Optimal strategies for DEC-POMDPs: An example

Paper ID: 956

Abstract

In this paper, a solution approach to solve a subclass of DEC-POMDPs is presented. For ease of exposition, the approach is presented by using the two-user multiaccess broadcast channel (2-MABC) as an example. The 2-MABC has been used as a benchmark for different DEC-POMDP algorithms; but none of the existing algorithms have been able to optimally solve the infinite horizon discounted 2-MABC. In contrast, the proposed approach transforms the 2-MABC into a POMDP in which the size of the unobserved state is two and the size of the action space is four. As such, it is straightforward to solve this transformed POMDP. The proposed approach is applicable to any DEC-POMDP where the common information (i.e., data known to all the agents) is increasing with time, but the size of the private information (total observations minus the common information) is not changing with time.

Introduction

Multi-agent systems in which agents cooperate to achieve a common goal arise in different application domains: stochastic control, artificial intelligence, operations research, and economics. Such models are called *dynamic teams* in stochastic control, artificial intelligence, and operations research literature, and *DEC-POMDP*s in artificial intelligence literature. In general, such systems are known to be NEXP complete (Bernstein, Zilberstein, and Immerman 2000; Goldman and Zilberstein 2004).

In light of the above complexity results, it is worthwhile to identify subclasses of DEC-POMDPs that are simpler. In particular, we are interested in identifying subclasses that can be converted to POMDPs. One such subclass was identified in the stochastic control literature by (Mahajan, Nayyar, and Teneketzis 2008). The motivation of this paper is to present that solution approach to the artificial intelligence community. This approach is based on solving the system from the point of view of a coordinator that observes data that is common knowledge to all agents. We call this approach the *coordinator approach*. In this paper, we apply this approach to a benchmark problem for DEC-POMDPs: the two-user multiaccess broadcast channel (2-MABC).

The two-user multiaccess broadcast channel is used as a benchmark for DEC-POMDP algorithms. However, none of the algorithms are able to solve this problem exactly. Most exact algorithms run into numerical difficulties after a horizon of four or five. See the section on literature overview for details. The coordinator approach transforms the 2-MABC into a POMDP with finite state and action spaces. The size of (unobserved) is 2 while the size of the action space is 4; thus, this transformed POMDP can be easily solved using standard POMDP algorithms. Hence, the coordinator approach allows us to identify subset of DEC-POMDPs that are equivalent to POMDPs and solve them using standard POMDP algorithms.

Notation

x_t denotes the value of a variable x at time t; $x_{1:t}$ denotes the sequence $x_1, x_2, \dots, x_t$. $\mathbb{P}(\cdot)$ denotes probability of an event; $\mathbb{E}\{\cdot\}$ denotes expectation of a random variable. For any $p \in [0, 1]$, $\bar{p}$ denotes $1 - p$. $\mathbb{N}$ denotes the set of whole numbers $\{0, 1, 2, \dots\}$.

The two-user multiaccess broadcast

System Model

A two-user multiple access broadcast system (2-MABC) is a communication system in which two user communicate to a single receiver over a broadcast medium. Time is slotted. At the beginning of each time slot, packets arrive stochastically at each user. The users have a buffer that can store one packet. If the buffer is empty when a new packet arrives, the packet is stored in the buffer; if the buffer already has a packet, the new packet is dropped. (In some applications, the old packet is dropped and the new packet is stored in the buffer).

After the packet arrival, each user decides whether or not to transmit. If both users transmit simultaneously, the transmissions interfere and the receiver cannot decode; if only one user transmits, the receiver can decode the transmitted packet. At the end of transmission, the receiver feeds back whether or not it successfully decoded a packet. In case of a successful decoding, the transmitting user removes the packet from her buffer. The above process is repeated at each slot.

The design objective is to choose decentralized transmission policies at both users to maximize the expected discounted number of successful packet transmissions over an

infinite horizon.

The salient features of the above model are: (i) Each user knows its own queue state but has only partial information about the queue state of the other user; (ii) The queue dynamics of the two users are coupled due to packet collision.

For user i, $i = 1, 2$, and time t, let $a_{i,t} \in \{0, 1\}$ denote the number of new packet arrivals, $x_{i,t} \in \{0, 1\}$ the number of packets in the buffer, and $u_{i,t} \in \{0, 1\}$ the number of transmitted packets.

The packet arrivals at both users are independent Bernoulli processes with arrival probability p_1 and p_2. Thus, for $i = 1, 2$,

$$a_{i,t} = \begin{cases} 0 & \text{with probability } (1 - p_i) \\ 1 & \text{with probability } p_i \end{cases}$$

Let $z_t \in \{0, 1\}$ indicate if the receiver successfully decoded a packet. Thus,

$$z_t = u_{1,t} \oplus u_{2,t} \tag{1}$$

where $\oplus$ denotes modulo 2 addition. The state of the buffer is updated according to

$$x_{i,t+1} = \max\{x_{i,t} - u_{i,t} z_t, a_{i,t}\} \tag{2}$$

At the end of the slot z_t is fed back to both users. The users choose their transmission decisions based on their histories of buffer states and channel feedback according to a transmission rule $g_{i,t}$ as follows

$$u_{i,t} = g_{i,t}(x_{i,1:t}, u_{i,1:t-1}, z_{1:t-1}) \tag{3}$$

such that $u_{i,t} \leq x_{i,t}$.

At time t, the system gets a reward z_t. We are interested in the following optimization problem.

Problem 1 *Given the arrival rates p_1 and p_2, choose transmission policies $\mathbf{g}_i := (g_{i,1}, g_{i,2}, \dots)$, $i = 1, 2$, of the form* (3) *that maximizes the expected discounted reward*

$$J(\mathbf{g}_1, \mathbf{g}_2) = \sum_{t=1}^{\infty} \mathbb{E}^{\mathbf{g}_1, \mathbf{g}_2} \{\beta^{t-1} z_t\}.$$

where $\beta \in (0, 1)$ is a discount factor and $\mathbb{E}^{\mathbf{g}_1, \mathbf{g}_2}$ denotes the expectation taken with respect to the joint probability measure induced on all the system variables from the choices of $(\mathbf{g}_1, \mathbf{g}_2)$.

We restrict attention to pure policies (also called deterministic policies). Since the optimization problem does not have any constraints, randomization cannot improve performance (DeGroot 1970, Chapter 8). Hence, restriction attention to pure policies is without loss.

Salient Features

In the above model, two agents cooperate to achieve a common, system-wide objective of maximizing a system-wide objective; hence, it is a *team problem* (Marschak and Radner 1972). The actions taken by one agent (whether to transmit or not) affect the observations of the other; hence it is a *dynamic* team problem (Ho 1980). Neither agent knows the observations of the other agent; hence, the system has a *non-classical information structure* (Witsenhausen 1971).

In the artificial intelligence community, team problems with non-classical information structure are called DEC-POMDPs. In the above model, if the agents combine their observations, the state of the system is fully observed; hence, it is a *DEC-MDP*. The observations of an agent depends only on its local state (the buffer size); hence the system has *independent observations*. The evolution of the local state of each agent depends on the actions of the other agent; hence the system has *dependent* transitions. DEC-MDPs with dependent transitions are known to be NEXP-complete (Bernstein, Zilberstein, and Immerman 2000).

Literature overview

Due to space limitations, we only review the literature pertinent to the multiaccess broadcast channel. The MABC is motivated by early satellite communication systems. Early investigations (Schoute 1976; Varaiya and Walrand 1979) of n-user MABC made three simplifying assumptions: the arrival rate at all users is identical; packet collision incurs a cost rather than retransmission; and the queue states are shared between the users with a delay. Later, (Hluchyj and Gallager 1981) removed the last two assumptions and derived lower bounds on optimal performance by restricting attention to *window protocols*. For the same system, (Ooi and Wornell 1996) determined upper bounds by assuming that the state of the queues were shared after a delay (which was a tunable parameter). For 2- and 3-MABC, the upper bounds of (Ooi and Wornell 1996) (with a suitably chosen value of delay) matched the lower bounds of (Hluchyj and Gallager 1981). Thus, the optimal transmission policy for 2- and 3-MABC with symmetric arrival rates is completely characterized.

The 2-MABC with asymmetric arrival rates was introduced as a benchmark problem by (Hansen, Bernstein, and Zilberstein 2004). They derived an exact dynamic programming approach for a general (finite-horizon) DEC-POMDP and applied that to a 2-MABC with $p_1 = 0.9$ and $p_2 = 0.1$. They were able to identify optimal strategy for a horizon of four, before running out of memory. (Szer, Charpillet, and Zilberstein 2005) presented a heuristic technique based on the A* algorithm to solve general DEC-POMDPs. Their MAA* algorithm was able to solve 2-MABC for horizon four *within a few minutes* (actual time was not reported). They were also able to compute approximate solutions for larger horizons. (Szer and Charpillet 2005) restricted attention to finite state controllers and proposed a heuristic best-first search algorithm to find the best finite state controllers. They found solutions to the infinite horizon 2-MABC for $p_1 = 0.9$, $p_2 = 0.1$, and discount factor $\beta = 0.9$. A finite-state controller is similar to the window protocol proposed by (Hluchyj and Gallager 1981). (Bernstein, Hansen, and Zilberstein 2005) presented a bounded policy iteration for DEC-POMDPs by restricting attention to finite state controllers and assuming that all agents have access to shared randomness. (Szer and Charpillet 2006) presented a point based approximation for dynamic programming approach. Using this approach they were able to solve a 2-MABC for horizon 5 exactly, and find approximate solutions for horizon up to 8. (Seuken and Zilberstein 2007) presented a

heuristic to sample the belief space and solved the 2-MABC for large horizon (100,00) in reasonable time (less than 24 hours). Similar performance was obtained by incremental pruning by (Dibangoye, Mouaddib, and Chaib-draa 2008).

In contrast, the coordinator approach transforms the 2-MABC into a POMDP whose (unobserved) state space is of size 2 and action space is of size 4; for a horizon T, the reachable set of belief states is of size $4 + 2(T + 1) + (T + 1)^2 \approx (T + 2)^2$. Hence, for finite horizon, the 2-MABC can be solved by solving an MDP in which the size of the state space is $(T + 2)^2$ and the size of the action space is 2; for infinite horizon, the 2-MABC can solved either by off-the-shelf POMDP algorithms, or by successive finite-state approximations of the reachable set.

Preliminary results

We first provide an alternative form of the model. Notice that $z_t = u_{1,t} \oplus u_{2,t}$. This implies that

$$u_{1,t} = z_t \oplus u_{2,t} \quad \text{and} \quad u_{2,t} = z_t \oplus u_{1,t}$$

As a result, any control policy of the form (3) can also be written as

$$u_{i,t} = g_{i,t}(x_{i,1:t-1}, u_{1,1:t-1}, u_{2,1:t-1}). \quad i = 1, 2. \qquad (4)$$

Hence, Problem 1 is equivalent to the following Problem:

Problem 2 *Maximize the reward functions of Problem 1 when the control laws are of the form* (4).

In Problem 2, both control agents know the control actions of the other agent after one unit delay. Such models are said to have *control sharing information structure*. The next result is applicable to any DEC-MDP with control sharing information structure. The proof relies on the following result:

Lemma 1 *The buffer state processes* $\{x_{i,t}, t = 1, 2, \dots\}$ *are conditionally independent given* $\{(u_{1,t-1}, u_{2,t-1}), t = 1, 2, \dots\}$*, i.e.,*

$$\begin{aligned}&\mathbb{P}(x_{1,1:t}, x_{2,1:t} | u_{1,1:t-1}, u_{2,1:t-1}) \\ &= \mathbb{P}(x_{1,1:t} | u_{1,1:t-1}, u_{2,1:t-1}) \mathbb{P}(x_{2,1:t} | u_{1,1:t-1}, u_{2,1:t-1})\end{aligned}$$

This lemma follows from the law of total probability and the system dynamics.

Proposition 1 *In Problem 2, restricting attention to control laws of the form*

$$u_{i,t} = g_{i,t}(x_{i,t}, u_{1,1:t-1}, u_{2,1:t-1}), \quad i = 1, 2 \qquad (5)$$

is without loss □

PROOF We prove the result for agent 1. The result for agent 2 follows from symmetry. Arbitrarily fix the policy of agent 1 of the form (4) and consider Problem 2 from the point of view of agent 1. Let $y_t := (x_{1,t}, u_{1,1:t-1}, u_{2,1:t-1})$. Lemma 1 implies that

1. $\mathbb{P}(y_{t+1} | y_{1:t}, u_{1,1:t}) = \Pr(y_{t+1} | y_t, u_{1,t})$
2. $\mathbb{E}\{z_t | y_{1:t}, u_{1,1:t}\} = \mathbb{E}\{z_t | y_t, u_{1,t}\}$

Thus, $\{y_t, t = 1, 2, \dots\}$ is a controlled Markov process with control action $u_{1,t}$. Moreover, the expected reward function is a function of y_t and $u_{1,t}$. This implies that for any fixed policy of agent 2, Problem 2 from the point of view of agent 1 is a MDP (Markov decision process). Thus, from the standard results in Markov decision process (Kumar and Varaiya 1986), restricting attention to control laws of the form

$$u_{1,t} = g_{1,t}(y_t) = g_{1,t}(x_{1,t}, u_{1,1:t-1}, u_{2,1:t-1})$$

is without loss. ■

Proposition 1 implies that Problem 2, and hence Problem 1, is equivalent to the following:

Problem 3 *Maximize the reward functions of Problem 1 when the control laws are of the form* (5).

The coordinator approach

In this section, we illustrate the coordinator approach for dynamic teams and DEC-POMDPs. The main idea is to coordinate the policies of all agents based on *shared information* $(u_{1,1:t-1}, u_{2,1:t-1})$. For that matter, we consider an alternative system, which we call the *coordinated system*, consisting of a *coordinator* and two *passive* agents. The coordinator observes

$$o_t = (u_{1,t-1}, u_{2,t-1})$$

at time t and chooses actions $(w_{1,t}, w_{2,t}) \in \{0, 1\}^2$ according to

$$(w_{1,t}, w_{2,t}) = h_t(o_{1:t}, w_{1,1:t-1}, w_{2,1:t-1}) \qquad (6)$$

Both agents are passive. Agent i observes $x_{i,t}$ and $w_{i,t}$ and generates $u_{i,t}$ according to

$$u_{i,t} = x_{i,t} w_{i,t}. \qquad (7)$$

The system dynamics and the reward are the same as in the model described in Section . Thus, the reward at time t is

$$z_t = (x_{1,t} w_{1,t}) \oplus (x_{2,t} w_{2,t}). \qquad (8)$$

In the above coordinated system, the only decision maker is the coordinator; the agents simply carry out the calculations prescribed by (7). The coordinator has to solve the following optimization problem.

Problem 4 *Given the arrival rates* p_1 *and* p_2*, choose coordination policy* $\mathbf{h} := (h_1, h_2, \dots)$ *of the form* (6) *that maximizes the expected discounted reward*

$$\hat{J}(\mathbf{h}) = \sum_{t=1}^{\infty} \mathbb{E}^{\mathbf{h}}\{\beta^{t-1} z_t\}.$$

The main advantage of constructing the above centralized coordinator system is that it is equivalent to the original system. Specifically, we have the following:

Proposition 2 *Any transmission policy* $(\mathbf{g}_1, \mathbf{g}_2)$ *for Problem 3 can be implemented in Problem 4 by a corresponding coordination policy* $\mathbf{h}$ *with identical expected reward. Conversely, any coordinator policy* $\mathbf{h}$ *for Problem 4 can be implemented in Problem 3 by a transmission policy* $(\mathbf{g}_1, \mathbf{g}_2)$ *for Problem 3 with identical expected reward.* □

PROOF We prove the first part of the proposition. The proof of the second part is similar. Consider a transmission policy $(\mathbf{g}_1, \mathbf{g}_2)$ for Problem 3. To implement this policy in Problem 4 set a coordination policy $\mathbf{h}$ for Problem 4 by choosing

$$\begin{bmatrix} w_{1,t} \\ w_{2,t} \end{bmatrix} = \begin{bmatrix} g_{1,t}(x_{1,t} = 1, u_{1,1:t-1}, u_{2,1:t-1}) \\ g_{2,t}(x_{2,t} = 1, u_{1,1:t-1}, u_{2,1:t-1}) \end{bmatrix} =: h_t(o_{1:t}, w_{1,1:t-1}, w_{2,1:t-1}) \quad (9)$$

Now consider Problem 3 and 4 for a specific realization of $x_{1,1}$, $x_{2,1}$, $\{a_{1,1:t}, t = 1, 2, \dots\}$, and $\{a_{2,t}, t = 1, 2, \dots\}$. The choice (9) of $\mathbf{h}$ implies that $\{(x_{1,t}, x_{2,t}, u_{1,t}, u_{2,t}, z_t), t = 1, 2, \dots\}$ are identical in Problems 3 and 4. Thus, any transmission policy $(\mathbf{g}_1, \mathbf{g}_2)$ for Problem 3 can be implemented in Problem 4 by choosing $\mathbf{h}$ according to (9). Furthermore, since the system variables in the two Problems are identical along all sample paths, the expected reward of the transmission policy $(\mathbf{g}_1, \mathbf{g}_2)$ for Problem 3 is identical to the expected reward of the coordination policy $\mathbf{h}$ for Problem 4. ■

The coordinator system as a POMDP

Proposition 2 implies that instead of investigating Problem 3, we can investigate the coordinated system of Problem 4. The latter is a centralized system with one decision maker: the coordinator. The hidden state of the system at time t is $(x_{1,t-1}, x_{2,t-1})$. The observation $o_t = (u_{1,t-1}, u_{2,t-1})$ of the coordinator is a function of the current state $(x_{1,t-1}, x_{2,t-1})$ and the past control action $(w_{1,t-1}, w_{2,t-1})$, given by (7). The coordinator chooses a control action $(w_{1,t}, w_{2,t})$ according to (6). The instantaneous reward z_t, given by (8), is a function of the next state $(x_{1,t}, x_{2,t})$ and the current control action $(w_{1,t}, w_{2,t})$. *Hence, Problem 4 is a POMDP.* Standard results on POMDPs (see (Kumar and Varaiya 1986)) imply that the coordinator can choose its action based on the belief state

$$\mathbb{P}\left(x_{1,t-1}, x_{2,t-1} \,\middle|\, \begin{matrix} u_{1,1:t-1}, u_{2,1:t-1} \\ w_{1,1:t-1}, w_{2,1:t-1} \end{matrix} \right)$$

Lemma 1 implies that this belief state decomposes as

$$\prod_{i=1,2} \mathbb{P}\left(x_{i,t-1} \,\middle|\, \begin{matrix} u_{1,1:t-1}, u_{2,1:t-1} \\ w_{1,1:t-1}, w_{2,1:t-1} \end{matrix} \right)$$

Furthermore, because $x_{i,t-1} \in \{0, 1\}$, the belief state is equivalent to $(\pi_{1,t}, \pi_{2,t})$ where

$$\pi_{i,t} = \mathbb{P}\left(x_{i,t-1} = 1 \,\middle|\, \begin{matrix} u_{1,1:t-1}, u_{2,1:t-1} \\ w_{1,1:t-1}, w_{2,1:t-1} \end{matrix} \right), \quad i = 1, 2$$

Hence, we have the following:

Proposition 3 *In problem 3, restricting attention to coordination policy of the form*

$$(w_{1,t}, w_{1,t}) = h_t(\pi_{1,t}, \pi_{2,t})$$

is without loss of optimality. Consequently, in Problem 2, restriction attention to a transmission policy of the form

$$u_{i,t} = g_{i,t}(x_{i,t}, \pi_{1,t}, \pi_{2,t}), \quad i = 1, 2$$

is without loss of optimality. □

PROOF The result follows from the fact that in POMDPs, we can choose a control action based on the belief state, and that the belief state is equivalent to $(\pi_{1,t}, \pi_{2,t})$. The second part of the result follows from the equivalence between Problems 2 and 3. ■

To compactly write the time evolution of the belief state $(\pi_{1,t}, \pi_{2,t})$ define operators A_1, A_2 from $[0, 1]$ to $[0, 1]$ such that for any $\pi \in [0, 1]$

$$A_i \pi = p_i + (1 - p_i)\pi, \quad i = 1, 2. \quad (10)$$

Note that, we can show by induction that

$$A_i^n \pi = 1 - \bar{p}_i^n \bar{\pi}, \quad i = 1, 2, \quad n \in \mathbb{N} \quad (11)$$

where $\bar{p}_i = (1 - p_i)$ and $\bar{\pi} = (1 - \pi)$.

Proposition 4 *The value of $(\pi_{1,t+1}, \pi_{2,t+1})$ depends only on the value of $(\pi_{1,t}, \pi_{2,t})$ and $(w_{1,t}, w_{2,t})$. Specifically:*

1. *When $(w_{1,t}, w_{2,t}) = (0, 0)$,*
$$(\pi_{1,t+1}, \pi_{2,t+1}) = (A_1\pi_{1,t}, A_2\pi_{2,t}).$$
2. *When $(w_{1,t}, w_{2,t}) = (1, 0)$,*
$$(\pi_{1,t+1}, \pi_{2,t+1}) = (p_1, A_2\pi_{2,t}).$$
3. *When $(w_{1,t}, w_{2,t}) = (0, 1)$,*
$$(\pi_{1,t+1}, \pi_{2,t+1}) = (A_1\pi_{1,t}, p_2).$$
4. *When $(w_{1,t}, w_{2,t}) = (1, 1)$,*
$$(\pi_{1,t+1}, \pi_{2,t+1}) = \begin{cases} (1, 1) & \text{with prob. } \pi_{1,t}\pi_{2,t} \\ (p_1, p_2) & \text{with prob. } 1 - \pi_{1,t}\pi_{2,t} \end{cases}$$

□

PROOF The result follows directly from the definition of $(\pi_{1,t}, \pi_{2,t})$ and the system dynamics. ■

Dynamic programming decomposition

Theorem 1 *Since the coordinated system is a POMDP, the optimal coordination policy is stationary (does not depend on time), that is*

$$(w_{1,t}, w_{2,t}) = h(\pi_{1,t}, \pi_{2,t})$$ □

The control law h can be found by the solution of the following fixed point equation:

$$V(\pi_{1,t}, \pi_{2,t}) = \max_{(w_1,w_2)\in\{0,1\}^2} \{W_{w_1,w_2}(\pi_1, \pi_2)\} \quad (12)$$

where

$$\begin{aligned} W_{00}(\pi_1, \pi_2) &= \beta V(A_1\pi_1, A_2\pi_2) \\ W_{01}(\pi_1, \pi_2) &= \pi_2 + \beta V(A_1\pi_1, p_2) \\ W_{10}(\pi_1, \pi_2) &= \pi_1 + \beta V(p_1, A_2\pi_2) \\ W_{11}(\pi_1, \pi_2) &= \pi_1 + \pi_2 - 2\pi_1\pi_2 + \beta\pi_1\pi_2 V(1, 1) \\ &\quad + \beta(1 - \pi_1\pi_2)V(p_1, p_2) \end{aligned}$$

This dynamic program corresponds to the dynamic program of a POMDP whose (unobserved) state space is of size two and action space is of size four. Such POMDPs can be easily solved numerically (Smallwood and Sondik 1973; Cassandra, Littman, and Zhang 1997). However, we can further simplify the numerical solution by analyzing the reachable set of belief states.

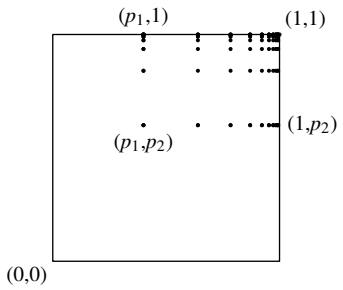

Figure 1: The reachable set of information states for $p_1 = 0.4$ and $p_2 = 0.6$.

Proposition 5 *Suppose the system starts in state* (p_1, p_2). *Then, the reachable set of* (π_1, π_2) *is countable and given by*

$$S = \{(1,1), (p_1,1), (1,p_2), (p_1,p_2)\} \bigcup \{(A_1^n p_1, p_2), (p_1, A_2^m p_2), (A_1^n p_1, A_2^m p_2) : n, m \in \mathbb{N}\} \quad (13)$$

PROOF This is an immediate consequence of Proposition 4. ■

The reachable set for $p_1 = 0.4$ and $p_2 = 0.6$ is shown in Figure 1. Notice that the reachable set is considerably smaller than whole space $[0,1]^2$. We need to solve the dynamic program of Theorem 1 only for $(\pi_1, \pi_2) \in S$, resulting in considerable computational savings. The resultant dynamic program has countable state space and finite action space, and can be solved using finite state approximations (White 1980).

A natural candidate for finite state approximation is restricting attention to the set

$$S_\ell = \{(1,1), (p_1,1), (1,p_2), (p_1,p_2)\} \bigcup \{(A_1^n p_1, p_2), (p_1, A_2^m p_2), (A_1^n p_1, A_2^m p_2) : n, m \in \{1, \dots, \ell\}\} \quad (14)$$

for some fixed value of ℓ. Such a restriction corresponds to a situation in which each user must transmit at least once within ℓ consecutive slots. For such a finite state approximation, we need to solve the dynamic program In this case, we only need to solve the dynamic program of Theorem 1 only for $(\pi_1, \pi_2) \in S_\ell$. The resultant dynamic program has a finite state and action spaces, and can be solved using standard value or policy iteration (Puterman 1994).

Remark about finite horizon

If we were interested in a finite horizon system, then the reachability analysis provides a crude way to model the system as a MDP. In a finite horizon 2-MABC with horizon T, the reachable set of the state space is S_{T+1}. Thus, we can transform it into an MDP where the state space is S_{T+1} and the action space is 4: hence the solution complexity is $O(4T|S_{T+1}|) = O(T^3)$. Thus, even with a very loose upper bound, the solution complexity is only increasing polynomially with the horizon.

The general approach

As the above example illustrates, it is possible to convert a non-trivial DEC-POMDP into a POMDP. As mentioned in the introduction, the general approach is presented in (Mahajan, Nayyar, and Teneketzis 2008). For completeness we briefly describe the main idea here.

Consider a general DEC-POMDP with n agents. Let $I_{k,t}$ denote all the data available with agent k at time t. The agents choose control actions $A_{k,t}$ as a function of $I_{k,t}$ according to

$$A_{k,t} = g_{k,t}(I_{k,t})$$

Partition this available data into two sets: common observations

$$C_t := \bigcap_k I_{k,t}$$

and local observations

$$L_{k,t} := I_{k,t} \setminus C_t$$

Suppose the following two assumptions are true.

Assumption 1 (Increasing common observations) *The common observations are increasing with time,* i.e.*,*

$$C_t \subseteq C_{t+1}, \quad \forall t$$

Assumption 2 (Finite local observations) *The local observations do not increase with time,* i.e.*,*

$$|L_{k,t}| = |L_{k,t+1}|, \quad \forall k, \forall t$$

View the system from the point-of-view of a coordinator that observes C_t. The coordinator chooses function sections

$$\hat{g}_{k,t}(\cdot) = g_{k,t}(\cdot, C_t)$$

according to

$$(\hat{g}_{1,t}, g_{2,t}, \dots, g_{k,t} = \psi_t(C_t)$$

and communicates these actions to all agents.[1] The agents are passive. They simply use the function sections $\hat{g}_{k,t}$ and compute $A_{k,t}$ from their local observations, *i.e,*,

$$A_{k,t} = \hat{g}_{k,t}(L_{k,t}).$$

Similar to Proposition 2, we can show that the coordinator's problem is equivalent to the original problem. However, the coordinator's problem is a POMDP. So, we can write a dynamic programming decomposition using standard approach. The details are presented in (Mahajan, Nayyar, and Teneketzis 2008).

[1] Since each agent observes the common information, they can simulate the actions of the coordinator and compute the function sections $\hat{g}_{k,t}$ on their own.

Conclusion

In this paper, the coordinator approach (Mahajan, Nayyar, and Teneketzis 2008) for transforming a DEC-POMDP (or dynamic team) to a POMDP was presented using the example of 2-MABC. In order for the coordinator approach to work, the system must satisfy Assumptions 1 and 2. The (unobserved) state space of the transformed POMP is the concatenation of the state space of the original system and the local observations of all agents; the action space of the POMDP is the concatenation of the space of functions from the local observation of an agent to its action space. Thus, the conversion of a DEC-POMDP to a POMDP comes at a significant increase in the size of the state and action space. The state of the art numerical algorithms for solving POMDPs are able to handle large state and action spaces. Hence, the coordinator approach provides an alternate way of solving a subset of DEC-POMDPs.

References

Bernstein, D. S.; Hansen, E. A.; and Zilberstein, S. 2005. Bounded policy iteration for decentralized POMDPs. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (UAI)*, 1287–1292.

Bernstein, D. S.; Zilberstein, S.; and Immerman, N. 2000. The complexity of decentralized control of markov decision processes. In *Proceedings of the 16th International Conference on Uncertainty in Artificial Intelligence (UAI)*, 32–27.

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

DeGroot, M. 1970. *Optimal Statistical Decisions*. McGraw Hills.

Dibangoye, J. S.; Mouaddib, A.-I.; and Chaib-draa, B. 2008. Incremental pruning heuristic for solving DEC-POMDPs. In *Proceedings of the Workshop on Multi-Agent Sequential Decision Making in Uncertain Domains (MSDM)*.

Goldman, C. V., and Zilberstein, S. 2004. Decentralized control of cooperative systems. *Journal of Artificial Intelligence Research* 143–174.

Hansen, E. A.; Bernstein, D. S.; and Zilberstein, S. 2004. Dynamic programming for partially observable stochastic games. In *Proceedings of the 19th national conference on artificial intelligence (AAAI)*, 709–715.

Hluchyj, M. G., and Gallager, R. G. 1981. Multiacces of a slotted channel by finitely many users. In *Proceedings of National Telecommunication Conference*, D.4.2.1–D.4.2.7.

Ho, Y.-C. 1980. Team decision theory and information structures. *Proceedings of the IEEE* 68(6):644–654.

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

Mahajan, A.; Nayyar, A.; and Teneketzis, D. 2008. Identifying tractable decentralized control problems on the basis of information structures. In *proceedings of the 46th Allerton conference on communication, control and computation*, 1440–1449.

Marschak, J., and Radner, R. 1972. *Economic Theory of Teams*. New Haven: Yale University Press.

Ooi, J. M., and Wornell, G. W. 1996. Decentralized control of a mutliple access broadcast channel: performance bounds. In *Proccedings of the 35th Conference on Decision and Control*, 293–298.

Puterman, M. 1994. *Markov decision processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons.

Schoute, F. C. 1976. Decentralized control in packet switched satellite communication. *IEEE Transactions on Automatic Control* AC-23(2):362–271.

Seuken, S., and Zilberstein, S. 2007. Memory-bounded dynamic programming for DEC-POMDPs. In *Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI)*, 2009–2015. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

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

Szer, D., and Charpillet, F. 2005. An optimal best-first search algorithm for solving infinite horizon DEC-POMDPs. In Gama, J. a.; Camacho, R.; Brazdil, P.; Jorge, A.; and Torgo, L., eds., *Machine Learning*, volume 3720 of *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin / Heidelberg. chapter 38, 389–399.

Szer, D., and Charpillet, F. 2006. Point-based dynamic programming for DEC-POMDPs. In *proceedings of the 21st national conference on Artificial intelligence (AAAI)*, 1233–1238. AAAI Press.

Szer, D.; Charpillet, F.; and Zilberstein, S. 2005. MAA*: A heuristic search algorithm for solving decentralized POMDPs. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*.

Varaiya, P., and Walrand, J. 1979. Decentralized control in packet switched satellite communication. *IEEE Transactions on Automatic Control* AC-24(5):794–796.

White, D. 1980. *Recent Developments in Markov Decision Processes*. Academic Press. chapter Finite state approximations for denumerable state infinite horizon discounted Markov processes: The method of successive approximations.

Witsenhausen, H. S. 1971. Separation of estimation and control for discrete time systems. *Proceedings of the IEEE* 59(11):1557–1566.