Forty-Eighth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 29 - October 1, 2010

Optimal transmission policies for two-user multiple access broadcast using dynamic team theory

Aditya Mahajan
Department of Electrical and Computer Engineering,
McGill University, Montreal, QC, Canada, H3A 2A7

***Abstract*—Optimal transmission policies for a two-user multiple access broadcast channel with binary feedback are investigated. The system is modeled as a dynamic team. Conditioned on the channel feedback, the past history of buffer states is redundant to each user. Once this redundant data is removed, the information structure, although non-classical, satisfies the sufficient conditions of Mahajan, Nayyar, and Teneketzis, 2008 for a dynamic team problem to be tractable. Using the idea of a virtual coordinator of Mahajan, Nayyar, and Teneketzis, a dynamic programming decomposition is presented. This dynamic program is defined over a countable state space and a finite action space. When the arrival rates of both users are symmetric, an optimal policy is identified by solving the dynamic program analytically. This policy matches the optimal window protocol proposed by Hluchyj and Gallager, 1981. Thus, this paper presents the an example of a non-trivial dynamic team with non-linear dynamics where an exact analytic solution is obtained.**

I. Introduction

A. System Model

Consider a two-user multiple access broadcast system (MABC) in which two users communicate to a common 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 average throughput over a finite or 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.

In this paper we analyze the two-user MABC using dynamic team theory and provide a dynamic programming decomposition. This dynamic program has countable state space and finite action space. When the arrival rates at both users are symmetric, we provide an analytic solution of the dynamic program.

B. Literature overview

Dynamic team theory has been used to study MABC as modeled above and its variants. Schoute [1] and Varaiya and Walrand [2] investigated MABC under the assumptions that packet collision incurs a cost rather than retransmission and that the queue states are shared between the users with a delay. Grizzle *et. al.* investigated MABC under the assumption that queue state and transmission decisions are shared between the users with one-step delay. In this paper, we do not make any assumption on delayed sharing of information.

Hluchyj and Gallager [3] investigated the two-user MABC with symmetric arrivals by restricting attention to *window protocols*. Thus, their approach provides a lower bound on optimal performance. Ooi and Wornell [4] investigated two-user MABC with symmetric arrivals under the assumption that queue state and transmission decisions are shared between the users with some delay. Thus, their approach provides an upper bound on performance. It turns out that the lower bound of [3] matches the upper bound of [4]. Hence, for the case of symmetric arrivals, the optimal policy is known.

For the case of asymmetric arrivals not much is known about the optimal policy. Various researchers [5]–[8]. have used MABC as a benchmark problem for the numerical algorithms for decentralized stochastic control problems. These algorithms are either heuristic and do not provide any optimality guarantees, or can compute the optimal policy only for small horizon (usually running out of memory at horizon four or five). These attempts highlight the difficulty of the seeming simple two-user MABC.

C. Contributions

The main contributions of this paper are two-fold.

1) We provide a dynamic programming decomposition of the two-user MABC. The dynamic program has a countable state space and finite action space. Hence, the optimal policy can be computed numerically using the standard methods for solving such dynamic programs [9]. This is in contrast to the existing attempts to numerically compute the optimal policies, which either relied on heuristics or could not handle large horizons.

978-1-4244-8214-6/10/$26.00 ©2010 IEEE

2) For the case of symmetric arrival rates, an optimal policy is obtained by analytically solving the dynamic program. The previous proof was based on identifying tight upper and lower bounds.

Thus, this paper presents an example of a non-trivial decentralized control problem with non-linear dynamics where an exact analytic solution has been obtained.

D. 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 $\{0,1,2,\dots\}$.

II. Problem formulation

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 $i = 1,2$,

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

and

$$x_{i,t+1} = (x_{i,t} - u_{i,t} z_t) \vee a_{i,t}$$

where $\vee$ denotes binary OR.

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 exclusive OR. 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

$$u_{i,t} = g_{i,t}(x_{i,1:t}, u_{i,1:t-1}, z_{1:t-1}) \tag{2}$$

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

At time t, the system gets a reward $r(u_{1,t} \oplus u_{2,t})$, where $r \ge 0$ is a normalizing constant. We are interested in the following optimization problem.

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

$$J_T(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}^{\mathbf{g}_1, \mathbf{g}_2}\left\{ \sum_{t=1}^{T} r(u_{1,t} \oplus u_{2,t}) \right\}$$

or the average expected reward per unit time as $T \to \infty$

$$\overline{J}(\mathbf{g}_1, \mathbf{g}_2) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}^{\mathbf{g}_1, \mathbf{g}_2}\left\{ \sum_{t=1}^{T} r(u_{1,t} \oplus u_{2,t}) \right\}.$$

where $\mathbb{E}^{\mathbf{g}_1, \mathbf{g}_2}$ 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 [10, Chapter 8]. Hence, restriction attention to pure policies is without loss.

III. Structure of optimal transmission policy

Problem 1 has a non-classical information structure [11] because both users do not know each other's buffer state. As a result, it is hard to compress each user's data, which is increasing with time, into an information state whose domain is not increasing with time. Unless such information states are identified, systems with non-classical information structure are intractable.

Sufficient conditions under which non-classical information structures are tractable were identified in [12]. The model of Section II does not satisfy these conditions. However, we show that conditioned on the channel feedback, the history of buffer states at both users are redundant, and after this redundant data is removed, the information structure, although non-classical, satisfies the conditions of [12]. Therefore, a tractable sequential decomposition is possible.

A. Feedback implies control sharing

Since the control actions are binary and the feedback is the XOR of the control actions, each user can recover the other user's control action from the feedback and its own control. In particular, user 1 can recover $u_{2,t}$ because $u_{2,t} = u_{1,t} \oplus z_t$ and user 2 can recover $u_{1,t}$ because $u_{1,t} = u_{2,t} \oplus z_t$. Hence,

$$(u_{i,1:t-1}, z_{1:t-1}) \leftrightarrow (u_{1,1:t-1}, u_{2,1:t-1}).$$

Consequently, we have the following:

Proposition 1: In problem 1, each user can use a transmission policy of the form

$$u_{i,t} = g_{i,t}(x_{i,1:t-1}, u_{1,1:t-1}, u_{2,1:t-1}). \tag{3}$$

B. Removing redundant data

Proposition 2: For any transmission policy of user 2 of the form (3), restriction attention to a transmission policy of the form

$$u_{1,t} = g_{1,t}(x_{1,t}, u_{1,1:t-1}, u_{2,1:t-1}) \tag{4}$$

at user 1 is without loss. By symmetry, the result is also true if the role of users 1 and 2 is interchanged.

Combining both the cases of the above, we have the following:

Corollary 1: Restricting attention to transmission policies of the form

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

is without loss

Due to lack of space, we omit the proof of Proposition 2. Corollary 1 implies that Problem 1 is equivalent to the following problem.

Problem 2: In Problem 1, determine optimal policies of the form (5).

The information structure of Problem 2 is same as Model A of [12]. Each user's data consists of two parts: shared observations $(u_{1,1:t-1}, u_{2,1:t-1})$ that are increasing with time, and private observation $x_{1,t}$ and $x_{2,t}$ that have a fixed size. So, Problem 2 can be sequentially decomposed using the methodology of [12].

For that matter, consider a virtual coordinator that observes the shared observations $(u_{1,1:t-1}, u_{2,1:t-1})$ and chooses how each user uses her private observations $x_{1,t}$ or $x_{2,t}$. Any policy for the coordinator can be implemented in original problem and vice versa. Hence, the two problems are equivalent. The coordinator's problem, which is centralized, sequentially decomposes by an appropriate choice of information state. The corresponding dynamic programming decomposition also determines optimal transmission policy for Problem 2.

C. Virtual coordinator

Consider the following modified problem. In the model described in Section II, in addition to the two users, a virtual coordinator is present. The coordinator has perfect recall, *i.e.*, it remembers its past observations and control actions. At time t, it observes $(u_{1,t-1}, u_{2,t-1})$ and chooses *partial functions*

$$\gamma_{i.t} : \{0,1\} \mapsto \{0,1\}$$

for user i, $i = 1, 2$. Each user then uses its assigned partial function to generate a transmission decision as follows:

$$u_{i,t} = \gamma_{i,t}(x_{i,t}).$$

Since $u_{i,t} \leq x_{i,t}$, $\gamma_{i,t}$ is equivalent to $\varphi_{i,t} = \gamma_{i,t}(1)$. Hence, we can write

$$u_{i,t} = \varphi_{i,t} x_{i,t}. \tag{6}$$

We call $(\varphi_{1,t}, \varphi_{2,t})$ the control action of the coordinator. These are similar to the partial functions defined in [12].

The coordinator chooses the current control action based on all past observations and past control actions as follows:

$$(\varphi_{1,t}, \varphi_{2,t}) = h_t(u_{1,1:t-1}, u_{2,1:t-1}, \varphi_{1,1:t-1}, \varphi_{2,1:t-1}) \tag{7}$$

The system dynamics and the reward are the same as in the model described in Section II. Thus, the reward at time t is $r((x_{1,t}\varphi_{1,t}) \oplus (x_{2,t}\varphi_{2,t}))$

In the above formulation, the only decision maker is the virtual coordinator; the users simply carry out the calculations prescribed by (6). The virtual coordinator has to solve the following optimization problem.

Problem 3: Given the arrival rates p_1 and p_2 and a time horizon T, choose coordination policy $\mathbf{h} := (h_1, \dots, h_T)$ of the form (7) that maximizes the expected total reward

$$J_T(\mathbf{h}) = \mathbb{E}^{\mathbf{h}}\left\{ \sum_{t=1}^{T} r((x_{1,t}\varphi_{1,t}) \oplus (x_{2,t}\varphi_{2,t})) \right\}.$$

or the average reward per unit time as $T \to \infty$

$$\overline{J}(\mathbf{h}) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}^{\mathbf{h}}\left\{ \sum_{t=1}^{T} r((x_{1,t}\varphi_{1,t}) \oplus (x_{2,t}\varphi_{2,t})) \right\}.$$

Problems 2 and 3 are equivalent.

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

$$\varphi_{i,t} = g_{i,t}(1, u_{1,1:t-1}, u_{2,1:t-1}). \tag{8}$$

Now consider Problem 2 and 3 for a specific realization of $x_{1,1}$, $x_{2,1}$, $a_{1,1:T}$, and $a_{2,1:T}$. The choice (8) of $\mathbf{h}$ implies that $x_{1,1:T}$, $x_{2,1:T}$, $u_{1,1:T}$, $u_{2,1:T}$, and $z_{1:T}$ are identical in Problems 2 and 3. Thus, any transmission policy $(\mathbf{g}_1, \mathbf{g}_2)$ for Problem 2 can be implemented in Problem 3 by choosing $\mathbf{h}$ according to (8). 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 2 is identical to the expected reward of the coordination policy $\mathbf{h}$ for Problem 3. ∎

D. Information states at the virtual coordinator

To sequentially decompose Problem 3, we define the following.

Definition 1: Let $\pi_{i,t}$ be the posterior belief of the coordinator that the buffer of user i is full, *i.e.*,

$$\pi_{i,t} = \mathbb{P}\left(x_{i,t} = 1 \,\middle|\, \begin{matrix} u_{1,1:t-1}, u_{2,1:t-1} \\ \varphi_{1,1:t-1}, \varphi_{2,1:t-1} \end{matrix} \right) \tag{9}$$

Proposition 4: Let A_i, $i = 1, 2$, be operators from $[0,1]$ to $[0,1]$ defined for any $\pi \in [0,1]$, as

$$A_i\pi = 1 - \bar{p}_i\bar{\pi} = p_i + \bar{p}_i\pi.$$

Then,

$$A_i^n\pi = 1 - \bar{p}_i^n\bar{\pi}.$$

Proof: The proof follows from induction. ∎

Proposition 5: The vector $(\pi_{1,t}, \pi_{2,t})$ evolves as follows:

1) When $(\varphi_{1,t}, \varphi_{2,t}) = (0,0)$,

$$(\pi_{1,t+1}, \pi_{2,t+1}) = (A_1\pi_{1,t}, A_2\pi_{2,t}).$$

2) When $(\varphi_{1,t}, \varphi_{2,t}) = (1,0)$,

$$(\pi_{1,t+1}, \pi_{2,t+1}) = (p_1, A_2\pi_{2,t}).$$

3) When $(\varphi_{1,t}, \varphi_{2,t}) = (0,1)$,

$$(\pi_{1,t+1}, \pi_{2,t+1}) = (A_1\pi_{1,t}, p_2).$$

4) When $(\varphi_{1,t}, \varphi_{2,t}) = (1,1)$,

$$(\pi_{1,t+1}, \pi_{2,t+1}) = \begin{cases} (1,1) & \text{if } x_{1,t} = x_{2,t} = 1 \\ (p_1, p_2) & \text{otherwise} \end{cases}$$

Proof: The result follows directly from the definition of $(\pi_{1,t}, \pi_{2,t})$. ∎

Proposition 6: 1) For any coordination policy of the form (7), the process $\{(\pi_{1,t}, \pi_{2,t})\}$ is a controlled Markov chain with control action $(\varphi_{1,t}, \varphi_{2,t})$, that is

$$\mathbb{P}\left(\pi_{1,t+1}, \pi_{2,t+1} \middle| \begin{array}{l} \pi_{1,1:t}, u_{1,1:t-1}, \varphi_{1,1:t-1} \\ \pi_{2,1:t}, u_{2,1:t-1}, \varphi_{2,1:t-1} \end{array}\right) = \mathbb{P}(\pi_{1,t+1}, \pi_{2,t+1} | \pi_{1,t}, \pi_{2,t}, \varphi_{1,t}, \varphi_{2,t})$$

2) The conditional instantaneous reward may be written as

$$\mathbb{E}\left\{r(u_{1,t} \oplus u_{2,t}) \middle| \begin{array}{l} \pi_{1,1:t}, u_{1,1:t-1}, \varphi_{1,1:t-1} \\ \pi_{2,1:t}, u_{2,1:t-1}, \varphi_{2,1:t-1} \end{array}\right\} = \mathbb{E}\{r(u_{1,t} \oplus u_{2,t}) | \pi_{1,t}, \pi_{2,t}, \varphi_{1,t}, \varphi_{2,t})$$

Proof:

1) The result follows from the update equation of Proposition 5.
2) From Definition 1, we have

$$\begin{aligned}\mathbb{P}\left(\pi_{1,t+1}, \pi_{2,t+1} \middle| \begin{array}{l} \pi_{1,1:t}, u_{1,1:t-1}, \varphi_{1,1:t-1} \\ \pi_{2,1:t}, u_{2,1:t-1}, \varphi_{2,1:t-1} \end{array}\right) &= r(\pi_{1,t}\varphi_{1,t}(1 - \pi_{2,t}\varphi_{2,t}) + (1 - \pi_{1,t}\varphi_{1,t})\pi_{2,t}\varphi_{2,t}) \\ &= \mathbb{P}(\pi_{1,t+1}, \pi_{2,t+1} | \pi_{1,t}, \pi_{2,t}, \varphi_{1,t}, \varphi_{2,t})\end{aligned}$$

∎

E. Main structural result

The above result implies that the vector $(\pi_{1,t}, \pi_{2,t})$ is an information state for Problem 3. Consequently, we have the following.

Theorem 1: In Problem 2, restricting attention to coordination policy of the form

$$(\varphi_{1,t}, \varphi_{2,t}) = h_t(\pi_{1,t}, \pi_{2,t}) \tag{10}$$

is with loss. Consequently, in Problem 2, restricting attention to a transmission policy of the form

$$u_{i,t} = g_{i,t}(x_{i,t}, \pi_{1,t}, \pi_{2,t}) \tag{11}$$

is without loss.

Proof: Proposition 6 implies that the coordinator's optimization problem can be viewed as an MDP in which the underlying Markov process is $(\pi_{1,t}, \pi_{2,t})$ and the instantaneous cost is $\mathbb{E}\{u_{1,t} \oplus u_{2,t} | \pi_{1,t}, \pi_{2,t}, \varphi_{1,t}, \varphi_{2,t}\}$. This MDP formulation implies the result of the theorem. ∎

The information state $(\pi_{1,t}, \pi_{2,t})$, exploits the special structure of the two-user MABC, and as such is simpler than the information state $\mathbb{P}(x_{1,t}, x_{2,t} \mid u_{1,1:t-1}, u_{2,1:t-1}, \varphi_{1,t-1}, \varphi_{2,1:t-1})$ proposed in [12]. This simplification of the information state was also reported in [13].

IV. Dynamic programming decomposition

A. Finite horizon

Since $(\pi_{1,t}, \pi_{2,t})$ is a controlled Markov process, Problem 3 sequentially decomposes as follows:

Theorem 2: In Problem 3, an optimal policy of the form (10) is given by the solution of the following dynamic program. For any $\pi_1, \pi_2 \in [0, 1]$,

$$V_{T+1}(\pi_1, \pi_2) = 0 \tag{12}$$

and for $t = T, T-1, \ldots, 1$,

$$\begin{aligned}V_t(\pi_1, \pi_2) = \max \big\{&W_{10,t}(\pi_1, \pi_2), \\ &W_{01,t}(\pi_1, \pi_2), W_{11,t}(\pi_1, \pi_2)\big\}\end{aligned} \tag{13}$$

where $W_{ij,t}$, $i, j \in \{0, 1\}$ denotes the expected future reward if $(\varphi_{1,t}, \varphi_{2,t})$ is chosen to be (i, j), *i.e.*,

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

Proof: The above dynamic program follows from the MDP formulation presented in the proof of Theorem 1. ∎

B. Properties of the value function

Proposition 7: The functions V_t, $W_{10,t}$, $W_{01,t}$, and $W_{11,t}$ satisfy the following properties: for all t

1) $W_{10,t}(\pi_1, \pi_2)$ is linear in π_1 and convex in π_2.
2) $W_{01,t}(\pi_1, \pi_2)$ is convex in π_1 and linear in π_2.
3) $W_{11,t}(\pi_1, \pi_2)$ is component-wise linear in π_1 and π_2.
4) $V_t(\pi_1, \pi_2)$ is component-wise convex in π_1 and π_2.

Proof: The proof proceeds by induction. The properties are true at $T + 1$. Suppose they are also true at $t + 1$. Then, 1) and 2) follow from the convexity of V_{t+1} and linearity of A_1 and A_2, 3) follows from definition, and 5) follows from the component-wise convexity of $W_{10,t}$, $W_{01,t}$, and $W_{11,t}$. ∎

C. Properties of the optimal policy

Definition 2: Partition the space $[0, 1]^2$ of realization of $(\pi_{1,t}, \pi_{2,t})$ into three disjoint regions $R_{10,t}$, $R_{01,t}$, and $R_{11,t}$ such that $R_{ij,t}$ denotes the region where $(\varphi_{1,t}, \varphi_{2,t}) = (i, j)$ is optimal, $i, j \in \{0, 1\}$.

We say that $R_{ij,t}$, $i, j \in \{0, 1\}$, is *element-wise convex* in π_1 if for any π_1', π_1'', π_2, $\lambda \in [0, 1]$ and

$$\pi_1 = \lambda\pi_1' + \bar{\lambda}\pi_1'',$$

such that (π_1', π_2), $(\pi_1'', \pi_2) \in R_{ij,t}$, then $(\pi_1, \pi_2) \in R_{ij,t}$. Element-wise convexity in π_2 is defined in a similar manner.

Proposition 8: For all t, the regions $R_{10,t}$ and $R_{11,t}$ are element-wise convex in π_1; the regions $R_{01,t}$ and $R_{11,t}$ are element-wise convex in π_2.

Proof: We prove the result for one case. The proof of other cases is similar. Suppose (π_1', π_2), $(\pi_1'', \pi_2) \in R_{10,t}$, $\lambda \in [0,1]$, and $\pi_1 = \lambda\pi_1' + \bar{\lambda}\pi_1''$. Then,

$$\begin{aligned} W_{10,t}(\pi_1, \pi_2) &\overset{(a)}{=} \lambda W_{10,t}(\pi_1', \pi_2) + \bar{\lambda} W_{10,t}(\pi_1'', \pi_2) \\ &\overset{(b)}{\geq} \lambda W_{01,t}(\pi_1', \pi_2) + \bar{\lambda} W_{01,t}(\pi_1', \pi_2) \\ &\overset{(c)}{\geq} W_{01,t}(\pi_1, \pi_2) \end{aligned}$$

where (a) is true because $W_{10,t}(\pi_1, \pi_2)$ is linear in π_1; (b) is true because (π_1', π_2), $(\pi_1'', \pi_2) \in R_{10,t}$; and (c) is true because $W_{01,t}(\pi_1, \pi_2)$ is convex in π_1. By a similar argument, we can show that $W_{10,t}(\pi_1, \pi_2) \geq W_{11,t}(\pi_1, \pi_2)$. Hence, $(\pi_1, \pi_2) \in R_{10,t}$. ∎

D. Infinite horizon

When $T = \infty$, the sequence of nested functions of Theorem 2 simplify as follows.

Theorem 3: In Problem 3, when $T = \infty$, the optimal coordination policy is stationary and given by the solution of the following fixed point equation.

$$\begin{aligned} v(\pi_1, \pi_2) + J^* = \max \big\{ & w_{10}(\pi_1, \pi_2), \\ & w_{01}(\pi_1, \pi_2), w_{11}(\pi_1, \pi_2) \big\} \end{aligned} \quad (14)$$

where J^* denotes the average reward per unit time, $v(\pi_1, \pi_2)$ is the differential reward at (π_1, π_2) and w_{ij}, $i, j \in \{0, 1\}$ are the expected differential reward if (φ_1, φ_2) is chosen to be (i, j), *i.e.*,

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

Proof: Since reward function is bounded, the result follows from standard dynamic programming arguments. See [14]. ∎

Similar to Definition 2, let R_{ij} be the region where $(\varphi_1, \varphi_2) = (i, j)$ is optimal. Then, Proposition 8 implies that R_{10} and R_{11} are element-wise convex in π_1 while R_{01} and R_{11} are element-wise convex in π_2.

E. Reduction to a countable state MDP

The information state (π_1, π_2) takes value in the set $[0,1]^2$. However, the reachable set of the information state is countable.

Proposition 9: Suppose the system starts in state (p_1, p_2). Then, the reachable set of (π_1, π_2) is countable and given by

$$\begin{aligned} S = \{&(1,1), (p_1, 1), (1, p_2), (p_1, p_2)\} \\ &\bigcup \{(A_1^n p_1, p_2), (p_1, A_2^n p_2), : n \in \mathbb{N}\} \end{aligned} \quad (15)$$

Proof: This is an immediate consequence of Proposition 5. ∎

The reachable set for $p_1 = p_2 = 0.4$ 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 3 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 the methods described in [9].

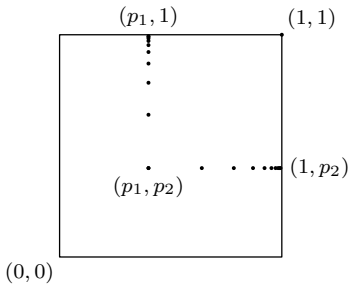

Fig. 1. Reachable set of (π_1, π_2) for $p_1 = p_2 = 0.4$.

In some applications, in particular sensor networks, it is reasonable to assume that each user must transmit at least once within m consecutive time slots, where m is finite. If we make such an assumption, then the reachable state space is finite and is given by

$$\begin{aligned} S_m = &\{(1,1), (p_1, 1), (1, p_2), (p_1, p_2)\} \\ &\bigcup \{(A_1^n p_1, p_2), (p_1, A_2^n p_2) : n = 1, \ldots, m\} \end{aligned} \quad (16)$$

In this case, we only need to solve the dynamic program of Theorem 3 only for $(\pi_1, \pi_2) \in S_m$. The resultant dynamic program has a finite state and action spaces, and can be solved using standard policy iteration [15].

V. Symmetric arrival rates

In this section, we focus on the infinite horizon case for symmetric arrival rates, *i.e.*, $p_1 = p_2 = p$. For this case, $A_1 = A_2$, so we denote both operators by A. An immediate consequence of symmetric arrival rates is that the optimal policy is symmetric: for any $\pi_1, \pi_2 \in [0,1]$

$$h(\pi_1, \pi_2) = h(\pi_2, \pi_1).$$

A. Optimal policies

We exploit the symmetry of the optimal policy to analytically solve the dynamic program of Theorem 3 and identify an optimal policy. For that matter, we define the following.

Definition 3: Let for $n \in \mathbb{N}$

$$f_n(x) = 1 + (1-x)^2 - (3+x)(1-x)^{n+1}.$$

and s_n denote the root of $f_n(x)$ that is between $[0,1]$. The function $f(n)$ is negative for $x \in [0, s_n]$ and positive in $x \in [s_n, 1]$. Thus, s_n is a decreasing sequence. Moreover, $s_0 \approx 0.70711$ and $s_1 \approx 0.34729$. Let τ be the root of $x = (1-x)^2$ that lies in $[0,1]$. Observe that $s_1 < \tau < s_0$ and $\tau \approx 0.38196$.

Theorem 4: An optimal policy for Problem 3 is: For $p \geq \tau$,

$$h^*(\pi_1, \pi_2) = \begin{cases} (1,0) & \text{if } \pi_1 > \pi_2, \\ (0,1) & \text{if } \pi_1 < \pi_2, \\ (1,0) \text{ or } (0,1) & \text{if } \pi_1 = \pi_2. \end{cases} \quad (17)$$

For $p \leq \tau$, let $m \in \mathbb{N}$ be such that $s_{m+1} \leq p \leq s_m$. Then,

$$h^*(\pi_1, \pi_2) = \begin{cases} (1,1) & \text{if } \pi_1 \leq A^m p, \pi_2 \leq A^m p, \\ (1,0) & \text{if } \pi_1 > \pi_2, \pi_1 > A^m p \\ (0,1) & \text{if } \pi_1 < \pi_2, \pi_2 > A^m p \\ (1,0) \text{ or } (0,1) & \text{if } \pi_1 = \pi_2 = 1. \end{cases} \quad (18)$$

The average reward per unit time is

$$J^* = \begin{cases} rp[1 - (2p^2 - 1)/D(p)] & \text{if } p \leq s_1, \\ r(1 - \bar{p}^2) & \text{if } s_1 \leq p; \end{cases} \quad (19)$$

where $D(p) := 1 + p^2 + p^3$.

See Appendix A for proof.

B. Qualitative properties

Although the optimal policy looks complicated, with different behavior depending on the value of p, essentially it has only two modes of operation. When $p \geq s_1$, the states $\{(p, Ap), (Ap, p)\}$ are absorbing and form a recurrence class in S. Within this recurrence class, the optimal policy is a round-robin policy. When $p \leq s_1$, the states $\{(1,1), (p,p), (Ap,p), (p,Ap)\}$ are absorbing and form a recurrence class in S. Within this recurrence class, the optimal policy is identical for all $p \leq s_1$. Each user transmits if it has a packet. If a collision occurs, both users know that both of them have a packet. So, they simply empty their buffer one-by-one, and then go back to "transmit if you have a packet".

Thus, for both $p \geq s_1$ and $p \leq s_1$, the optimal policy *restricted to the recurrent states* is a window protocol [3] and identical to the one proposed by Hluchyj and Gallager.

According to the optimal policy, irrespective of the value of p each user gets a transmission opportunity (*i.e.*, $w_{i,t} = 1$) at least once in two consecutive time slots. This property enforces a two-step delayed state sharing information structure. So, if we could establish this property beforehand, we could use the dynamic programming decomposition for delayed state sharing information structure [16]. In fact, Schoute [1] and Varaiya and Walrand [2] assumed this property while deriving structural properties for their model. However, we could find a direct way of proving this property.

VI. Conclusion

We presented a dynamic programming decomposition for finding optimal decentralized transmission policies for two-user MABC. This dynamic program has a countable state space and a finite action space. When both users have symmetric arrival rates, we find an analytic solution to the dynamic program, thereby identifying an optimal policies. Our approach differs from existing approaches in the literature, which have focused on either a restriction of the model, or a relaxation of the model, or on heuristic approximations. In contrast, in this paper we provide a direct and explicit solution.

The model of this paper is an example of a dynamic team problem in which the search of an optimal solution is tractable. The insights provided by this example may also be useful for understanding general dynamic team problems.

Acknowledgement

The author is indebted to Serdar Yüksel and Ashutosh Nayyar for valueable discussions.

Appendix A
Proof of Theorem 4

The proof of the theorem is purely algebraic. We simply guess the differential reward functions and show that they satisfy the dynamic programming equations of Theorem 3. At some places, the proof exploits a recursive property of $f_n(x)$, namely.

Lemma 1: For any $n \in \mathbb{N}$

$$f_{n+1}(x) - (1-x) f_n(x) = p(1 + \bar{p}^2)$$

The rest is simply elementary (but tedious) algebra.

We prove the cases, $p \geq \tau$, $\tau > p \geq s_1$, $s_1 > p \geq s_2$, etc. separately.

A. Case 1: $p \geq \tau$

For this case, the differential reward functions are

$$\begin{aligned} v(p, A^n p) = v(A^n p, p) &= r(1 - \bar{p}^{n+1}), \quad n > 1 \\ v(p,1) = v(1,p) &= r, \\ v(1,1) &= r(1 + \bar{p}^2), \\ v(p,p) &= rp \end{aligned}$$

To show that h^* is optimal, we need to show two things. First

$$\begin{aligned} w_{01}(p, A^n p) &= r(1 - \bar{p}^{n+1}) + v(Ap, p) = v(p, A^n p) + J^*, \\ w_{01}(p,1) &= r + v(Ap, p) = v(p,1) + J^*, \\ w_{01}(1,1) &= r + v(1,p) = v(1,1) + J^*, \\ w_{01}(p,p) &= rp + v(p, Ap) = v(p,p) + J^*. \end{aligned}$$

which is easy to verify.

Next we show that if $\pi_1 \leq \pi_2$,

$$w_{01}(\pi_1, \pi_2) \geq \max\{w_{10}(\pi_1, \pi_2), w_{11}(\pi_1, \pi_2)\}$$

We show this on a case-by-case basis:

1) For $(\pi_1, \pi_2) = (p, A^n p)$, we have

$$w_{01}(p, A^n p) - w_{10}(p, A^n p) = rp\bar{p}(1 - \bar{p}^n) \geq 0$$

Furthermore,

$$w_{01}(p, Ap) - w_{11}(p, Ap) = rp^2[1 + \bar{p}(1 - (2 + \bar{p})\bar{p}^n)]$$

Observe that

$$1 - (2 + \bar{p})\bar{p}^2 = (1 + \bar{p})(p - \bar{p}^2).$$

Thus, for $p \geq \tau$, this term is positive. Hence, for $n \geq 2$,

$$1 - (2 + \bar{p})\bar{p}^n \geq 1 - (2 + \bar{p})\bar{p}^2 \geq 0.$$

For $n = 1$, the term in the square brackets is $-p^3 + 5p^2 - 6p + 1$ which is positive for $p \geq 0.19806$. Hence, h^* is optimal at $(p, A^n p)$ and $(A^n p, p)$, $n > 1$.

2) For $(\pi_1, \pi_2) = (p, 1)$, we have

$$w_{01}(p, 1) - w_{10}(p, 1) = r\,\bar{p} \geq 0$$

and

$$w_{01}(p, 1) - w_{11}(p, 1) = rp^2(1 + \bar{p}) \geq 0$$

Hence, h^* is optimal at $(p, 1)$ and $(1, p)$

3) For $(\pi_1, \pi_2) = (1, 1)$, we have

$$w_{01}(1, 1) - w_{10}(1, 1) = 0$$

and

$$w_{01}(1, 1) - w_{11}(1, 1) = rp(1 + \bar{p}) \geq 0$$

Hence, h^* is optimal at $(1, 1)$.

4) For $(\pi_1, \pi_2) = (p, p)$, we have

$$w_{01}(p, p) - w_{10}(p, p) = 0$$

and

$$w_{01}(p, p) - w_{11}(p, p) = rp^2(p - \bar{p}^2)$$

which is positive for $p \in [0, 1]$ such that $p > \bar{p}^2$ or $p > \tau$. Hence, h^* is optimal at $(1, 1)$.

Thus, we have proved that h^* is optimal for $p \geq \tau$.

B. Case 2: $s_1 \leq p \leq \tau$

For this case, the differential reward functions are

$$\begin{aligned} v(p, A^n p) = v(A^n p, p) &= r(1 - \bar{p}^{n+1}), n > 1 \\ v(p, 1) = v(1, p) &= r, \\ v(1, 1) &= r(1 + \bar{p}^2), \\ v(p, p) &= r\,\bar{p}^2 \end{aligned}$$

As before, it is easy to verify that

$$\begin{aligned} w_{01}(p, A^n p) &= r(1 - \bar{p}^n) + v(Ap, p) = v(p, A^n p) + J^*, \\ w_{01}(p, 1) &= r + v(Ap, p) = v(p, 1) + J^*, \\ w_{01}(1, 1) &= r + v(1, p) = v(1, 1) + J^*, \\ w_{11}(p, p) &= 2rp\bar{p} + p^2 v(p, p) + (1 - 2p^2)v(1, 1) \\ &= v(p, p) + J^*. \end{aligned}$$

Next, we will show that for $\pi_1 < \pi_2$, and $\pi_1 = \pi_2 = 1$,

$$w_{01}(\pi_1, \pi_2) \geq \max\{w_{10}(\pi_1, \pi_2), w_{11}(\pi_1, \pi_2)\}$$

while for $\pi_1 = \pi_2 = p$,

$$w_{11}(p, p) \geq \max\{w_{01}(p, p), w_{10}(p, p)\}$$

As before, we show this on a case-by-case basis:

1) For $(\pi_1, \pi_2) = (p, A^n p)$, we have

$$w_{01}(p, A^n p) - w_{10}(p, A^n p) = rp\bar{p}(1 - \bar{p}^n) \geq 0$$

Furthermore,

$$\begin{aligned} w_{01}(p, Ap) - w_{11}(p, Ap) &= r(1 - 2\bar{p}^2 - p\bar{p}^{n+1}) \\ &= r[f_n(p) - 3\bar{p}^2(1 - \bar{p}^{n-1})] \end{aligned}$$

which is positive for $p \geq s_1$. Hence, h^* is optimal at $(p, A^n p)$ and $(A^n p, p)$.

2) For $(\pi_1, \pi_2) = (p, 1)$

$$w_{01}(p, 1) - w_{10}(p, 1) = r\,\bar{p} \geq 0$$

Furthermore,

$$w_{01}(p, 1) - w_{11}(p, 1) = -rf_0(\bar{p})$$

which is positive if $\bar{p} < s_0$, or equivalently, $p > 1 - s_0$. Since $p > s_1 > 1 - s_0$, $w_{01}(p, 1) - w_{11}(p, 1) \geq 0$. Hence, h^* is optimal at $(p, 1)$ and $(1, p)$.

3) For $(\pi_1, \pi_2) = (1, 1)$, the calculations are the same as for $p \geq \tau$.

4) For $(\pi_1, \pi_2) = (p, p)$,

$$\begin{aligned} w_{01}(p, p) - w_{10}(p, p) &= w_{11}(p, p) - w_{01}(p, p) \\ &= r(\bar{p}^2 - p) \end{aligned}$$

which is positive if $p < \bar{p}^2$, or equivalently, $p < \tau$. Hence, h^* is optimal at (p, p).

Thus, we have proved that h^* is optimal for $s_1 \leq p \leq \tau$.

C. General case: $s_{m+1} \leq p < s_m$, $m \in \mathbb{N}$

For this case,

$$J^* = rp[1 - f_0(p)/D(p)]$$

where $D(p) := 1 + p^2 + p^3$. The differential reward functions are

$$\begin{aligned} v(p, 1) = v(1, p) &= J^*, \\ v(1, 1) &= r, \\ v(p, p) &= rf_1(p)/D(p), \\ v(A^n p, p) = v(p, A^n p) &= \begin{cases} c_*(n) & \text{if } n \leq m, \\ c^*(n) & \text{if } n > m \end{cases} \end{aligned}$$

where

$$\begin{aligned} c_*(n) &= \frac{\bar{p}}{p}(1 - \bar{p}^n)J^* + r\,\bar{p}^{n+1} - r\,\bar{p} + v(p, p), \\ c^*(n) &= r(1 - \bar{p}^{n+1}) + c_*(1) - J^* \end{aligned}$$

The functions c_* and c^* satisfy the following property.

Lemma 2: For any $n \in \mathbb{N}$,

$$\begin{aligned} c_*(n + 1) - c_*(n) &= p\,\bar{p}^{n+1} f_0(p)/D(p), \\ c^*(n + 1) - c^*(n) &= p\,\bar{p}^{n+1} \end{aligned}$$

which follow from elementary algebra. Using the above recursion, we can express $c_*(n)$ and $c^*(n)$ in terms of $c_*(1)$ and $c^*(1)$.

To show that h^* is optimal, we need to show two things. First

$$\begin{aligned}
w_{11}(p, Ap) &= r(p + Ap + pAp) + pApv(1,1) \\
&\quad + (1 - pAp)v(p, p) = v(p, Ap) + J, \\
w_{01}(p, A^n p) &= rA^n p + v(Ap, p) = v(p, A^n p) + J, \quad n > 1 \\
w_{01}(p, 1) &= r + v(Ap, p) = v(p, 1) + J^*, \\
w_{01}(1, 1) &= r + v(1, p) = v(1, 1) + J^*, \\
w_{11}(p, p) &= 2rp\bar{p} + p^2 v(1, 1) + (1 - p^2)v(p, p) \\
&= v(p, p) + J^*.
\end{aligned}$$

which is easy to verify. Second, when $\pi_1 < A^m p$ and $\pi_2 < A^m p$,

$$w_{11}(\pi_1, \pi_2) \geq \max\{w_{10}(\pi_1, \pi_2), w_{01}(\pi_1, \pi_2)\}$$

otherwise when $\pi_1 \geq \pi_2$,

$$w_{01}(\pi_1, \pi_2) \geq \max\{w_{10}(\pi_1, \pi_2), w_{11}(\pi_1, \pi_2)\}$$

We show these on a case-by-case basis.

1) For $(\pi_1, \pi_2) = (p, A^n p)$, $n \leq m$, we have

   $$w_{11}(p, A^n p) - w_{10}(p, A^n p) = -rp(1 - \bar{p}^{n+1}) f_0(p)/D(p)$$

   which is positive since $p \leq \tau < s_0$. Furthermore,

   $$w_{11}(p, A^n p) - w_{01}(p, A^n p) = -rp^2 f_n(p)/D(p)$$

   which is positive for $p \leq s_n$. Since $n \leq m$, $s_n \geq s_m$. By assumption $p < s_m$. Thus, $p \leq s_n$ and hence h^* is optimal at $(p, A^n p)$ and $(A^n p, p)$, $n \leq m$.
2) For $(\pi_1, \pi_2) = (p, A^n p)$, $n > m$, we have

   $$\begin{aligned}
   w_{01}(p, A^n p) - w_{10}(p, A^n p) &= rp\left[-\frac{f_0(p)}{D(p)} - \bar{p}^{n+1}\right] \\
   &\geq p\left[\frac{-f_0(p)}{D(p)} - \bar{p}\right] \\
   &= p\left[\frac{-pf_1(p)}{D(p)}\right]
   \end{aligned}$$

   which is positive for $p \in [0, s_1]$. Furthermore,

   $$w_{01}(p, A^n p) - w_{11}(p, A^n p) = rp^2 f_n(p)/D(p)$$

   which is positive when $p \geq s_n$. Since $n \geq m + 1$, $s_{m+1} \geq s_n$. By assumption, $p \geq s_{m+1}$. Thus, $p \geq s_n$ and h^* is optimal at $(p, A^n p)$, $(A^n p, p)$ for $n > m$.
3) For $(\pi_1, \pi_2) = (p, 1)$, we have

   $$w_{01}(p, 1) - w_{10}(p, 1) = r\bar{p} \geq 0$$

   Furthermore,

   $$w_{01}(p, 1) - w_{11}(p, 1) = rp^2(1 + \bar{p}^2)/D(p) \geq 0$$

   Hence h^* is optimal at $(p, 1)$ and $(1, p)$
4) For $(\pi_1, \pi_2) = (1, 1)$, we have

   $$w_{01}(1, 1) - w_{10}(1, 1) = 0.$$

   Furthermore,

   $$w_{01}(1, 1) - w_{11}(1, 1) = rp(1 + p)(1 + \bar{p}^2)/D(p) \geq 0.$$

   Hence, h^* is optimal at $(1, 1)$.
5) For $(\pi_1, \pi_2) = (p, p)$, we have

   $$w_{11}(p, p) - w_{01}(p, p) = rp^2(1 - 2p^2)/D(p)$$

   which is positive since $p < s_0$. Furthermore, by symmetry $w_{10}(p, p) = w_{01}(p, p)$, so $w_{11}(p, p) - w_{01}(p, p) \geq 0$. Hence, h^* is optimal at (p, p).

Thus, we have proved that h^* is optimal for $s_{m+1} \leq p < s_m$, $m \in \mathbb{N}$.

REFERENCES

[1] F. C. Schoute, "Decentralized control in packet switched satellite communication," *IEEE Trans. Autom. Control*, vol. AC-23, no. 2, pp. 362–271, Apr. 1976.

[2] P. Varaiya and J. Walrand, "Decentralized control in packet switched satellite communication," *IEEE Trans. Autom. Control*, vol. AC-24, no. 5, pp. 794–796, Oct. 1979.

[3] M. G. Hluchyj and R. G. Gallager, "Multiacces of a slotted channel by finitely many users," in *Proceedings of National Telecommunication Conference*, 1981, pp. D.4.2.1–D.4.2.7.

[4] J. M. Ooi and G. W. Wornell, "Decentralized control of a mutliple access broadcast channel: performance bounds," in *Proccedings of the 35th Conference on Decision and Control*, Kobe, Japan, 1996, pp. 293–298.

[5] E. A. Hansen, D. S. Bernstein, and S. Zilberstein, "Dynamic programming for partially observable stochastic games," in *Proceedings of the 19th national conference on artificial intelligence (AAAI)*, San Jose, CA, Jul. 2004, pp. 709–715.

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

[7] D. Szer and F. Charpillet, "Point-based dynamic programming for dec-pomdps," in *proceedings of the 21st national conference on Artificial intelligence (AAAI)*. AAAI Press, 2006, pp. 1233–1238.

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

[9] L. I. Sennott, "The computation of average optimal policies in denumerable state Markov decision chains," *Advances in Applied Probability*, vol. 29, pp. 114–137, 1997.

[10] M. DeGroot, *Optimal Statistical Decisions*. McGraw Hills, 1970.

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

[12] A. Mahajan, A. Nayyar, and D. Teneketzis, "Identifying tractable decentralized control problems on the basis of information structures," in *proceedings of the 46th Allerton conference on communication, control and computation*, Sep. 2008, pp. 1440–1449.

[13] A. Anastasopoulos, "A sequential transmission scheme for multiple access channel with noiseless feedback," 2009, preprint.

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

[15] R. A. Howard, *Dynamic Programming and Markov Processes*. The M.I.T. Press, 1960.

[16] 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.