mathscr{L}_2^n \times \dots \mathscr{L}_T^n)$.

2) Infinite horizon case: Consider Model A that operates for an infinite horizon $T \to \infty$. We assume that the system is time homogeneous, that is

(i) all system variables take values in time-invariant spaces; so $\mathcal{X}_t$, $\mathcal{Y}_t$, $\mathcal{Z}_t^k$, $\mathcal{M}_t^k$, $\mathcal{U}_t^k$, $\mathcal{W}_k$, $\mathcal{Q}_t$, and $\mathcal{N}_t^k$ do not depend on t and can be written as $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}^k$, $\mathcal{M}^k$, $\mathcal{U}^k$, $\mathcal{W}_k$, $\mathcal{Q}$, and $\mathcal{N}^k$, respectively.

(ii) the plant function f_t, the observation functions c_t and h_t^k, and the instantaneous cost function ρ_t do not depend on t and can be written as f, c, h^k, and ρ respectively.

(iii) The statistics of the primitive random variables are independent of time. Thus, P_{W_t}, P_{Q_t}, and $P_{N_t^k}$ do not depend on t and can be written as P_W, P_Q, and P_{N^k}.

The collection of control laws and memory update rules for all agents for all time is called a design or a strategy and denoted by $(\boldsymbol{G}, \boldsymbol{L})$. The performance of a design is quantified by the expected discounted cost over an infinite horizon under that design; this cost is given by

$$J^\beta(\boldsymbol{G}, \boldsymbol{L}) := \mathbb{E}\left\{ \sum_{t=1}^{\infty} \beta^{t-1} \rho(X_t, U_t^1, \dots, U_t^n) \,\middle|\, \boldsymbol{G}, \boldsymbol{L} \right\} \tag{8}$$

where β is the *discount factor* which takes values in the interval $(0, 1)$. We are interested in the following optimization problem.

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.

Problem 2 (The infinite horizon problem): Given the plant function f, the observation functions c and h^k, the cost function ρ and the statistics P_{X_1}, P_W, P_Q, P_{N^k}, $k = 1, \dots, n$, of the primitive random variables, determine a design $(\boldsymbol{G}^*, \boldsymbol{L}^*)$ that is optimal with respect to the performance criterion of (8), i.e.,

$$J^\beta(\boldsymbol{G}^*, \boldsymbol{L}^*) = J^{\beta,*} := \inf_{\boldsymbol{G},\boldsymbol{L} \in (\bar{\mathscr{G}}, \bar{\mathscr{L}})} J^\beta(\boldsymbol{G}, \boldsymbol{L}) \tag{9}$$

where $\bar{\mathscr{G}} := (\mathscr{G}^1 \times \mathscr{G}^1 \times \dots) \times \dots \times (\mathscr{G}^n \times \mathscr{G}^n \times \dots)$ and $\bar{\mathscr{L}} := (\mathscr{L}^1 \times \mathscr{L}^1 \times \dots) \times \dots \times (\mathscr{L}^n \times \mathscr{L}^n \times \dots)$.

B. Model B

In model B we assume that the agents do not receive a common message. The rest of the model is same as model A. The plant update is given by (1) and each agent receives a private message according to (3). The control law and memory update rules of each agent do not depend on common messages and are given by

$$U_t^k = g_t^k(Z_t^k, M_{t-1}^k), \tag{10}$$

$$M_t^k = l_t^k(Z_t^k, M_{t-1}^k). \tag{11}$$

The instantaneous cost and statistics of the noise are the same as in model A.

For model B, we are interested in two optimization problems, one for finite horizon, and one for infinite horizon. These problems are similar those for model A.

C. Salient features of the models

Models A and B capture the salient features of decentralized systems that arise in different engineering applications. These features include

1) In decentralized systems agents have access to different information. Nevertheless, there may be certain information about the state of the system that is commonly known. Model A captures this situation.
2) The agents do not know all the observations and the control actions of other agents.
3) The actions taken by one agent affect the observations of other agents.
4) The agents are coupled through the plant dynamics and the common objective.

Hence a study of these models is useful for a large class of decentralized problems.

These models are also general enough to encompass other special classes of decentralized systems that have been considered in the literature. As an example, we show that the delayed sharing pattern is a special case of model A.

D. Delayed sharing pattern as an instance of model A

The delayed sharing pattern considered in [3], [4] consists of a plant and n agents/controllers.[1] The state of the plant is denoted by X_t and the control actions of agent k, $k = 1, \dots, n$, is denoted by U_t^k. The plant evolves according to

$$X_{t+1} = f_t(X_t, U_t^1, \dots, U_t^n, W_t), \tag{12}$$

Each agent receives two messages at time t: a common message consisting of private observations and control actions of all the agents at time $t-d$; and a private message consisting of noisy observations of the the state of the plant. More precisely, the private messages Z_t^k are given by

$$Z_t^k = h_t^k(X_t, N_t^k), \quad k = 1, \dots, n \tag{13}$$

where N_t^k denotes the observation noise. The common message Y_t is given by

$$Y_t = \big((Z_{t-d}^1, U_{t-d}^1), \dots, (Z_{t-d}^n, U_{t-d}^n)\big) \tag{14}$$

Each agent has perfect recall and generates its control actions according to

$$U_t^k = g_t^k(Z_1^k, \dots, Z_t^k, U_1^k, \dots, U_{t-1}^k, Y_1, \dots, Y_t)$$

which is equivalent to

$$U_t^k = g_t^k(Z_{t-d+1}^k, \dots, Z_t^k, U_{t-d+1}^k, \dots, U_{t-1}^k, Y_1, \dots, Y_t). \tag{15}$$

Thus, we can think of $(Z_{t-d+1}^k, \dots, Z_{t-1}^k, U_{t-d+1}^k, \dots, U_{t-1}^k)$ as the memory of agent k. At each time an instantaneous cost $\rho_t(X_t, U_t^1, \dots, U_t^n)$ is incurred.

The delayed sharing pattern has some similarties with model A but is not exactly the same. However, it can be considered as an instance of model A by a suitable expansion of the state space. Given a system with delayed sharing pattern, define

$$\bar{N}_t := (N_t^1, \dots, N_t^n) \tag{16}$$

$$\bar{U}_t := (U_t^1, \dots, U_t^n) \tag{17}$$

$$\bar{X}_t := (X_{t-d}, \dots, X_t, \bar{U}_{t-d}, \dots, \bar{U}_{t-1}, \bar{N}_{t-d}) \tag{18}$$

$$\bar{W}_t := (W_t, \bar{N}_{t-d+1}) \tag{19}$$

Then,

$$\bar{X}_{t+1} = (X_{t-d+1}, \dots, X_{t+1}, \bar{U}_{t-d+1}, \dots, \bar{U}_t, \bar{N}_{t-d+1})$$

The component X_{t+1} of $\bar{X}_{t+1}$ is generated according to (12); the component $\bar{U}_t$ is a controlled input, the component $\bar{N}_{t-d+1}$ is a component of $\bar{W}_t$, and all other components of $\bar{X}_{t+1}$ are part of $\bar{X}_t$. Thus, we can determine a function $\bar{f}_t$ such that

$$\bar{X}_{t+1} = \bar{f}_t(\bar{X}_t, \underbrace{U_t^1, \dots, U_t^n}_{\bar{U}_t}, \bar{W}_t) \tag{20}$$

The common observations of all the agnets can be written as

$$\begin{aligned} Y_t &:= (Z_{t-d}^1, \dots, Z_{t-d}^n, U_{t-d}^1, \dots, U_{t-d}^n) \\ &= \big(h_{t-d}^1(X_{t-d}, N_{t-d}^1), \dots, h_{t-d}^n(X_{t-d}, N_{t-d}^n), \\ &\qquad\qquad U_{t-d}^1, \dots, U_{t-d}^n\big) \\ &=: \bar{c}(\bar{X}_t) \end{aligned} \tag{21}$$

The private messages given by (13) can be written as

$$Z_t^k = h_t^k(X_t, N_t^k) =: \bar{h}_t^k(\bar{X}_t, N_t^k) \tag{22}$$

[1] We use a slightly different notation than [4].

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.

which is a special case of (3). Furthermore, the control laws of (15) is a special case of (4). The instantaneous cost can be written as $\rho_t(X_t, U_t^1, \dots, U_t^n) =: \bar{\rho}_t(\bar{X}_t, U_t^1, \dots, U_t^n)$. Thus, the system with the above regrouping and $\bar{X}_t$ as the state of the system is an instance of model A. This model is equivalent to the delayed sharing pattern, so the delayed sharing model is also an instance of model A.

E. Solution philosophy of centralized problems

In this paper, we show that when viewed appropriately, decentralized problems of model A and B can be considered as partially observed centralized systems. In order to explain our results, we first briefly present the main ideas for partially observed centralized systems.

A common model of centralized control problem is the partially observable Markov decision process (POMDP) which consists of a plant and a controller. The state of the plant is denoted by X_t which takes values in $\mathcal{X}$. The state evolves as follows:

$$X_{t+1} = f_t(X_t, U_t, W_t) \tag{23}$$

where U_t is the control action taking values in $\mathcal{U}$, and $\{W_t, t = 1, \dots, T\}$ is an independent noise process.

The state of the plant is imperfectly observed by a controller. The controller's observation at time t is denoted by Y_t and takes values in $\mathcal{Y}$. These observations are generated according to

$$Y_t = h_t(X_t N_t) \tag{24}$$

The controller has perfect recall, that is, it remembers all its past actions and observations. At each time t, the controller generates a control action according to

$$U_t = g_t(Y^t, U^{t-1}) \tag{25}$$

where $Y^t = Y_1, \dots, Y_t$ and $U^{t-1} = U_1, \dots, U_{t-1}$.

A *control policy* or a *design* for a finite time horizon T is given by a collection of functions :

$$g_t : \mathcal{Y}^t \times \mathcal{U}^{t-1} \to \mathcal{U} \tag{26}$$

At each instant, the system incurs a cost $\rho_t(X_t, U_t)$ that depends on the state and the action taken at that time. The initial state X_1 of the plant is a random variable with PDF P_{X_1}. The plant disturbance $\{W_1, t = 1, \dots\}$ and the observation noise $\{N_1, t = 1, \dots\}$ are independent processes which are also independent of X_1. The PDF of W_t and N_t are given by P_{W_t} and P_{N_t} respectively.

We are interested in two optimization problem, one for finite horizon and one for infinite horizon. For the finite horizon problem, the objective is to determine a design $G : = (g_1, \dots, g_T)$ to minimize a total expected cost given by

$$J_T(G) := \mathbb{E}\left\{ \sum_{t=1}^{T} \rho_t(X_t, U_t) \,\middle|\, G \right\}$$

For the infinite horizon problem, the system is assumed to be time homogeneous and the objective is to determine a design $G := (g_1, g_2, \dots)$ to minimize (or minimize within an ε) a total expected discounted cost given by

$$J^\beta(G) := \mathbb{E}\left\{ \sum_{t=1}^{\infty} \beta^{t-1} \rho(X_t, U_t) \,\middle|\, G \right\}$$

Choosing an optimal design for both finite and infinite horizons in a functional optimization problem. Markov decision theory provides a systematic methodology for determining optimal designs for POMDPs. For centralized problems, this methodology is called dynamic programming; in general, it is called sequential decomposition.

For finite horizon problem, Markov decision theory provides a decomposition of the one-shot optimization problem into a sequence of smaller problems; each step of this decomposition solves a series of parametric optimization problems. The notion of information state is key to obtaining such a sequential decomposition. For POMDPs, an information state is the controller's belief on the state of the system given its past observations and actions, i.e.,

$$\pi_t = \Pr\left\{ X_t \,\middle|\, Y^t, U^{t-1} \right\} \tag{27}$$

The information state π_{t+1} at time $t+1$ is a function of the information state π_t at time t, the control action U_t at time t and the observation Y_{t+1} at time $t+1$; so, it can be written as

$$\pi_{t+1} = F_t(\pi_t, U_t, Y_{t+1}). \tag{28}$$

See [1] for the exact functional form of F_t.

Markov decision theory shows that there is no loss of optimality in restricting attention to control laws of the form

$$U_t = g_t(\pi_t). \tag{29}$$

Further, optimal control laws can be determined by the solution of the following optimality equations, called the dynamic program

$$V_{T+1}(\pi) = 0 \tag{30a}$$

$$V_t(\pi) = \inf_{u_t \in \mathcal{U}} \big[\mathbb{E}\{\pi_t(X_t, U_t) + V_{t+1}(\pi_{t+1}) \,|\, \pi_t = \pi, U_t = u_t\} \big] \tag{30b}$$

The $\arg\inf$ at each step determines the corresponding optimal control action.

For the infinite horizon problems, Markov decision theory shows that for a time homogeneous system there is no loss of optimality in restricting attention to time-invariant designs of the form (29), i.e., designs where $g_t = g$ for all t. The optimal time-invariant design can be determined from the unique uniformly bounded fixed point of the following functional equation

$$V(\pi) = \inf_{u \in \mathcal{U}} \big[\mathbb{E}\{\rho(X, U) + \beta V(F(\pi, U, Y)) \,|\, \pi, U = u\} \big] \tag{31}$$

where $F(\cdot)$ is the time invariant version of F_t from (28), $Y = h(f(X, U, W), N)$ and the expectation is with respect to the measure on (X, W, N) given by $\pi \cdot P_W \cdot P_N$.

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.

The above described sequential decomposition provides a systematic way to search for an optimal design efficiently. For finite horizon problems where the system variables are finite valued, a brute force search for an optimal design requires computing the performance of approximately $|\mathcal{U}|^{(|\mathcal{Y}|\times|\mathcal{U}|)^T}$ designs. The dynamic programming equations can be solved by exploiting the piecewise linearity and convexity of the V_t functions [5]. In the worst case, this solution requires approximately $|\mathcal{U}|^{|\mathcal{Y}|^T}$ computations. However, for specific instances, the solution can be found in computations that are polynomial in T (see [6]).

For infinite horizon POMDPs, even when the system variables are finite valued, optimal designs cannot be searched in a brute force manner since there are infinitely many designs. The functional equations of (31) can be efficiently approximated using randomized algorithms that discretize the belief space. It is shown in [8] that the worst case complexity of solving discounted cost POMDPs with finite state and action space is polynomial in $|\mathcal{X}|$ and $|\mathcal{U}|$. There are other results which exploit the special structure of POMDPs arising in specific application domains to solve finite and infinite POMDPs more efficiently.

The advantages of sequential decomposition for centralized problems motivates obtaining the sequential decomposition of decentralized problems.

III. Sequential Decomposition

For a centralized stochastic control problem, the sequential decomposition of provides computational advantages in finding an optimal policy for both finite and infinite horizon problems (see [7], [8]). In certain cases the optimality equations of the sequential decomposition can be used to identify qualitative properties of optimal control laws. For example, for sequential hypothesis testing problem [9] the optimality equations are used to prove that an optimal decision rule is of a "threshold type"; for centralized LQG (linear quadratic and Gaussian) problems the optimality equations can be used to prove that optimal control laws are affine. Such qualitative properties significantly simplify the search for optimal designs.

In this section, we present sequential decomposition for both finite and infinite horizon multi-agent problems described in models A and B. As in the case of a centralized problem, the sequential decomposition proceeds backward in time — it first finds optimal strategies for all realizations of an information state at the last time step and then for the previous time step and so on. The decentralized nature of our models implies that each step of the decomposition is a functional optimization problem unlike the parametric optimization obtained in centralized problems.

A. Model A : Finite Horizon Case

Consider the finite horizon case of model A. At each time instant, the optimal strategy choice of an agent depends on the strategies of other agents. If the strategies of all agents are to be obtained sequentially, the agents should be able to agree on the choice of strategies of all agents. In other words, every agent must be able to carry out the same sequential decomposition. Hence, the sequential decomposition must be based on information that is available to all agents. At each time step t, this "common information" is the collection of all past control and memory-update functions and the sequence of past common messages $(Y_1, \dots, Y_t)$. Equivalently, we can assume that the sequential decomposition is carried out by a fictitious agent who has access to the common information. We call this agent the *common agent*.

The derivation of the sequential decomposition proceeds in three stages:

1) Formulate the problem as a centralized stochastic control problem from the point of view of the common agent who knows the common information.
2) Show that the centralized stochastic problem is a POMDP.
3) Identify an information state for the resulting POMDP and use this information state to obtain a sequential decomposition.

Below, we elaborate on each of these stages.

Stage 1: At each time step, the common agent knows all the common messages received so far. The common agent then selects functions that map each agent's private information to its actions. That is, for agent k at time t, the common agent selects the following *partial functions* $\hat{g}_t^k$ and $\hat{l}_t^k$,

$$\hat{g}_t^k : \mathcal{Z}_t^k \times \mathcal{M}_{t-1}^k \to \mathcal{U}_t^k$$
$$\hat{l}_t^k : \mathcal{Z}_t^k \times \mathcal{M}_{t-1}^k \to \mathcal{M}_t^k$$

based on $(Y_1, \dots, Y_t)$. Once these functions are selected, each agent simply receives its private message and uses the selected functions to determine its control action and update its memory according to

$$U_t^k = \hat{g}_t^k(Z_t^k, M_{t-1}^k),$$
$$M_t^k = \hat{l}_t^k(Z_t^k, M_{t-1}^k).$$

The system then incurs a cost $\rho_t(X_t, U_t^1, \dots, U_t^n)$. Viewed like this, the problem is now a centralized problem with the common agent as the only decision maker.

Stage 2: Define the following random vectors:

Definition 1:

$$S_t^0 := (X_t, M_{t-1}^1, M_{t-1}^2, \dots, M_{t-1}^n, U_{t-1}^1, U_{t-1}^2, \dots, U_{t-1}^n), \tag{32a}$$

and for $k = 1, \dots, n+1$,

$$\begin{aligned} S_t^k := (&X_t, M_t^1, M_t^2, \dots, M_t^{k-1}, \\ &M_{t-1}^k, M_{t-1}^{k+1}, \dots, M_{t-1}^n, U_t^1, U_t^2, \dots, U_t^{k-1}). \end{aligned} \tag{32b}$$

These random variables evolve with time as follows:

Claim 1: For each time t there exists a function $\hat{f}_t^0$ such that

$$S_t^1 = \hat{f}_t^0(S_t^0), \tag{33a}$$

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.

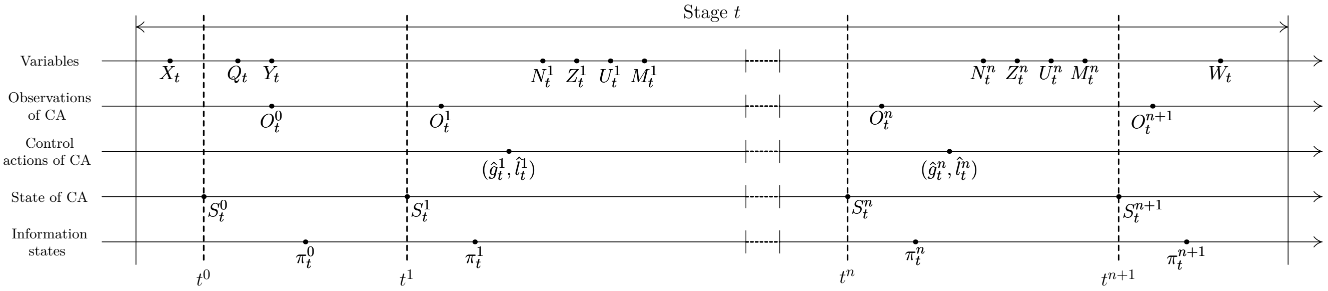

Fig. 1. Sequential ordering of the different variables in the system. In the labels CA is an abbreviation for common agent.

and for each time t, and each $k = 1, \ldots, n$, there exist function $\hat{f}_t^k$ such that

$$S_t^{k+1} = \hat{f}_t^k(S_t^k, \hat{g}_t^k, \hat{l}_t^k, N_t^k), \tag{33b}$$

and for each t, there exists a function $\hat{f}_t^{n+1}$ such that

$$S_{t+1}^0 = \hat{f}_t^{n+1}(S_t^{n+1}, W_t). \tag{33c}$$

Proof: This followed immediately from the functional relation between the different components of S_t^k. A detailed proof is presented in [10]. ■

Using these random variables, the common message can be written as

$$Y_t = c_t(X_t, U_{t-1}^1, \ldots, U_{t-1}^n, Q_t) =: \hat{c}_t(S_t^0, Q_t). \tag{34}$$

The update equations (33) of Claim 1 and the observation equation (34) imply that, from the point of view of the common agent, the variables S_t^k, $k = 1, \ldots, n+1$, $t = 1, \ldots, T$, are *states sufficient for input-output mapping of the system* . Thus, from the point of view of the common agent, the system is a partially observed system. To see this, we refine the notion of time. Consider that each agent makes its observation and takes an action at a different time step. Thus, in the time interval from t to $t+1$, we consider $n+2$ smaller time steps marked by: $t^0, t^1, \ldots, t^{n+1}$ (see Figure 1). All agents get the common message at t^0. Just after t^1, the common agent selects the partial functions for agent 1; then, the first agent gets his private observation, uses the selected partial functions to take its control action and update its memory. Similarly, for the k-th agent, just after t^k, the common agent selects the partial functions for agent k; then, the k-th agent gets his private observation, uses the selected partial functions to take its control action and update its memory.

We can assume that after each time step t^k, $k = 0, \ldots, n+1$ the common agent makes observations, which are denoted by O_t^k, where the observations at time t^k, $k = 1, \ldots, n+1$ are null. The common agent's observation functions can be written as:

$$O_t^0 = Y_t = c_t(X_t, U_{t-1}^1, U_{t-1}^2, \ldots, U_{t-1}^n, Q_t) =: \hat{h}_t^0(S_t^0, Q_t) \tag{35a}$$

and for $k = 1, \ldots, n+1$,

$$O_t^k = 0 =: \hat{h}_t^k(S_t^k) \tag{35b}$$

Thus, $\{S_t^k,\ k = 0, \ldots, n+1,\ t = 1, \ldots, T\}$ is a partially observed controlled Markov process and the common agents observations depend only on the current state of the controlled Markov chain.

Furthermore, the instantaneous cost can be written as follows:

Claim 2: $\rho_t(X_t, U_t^1, \ldots, U_t^n) = \hat{\rho}_t(S_t^{n+1})$.

Proof: This follows from the fact that X_t and $U_t^1, \ldots, U_t^n$ are components of S_t^{n+1}. ■

The total cost of the system can be written as:

$$J_T = \mathbb{E}\left\{ \sum_{t=1}^{T} \hat{\rho}_t(S_t^{n+1}) \right\}. \tag{36}$$

Thus from the common agent's perspective the system is a standard POMDP.

Stage 3: When the decentralized problem of model A is viewed from the perspective of the fictitious common agent, it becomes a centralized problem with an enlarged state and action spaces. In particular, the state space incorporates not only the state of the plant but also the state of the private information of each actual agent while the action space is no longer the space of control action taken but the space of functions used to map private information to control actions. Since the model A assumes perfect memory for common information, the common agent has perfect recall. Hence, from the perspective of the common agent, the problem is a POMDP. One can now use the standard POMDP results to come up with an information state and the sequential decomposition. Next we describe the information state and the sequential decomposition for this POMDP.

Information State: Let $\mathbf{O}_t^k$ denote all the observations received by common agent till time t^k. That is,

$$\begin{aligned} \mathbf{O}_t^k &= (O_1^0, O_1^1, .., O_1^{n+1}, O_2^0, .., O_2^{n+1}, .., O_t^0, ..O_t^k) \\ &= (Y_1, \ldots, Y_t). \end{aligned} \tag{37}$$

Similarly let $\mathbf{P}_t^k$ be the set of all partial functions selected by the controller before time t^k. That is,

$$\begin{aligned} \mathbf{P}_t^k = (&\hat{g}_1^1, \hat{l}_1^1, \ldots, \hat{g}_1^n, \hat{l}_1^n, \hat{g}_2^1, \hat{l}_2^1, \ldots, \hat{g}_2^n, \hat{l}_2^n, \ldots, \\ &\hat{g}_t^1, \hat{l}_t^1, \ldots, \hat{g}_t^{k-1}, \hat{l}_t^{k-1}) \end{aligned} \tag{38}$$

Then, the information state of the common agent is his belief on the current state given the past observations and

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.

actions and is given by

$$\pi_t^k = \Pr\left\{ S_t^k \;\middle|\; \mathbf{O}_t^k, \mathbf{P}_t^k \right\} \tag{39}$$

Since the common agent's problem is now a POMDP, it follows that the next belief depends only on the current belief, the current action and the next observation of the common agent. Similar to (28), we can write

$$\pi_t^1 = F_t^1(\pi_t^0) \tag{40a}$$

and for $k = 1, \ldots, n$,

$$\pi_t^{k+1} = F_t^k(\pi_t^k, \hat{g}_t^k, \hat{l}_t^k) \tag{40b}$$

and

$$\pi_{t+1}^0 = F_{t+1}^0(\pi_t^{n+1}, Y_t) \tag{40c}$$

We refer the reader to [10] for the exact functional form of F_t^k, $k = 0, \ldots, n+1$. Using these information states, the optimality equations of the problem at the common agent are given by the following.

Theorem 1: Let $\tilde{\pi}_t^k$ denote a PMF (probability mass function) on S_t^k, $k = 0, \ldots, n+1$, and $t = 1, \ldots, T$. Define the following functions:

$$V_{T+1}^0(\tilde{\pi}_{T+1}^0) = 0 \tag{41a}$$

and for $t = 1, 2, \ldots, T$ and $k = 1, 2, \ldots, n$,

$$\begin{aligned} &V_t^{n+1}(\tilde{\pi}_t^{n+1}) = \\ &\qquad \mathbb{E}\left\{ \hat{\rho}_t(S_t^{n+1}) + V_{t+1}^0(\pi_{t+1}^0) \;\middle|\; \pi_t^{n+1} = \tilde{\pi}_t^{n+1} \right\} \end{aligned} \tag{41b}$$

$$V_t^k(\tilde{\pi}_t^k) = \inf_{\hat{g}_t^k, \hat{l}_t^k} \left[\mathbb{E}\left\{ V_t^{k+1}(\pi_t^{k+1}) \;\middle|\; \pi_t^k = \tilde{\pi}_t^k, \hat{g}_t^k, \hat{l}_t^k \right\} \right] \tag{41c}$$

$$V_t^0(\tilde{\pi}_t^0) = \mathbb{E}\left\{ V_t^1(\pi_t^1) \;\middle|\; \pi_t^0 = \tilde{\pi}_t^0 \right\} \tag{41d}$$

The $\arg\inf$ at each step in (41c) determines the optimal partial functions.

The above sequential decomposition is similar to the one in POMDPs. Observe that the infimum at each step is over the set of partial functions and not parameters.

Since the information states in the above decomposition depends on common information which is available to all agents, the sequential decomposition can be carried out by all agents. If all agents use identical rules for breaking ties, they will come up with identical solutions for the optimality equations. An agent can then implement its own partial functions to select its control action and use the partial functions of all agents to track the evolution of information state.

B. Sequential Decomposition for Model B

As mentioned before, model B is a special case of model A where there are no common messages. Therefore, one can use the same arguments as for model A to arrive at a sequential decomposition with the only modification being the removal of the common messages. Specifically, the three stages in the argument for model A become modified as follows:

Stage 1: At each time step, the common agent has to select each agent's control law and memory update rule, that is, functions that map each agent's information to its actions.

$$\begin{aligned} g_t^k &: \mathcal{Z}_t^k \times \mathcal{M}_{t-1}^k \to \mathcal{U}_t^k \\ l_t^k &: \mathcal{Z}_t^k \times \mathcal{M}_{t-1}^k \to \mathcal{M}_t^k \end{aligned}$$

Stage 2: The definition of the states S_t^k and the refined notion of time are same as in model A. However, the common agent observations are null at all times. The system is then an unobserved controlled Markov decision process.

Stage 3: The information state is now given as:

$$\pi_t^k = \Pr\left\{ S_t^k \;\middle|\; \mathbf{P}_t^k \right\} \tag{42}$$

where

$$\begin{aligned} \mathbf{P}_t^k = (g_1^1, l_1^1, \ldots, g_1^n, l_1^n, g_2^1, l_2^1, \ldots, g_2^n, l_2^n, \ldots, \\ g_t^1, l_t^1, \ldots, g_t^{k-1}, l_t^{k-1}) \end{aligned} \tag{43}$$

The optimality equations for model B are given by the following.

Theorem 2: Let $\tilde{\pi}_t^k$ denote a PMF (probability mass function) on S_t^k, $k = 0, \ldots, n+1$, and $t = 1, \ldots, T$. Define the following functions:

$$V_{T+1}^0(\tilde{\pi}_{T+1}^0) = 0 \tag{44a}$$

For $t = 1, 2, \ldots, T$ and $k = 1, 2, \ldots, n$

$$\begin{aligned} &V_t^{n+1}(\tilde{\pi}_t^{n+1}) = \\ &\qquad \mathbb{E}\left\{ \rho_t(S_t^{n+1}) + V_{t+1}^0(\pi_{t+1}^0) \;\middle|\; \pi_t^{n+1} = \tilde{\pi}_t^{n+1} \right\} \end{aligned} \tag{44b}$$

$$V_t^k(\tilde{\pi}_t^k) = \inf_{g_t^k, l_t^k} \left[\mathbb{E}\left\{ V_t^{k+1}(\pi_t^{k+1}) \;\middle|\; \pi_t^k = \tilde{\pi}_t^k, g_t^k, l_t^k \right\} \right] \tag{44c}$$

$$V_t^0(\tilde{\pi}_t^0) = \mathbb{E}\left\{ V_t^1(\pi_t^1) \;\middle|\; \pi_t^0 = \tilde{\pi}_t^0 \right\} \tag{44d}$$

The $\arg\inf$ at each step in (44c) determines the optimal partial functions.

C. Models A and B: Infinite Horizon Cases

The common agent's problem described above for the finite horizon cases can also be extended to the infinite horizon cases. Essentially, from the common agent's perspective, the system is still a partially observed controlled Markov process but it now has a discounted cost over infinite horizon as the optimization criterion. Under the assumptions of time-homogeneity described in the problem formulation, the problem is equivalent to time-invariant infinite horizon POMDP. The optimality equations of Theorems 1 and 2 can be extended to the infinite horizon in the standard way. For model A, the infinite horizon designs are given by the following functional equations

$$V^0(\tilde{\pi}^0) = \mathbb{E}\left\{ V^1(\pi^1) \;\middle|\; \pi^0 = \tilde{\pi}^0 \right\},$$

for $k = 1, \ldots, n$

$$V^k(\tilde{\pi}^k) = \inf_{\hat{g}^k, \hat{l}^k} \left[\mathbb{E}\left\{ V^{k+1}(\pi^{k+1}) \;\middle|\; \pi^k = \tilde{\pi}^k, \hat{g}^k, \hat{l}^k \right\} \right],$$

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.

and

$$V^{n+1}(\tilde{\pi}^{n+1}) = \mathbb{E}\left\{ \hat{\rho}(S^{n+1}) + \beta V^0(\pi^0) \;\middle|\; \pi^{n+1} = \tilde{\pi}^{n+1} \right\}.$$

The information states π^k, $k = 0, \dots, n+1$, are related by a time-invariant version of (40). For model B, we can simply replace $(\hat{g}^k, \hat{l}^k)$ in the above equations with (g^k, l^k). The detailed derivation of these equations is shown in [10].

IV. An example — multiaccess broadcast system

A. Problem formulation

Consider the following two user multiaccess broadcast system where the users transmit over a shared channel. The system operates in discrete time. Both users have a buffer of size 1 and if they have a packet, they can transmit it at the beginning of the time slot. If only one user transmits, the packet is received successfully. If both users transmit simultaneously, a collision occurs and no packet is received. Packets arrival at each user is an independent Bernoulli process with probabilities of arrival p_1 and p_2, respectively.

At the end of each time step, both users receive a feedback describing the channel use in that time step, that is, both users know whether there was a successful transmission, a collision or that the channel was idle.

The objective of the problem is to design transmission protocols for both users to maximize the throughput of the system, that is, to maximize the probability of successful transmission in each time step.

Formally, the above model can be described as follows: for the i^{th} user, let $B^i_t \in \{0, 1\}$ be the state of the buffer at the beginning of the t^{th} time step, let $A^i_t \in \{0, 1\}$ be the number of packet arrivals during the t^{th} time step and let $U^i_t \in \{0, 1\}$ be the transmission decision made (0 representing no transmission and 1 representing a transmission). Let R_t be the channel feedback received at the end of the t^{th} time step. $R_t = 0$ represents idle channel, $R_t = i$, $(i = 1, 2)$ represents a successful transmission by the i^{th} user and $R_t = 3$ represents a collision. At each time step, each user decides its action based on the current state of its buffer and all the past channel feedback messages received. That is,

$$U^i_t = g^i_t(B^i_t, R_1, \dots, R_{t-1}). \tag{46}$$

The number of packets in the buffer of user i change as follows

a) $B^k_{t+1} = B^k_t$ if both users transmit, i.e., $U^k_t = U^{k'}_t = 1$, $k' \neq k$.
b) $B^k_{t+1} = A^k_t$ if only user k transmits, i.e., $U^k_t = 1, U^{k'}_t = 0$, $k' \neq k$.
c) $B^k_{t+1} = \min(1, B^k_t + A^k_t)$ if user k does not transmit, i.e., $U^k_t = 0$.

Also, the channel feedback R_t is a function only of the current actions $U^1_t$and U^2_t of the users. In particular,

$$R_t = \begin{cases} 0, & \text{if no user transmits, i.e., } U^1_t = U^2_t = 0, \\ 1, & \text{if only user 1 transmits, i.e., } U^1_t = 1 \text{ and } U^2_t = 0, \\ 2, & \text{if only user 2 transmits, i.e., } U^1_t = 0 \text{ and } U^2_t = 1, \\ 3, & \text{if both users transmit, i.e., } U^1_t = U^2_t = 1. \end{cases}$$

The reward earned in a time step is given as $r(U^1_t, U^2_t)$ where $r(a, b) = 1$ if $a \neq b$ and 0 otherwise. The overall objective is to maximize the total reward over a finite horizon T.

The above problem can be formulated as an instance of model A as follows :

1) The state of the system can be described as

$$X_t = (B^1_t, B^2_t) \tag{47}$$

Given the current state and the users' decisions, the state at next time step is determined by the independent arrival process, that is,

$$X_{t+1} = f_t(X_t, U^1_t, U^2_t, A^1_t, A^2_t) \tag{48}$$

2) R_{t-1} is the common message (Y_t) between users at time t. Clearly,

$$Y_t = R_{t-1} = c_t(U^1_{t-1}, U^2_{t-1}) \tag{49}$$

3) Additionally, each user observes the state of his own buffer. Thus B^i_t is the private message (Z^i_t) of i^{th} user.

$$Z^i_t = B^i_t = h^i_t(X_t) \tag{50}$$

4) Since the users don't recall past states of buffer, they have no memory $(M^i_t = \emptyset)$.
5) The instantaneous cost $\rho_t(X_t, U^1_t, U^2_t) = -r(U^1_t, U^2_t)$.

Thus, the multiaccess broadcast problem can be viewed as an instance of model A. We can use the results of Section III-A to obtain a sequential decomposition of the problem. This sequential decomposition is described in the next section.

B. Sequential decomposition of multiaccess broadcast

The common agent for this problem knows all the past channel feedback messages. The actions for the common agent are the partial functions that map the state of the user's buffer to a transmission decision. Since both the state of buffer and the action take values in {0,1}, the partial functions that the common agent selects are of the form

$$\hat{g}^k_t : \{0, 1\} \to \{0, 1\} \tag{51}$$

Following the arguments of sequential decomposition of Model A, we can define the states S^k_t as:

$$\begin{aligned} S^0_t &:= (B^1_t, B^2_t, U^1_{t-1}, U^2_{t-1}), \\ S^1_t &:= (B^1_t, B^2_t), \\ S^2_t &:= (B^1_t, B^2_t, U^1_t), \\ S^3_t &:= (B^1_t, B^2_t, U^1_t, U^2_t). \end{aligned} \tag{52}$$

It is straightforward to see that there exist deterministic functions $\hat{f}^k_t$, $k = 0, 1, 2$, such that

$$S^1_t = \hat{f}^0_t(S^0_t) \tag{53a}$$

$$S^2_t = \hat{f}^1_t(S^1_t, \hat{g}^1_t), \tag{53b}$$

$$S^3_t = \hat{f}^2_t(S^2_t, \hat{g}^2_t), \tag{53c}$$

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.

and for each t, there exists a function $\hat{f}_t^3$ such that

$$S_{t+1}^0 = \hat{f}_t^3(S_t^3, A_t^1, A_t^2). \tag{53d}$$

The common agent's observations are:

$$O_t^0 = Y_t = c_t(U_{t-1}^1, U_{t-1}^2) = \hat{h}_t^0(S_t^0) \tag{54a}$$

and for $k = 1, \dots, 3$,

$$O_t^k = \hat{h}_t^k(S_t^k) = 0; \tag{54b}$$

and the cost is

$$\rho_t(U_t^1, U_t^2) = \hat{\rho}_t(S_t^3) \tag{55}$$

We can now use equation (39) to write the information states at each instant in the refined time (t^0, t^1, t^2, t^3). A simple substitution gives the following information states

$$\pi_t^0 = \Pr\left\{ B_t^1, B_t^2, U_{t-1}^1, U_{t-1}^2 \mid R^{t-1} \right\} \tag{56a}$$
$$\pi_t^1 = \Pr\left\{ B_t^1, B_t^2 \mid R^{t-1} \right\} \tag{56b}$$
$$\pi_t^2 = \Pr\left\{ B_t^1, B_t^2, U_t^1 \mid R^{t-1} \right\} \tag{56c}$$
$$\pi_t^3 = \Pr\left\{ B_t^1, B_t^2, U_t^1, U_t^2 \mid R^{t-1} \right\} \tag{56d}$$

where $R^{t-1} = (R_1, \dots, R_{t-1})$. It can be easily verified that as in POMDP the information state at the next stage depends only on the current information state, the action and the observation of the common agent ([10]). The sequential decomposition of Theorem 1 can be written as follows for this problem

$$V_{T+1}^0(\tilde{\pi}_{T+1}^0) = 0 \tag{57a}$$

and for $t = 1, 2, \dots, T$ and $k = 1, 2$,

$$V_t^3(\tilde{\pi}_t^3) = \mathbb{E}\left\{ \hat{\rho}_t(S_t^3) + V_{t+1}^0(\pi_{t+1}^1) \mid \pi_t^3 = \tilde{\pi}_t^3 \right\} \tag{57b}$$
$$V_t^k(\tilde{\pi}_t^k) = \inf_{\hat{g}_t^k} \left[\mathbb{E}\left\{ V_t^{k+1}(\pi_t^{k+1}) \mid \pi_t^k = \tilde{\pi}_t^k, \hat{g}_t^k \right\} \right] \tag{57c}$$
$$V_t^0(\tilde{\pi}_t^0) = \mathbb{E}\left\{ V_t^1(\pi_t^1) \mid \pi_t^0 = \tilde{\pi}_t^0 \right\} \tag{57d}$$

In equations (57b) and (57d) we do not have to choose a control action. In (57c), we have to perform a minimization over a set of four possible choices of $\hat{g}_t^k$. Moreover, the state at each time S_t^k belongs to a finite set of small cardinality (at most 8) In principle, the four optimality equations can be combined by straightforward substitution to yield a single minimization at each step where the minimization is over all possible choices of $\hat{g}_t^1$ and $\hat{g}_t^2$ together.

Thus, we can transform the original decentralized problem into a POMDP by considering the problem from the common agent's perspective. Since each agents private information lies in a small finite set and the number of control actions is limited to two (transmit or not), the POMDP formulation involves only small state and action spaces. This allows using the available numerical techniques for solving moderately sized POMDPs to be applied for this decentralized problem. In the next section, we generalize these ideas to identify cases where the finite and infinite horizon problem for models A and B can be efficiently solved with the POMDP numerical methods.

V. Tractability of the sequential decomposition

In general optimality equations for POMDPs are hard to solve. However, when the state and action spaces are finite, efficient approximation techniques exist for both finite horizon [7] and infinite horizon [8] problems. This means that if the POMDPs corresponding to models A and B have finite state and action space, approximately optimal efficient solutions of these models can be computed.

Consider the finite horizon problem for model A when all system variables take values in a finite set. In this case, the states S_t^k, $k = 0, \dots, n+1$, sufficient for input-output mapping also take values in a finite set. Furthermore, the partial functions $\hat{g}_t^k$ and $\hat{l}_t^k$ are functions where the domain and the range are finite sets. Thus, there are finitely many possibilities for these partial functions. Hence, the POMDP formulated from the common agent's perspective has finite state and action spaces. Consequently, optimal solution for these POMDPs can be found using the computational techniques of [7]. Recall however that the conversion of the decentralized model A into a POMDP comes at the cost of the increase in the dimensionality of the state and action spaces. In particular, the action space is exponential in the sizes of the private messages and the memory of the agents. This means that optimal solutions for model A can be computed for a much smaller dimensionality of the state and action spaces (of the underlying decentralized problem) than those for POMDPs. Similar results and concerns hold for the finite horizon case of model B.

For infinite horizon problems, the system variables take values in time-invariant spaces. When all of these spaces are finite sets, the POMDP at the common agent has finite state and action spaces. This follows from arguments similar to those in the finite horizon case. Such POMDPs can be solved approximately using randomized algorithms [8]. The complexity of these algorithms in polynomial in the size of the state and the action spaces.

When the system variables in models A and B are continuous, the problem from the common agent's perspective is similar to a POMDP where the state space is continuous and the action space is the space of functions over continuous spaces. Computational algorithms for such POMDPs have not been considered in the literature, mainly because such POMDPs do not arise in centralized problems. The models usually considered for centralized POMDPs involve systems that are described by finite state Markov chains with finite action space and finite observation space.

Notice that decentralized problems can be identified as belonging to models A and B on the basis of their information structures. In addition, if the system variables take values in finite sets, we can obtain a sequential decomposition that can be solved efficiently.

VI. Conclusion

Analysis of decentralized systems presents both conceptual and computational challenges. In this paper, we described two general models of decentralized systems that encompass many practical models. We used the notion of

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.

common information between control agents as a key idea to formulate the decentralized problem as a POMDP. This allowed us to utilize Markov decision theory to come up with information states and sequential decomposition. We further observed that in cases where the private information of agents and their action spaces belong to small finite sets, one can use the computational methods used for POMDPs for carrying out the sequential decomposition we achieved.

The idea of common information also provides the unifying feature in sequential decompositions of many decentralized problems considered in literature. In some cases, the common information may be same as information available at one of the actual control agents. This implies, the problems can be converted to POMDPs from one actual agents perspective, as shown in [11], [12]. Sequential decompositions obtained for decentralized control problems when all agents know the state of the system with k step delay [13] can also be seen as utilizing the notion of common information. In cases when no common information is present (model B), our fictitious agent corresponds to a system designer who is sequentially deciding each agents strategy. The viewpoint of the system designer was used to identify information states for two agent systems equivalent to model B in [14], [15], [16].

The idea of using common information to identify information states for sequential decomposition might also be useful in systems which do not have an explicit common message. For such systems common knowledge between the agents, in the sense of Aumann [17] could play the role analogous to the role of common information in model A. However, we are not aware of any results on explicitly identifying common knowledge between decentralized dynamical systems.

Acknowledgements

This research was supported in part by NASA Grant NNX06AD47G.

References

[1] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation Identification and Adaptive Control*. Prentice Hall, 1986.

[2] H. S. Witsenhausen, "On the sequences of pairs of dependent random variables," *SIAM Journal of Applied Mathematics*, vol. 28, pp. 100–113, 1975.

[3] ——, "Separation of estimation and control for discrete time systems," *Proc. IEEE*, vol. 59, no. 11, pp. 1557–1566, Nov. 1971.

[4] P. Varaiya and J. Walrand, "On delayed sharing patterns," *IEEE Trans. Autom. Control*, vol. 23, no. 3, pp. 443–445, 1978.

[5] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Operations Research*, vol. 11, pp. 1071–1088, 1973.

[6] M. L. Littman, "Algorithms for sequential decision making," Ph.D. dissertation, Brown University, May 1996.

[7] A. Cassandra, M. L. Littman, and N. L. Zhang, "Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes," in *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 1997.

[8] J. Rust, "Using randomization to break the curse of dimensionality," *Econometrica*, vol. 65, no. 3, pp. 487–516, 1997.

[9] A. Wald, *Sequential hypothesis testing*. Wiley, 1947.

[10] A. Mahajan, A. Nayyar, and D. Teneketzis, "Information structures and decentralized control," *in preparation*.

[11] J. C. Walrand and P. Varaiya, "Optimal causal coding—decoding problems," *IEEE Trans. Inf. Theory*, vol. 29, no. 6, pp. 814–820, Nov. 1983.

[12] ——, "Causal coding and control of markov chains," *System and Control Letters*, vol. 3, pp. 189–192, 1983.

[13] M. Aicardi, F. Davoli, and R. Minciardi, "Decentralized optimal control of markov chains with a common past information set," *IEEE Trans. Autom. Control*, vol. 32, no. 11, pp. 1028–1031, 1987.

[14] A. Mahajan and D. Teneketzis, "On globally optimal encoding, decoding and memory update for noisy real-time communication systems," *submitted to IEEE Trans. Inf. Theory*, Jan. 2006, available as Control Group Report CGR-06-03, Department of EECS, University of Michigan, Ann Arbor, MI 48109-2122.

[15] ——, "Optimal performance of feedback control systems with limited communication over noisy channels," *submitted to SIAM Journal of Control and Optimization*, Dec. 2006, available as Control Group Report CGR-06-07, Department of EECS, University of Michigan, Ann Arbor, MI 48109-2122.

[16] A. Mahajan, "Sequential decomposition of sequential dynamic teams: applications to real-time communication and networked control systems," Ph.D. dissertation, University of Michigan, Ann Arbor, MI, 2008.

[17] R. J. Aumann, "Agreeing to disagree," *Annals of Statistics*, no. 4, pp. 1236–39, 1976.

Authorized licensed use limited to: Yale University. Downloaded on April 10, 2009 at 15:33 from IEEE Xplore. Restrictions apply.