Stochastic games with one step delay sharing information pattern with application to power control

Eitan Altman, Vijay Kamble, Alonso Silva

Stochastic Games with One-Step Delay Sharing Information Pattern with Application to Power Control

Eitan Altman, Vijay Kamble and Alonso Silva

Abstract

Non-cooperative game theory has gained much interest as a paradigm for decentralized control in communication networks. It allows to get rid of the need for a centralized controller. Decentralizing the decision making may result in situations where agents (decision makers) do not have the same view of the network: the information available to agents vary from one agent to another. The global view of the network state cannot be available to an agent as fast as the information on its local state. Incorporating into the decentralized control paradigm this information asymmetry renders it applicable to a much wider class of situations. In this paper we model the above information asymmetry using the one-step delay sharing information pattern from team theory and generalize it to the context of non-cooperative games. We study its properties and apply it to distributed power control problem.

1 Introduction

A well known paradigm for decentralized control that had been studied extensively is team theory: it replaces a central controller by agents that may work in a distributed way to achieve a common objective. We consider in this paper a distributed power control problem within the framework of the one-step delay sharing information pattern. It models decision problems by various agents where each agent has some local information on its own environment. It is assumed that this information is available to each agent instantaneously where as the information concerning the rest of the system is available to an agent after one unit of time.

The team problem with decentralized information can be transformed into an equivalent Partially Observable Markov Decision Process (PO-MDP), that can be solved using dynamic programming once we transform it to an equivalent Completely Observable Markov Decision Process (CO-MDP) see [5, 2, 3, 4]. The problem is that this transformation comes at a cost of enlarging the state space. In many problems involving decentralized information, the whole history has to be taken as a state which implies that the state space grows exponentially in the time horizon. An important challenge has been to identify information structure for which the dimension of the state space does not grow.

In this paper we introduce a special case of a one-step delay sharing information problem and then use it not only in the standard team context but also in the context of non-cooperative stochastic games.

Our first contributions is to extend the known solution framework for solving the one-step delay sharing team problem to (i) the game setting and (ii) to the case where there are additional constraints (both in the team as well as in the game settings).

These setting involve some non-trivial problems appear both related to the fact that we cannot restrict to pure policies anymore: we need to *randomize*. The first problem related to randomization is that it is not obvious to go from randomizations of actions in the original model to randomization between the "equivalent actions" (that are in fact policies) in the transformed equivalent model. A second problem that arises in the team problem consists on the fact that optimal randomized policies require joint randomizations by agents, which may not be possible. Therefore unlike the non-constrained case, the transformed team problem is not a standard CO-MDP.

We provide answers to both problems in the paper and then use the results to solve a power control problem.

2 The original model

We start by introducing the original Markov Decision Process (MDP) with one-step delay sharing information pattern. We consider N players and we denote by $\mathscr{P}(G)$ the set of probability measures over a measurable set G.

2.1 The original problem

1. $\mathbf{X}^i$ is the local state space of agent i, and $\mathbf{X} = \prod_{j=1}^N \mathbf{X}^j$ is the global state space,
2. $\mathbf{A}^i$ is the action space of agent i and $\mathbf{A} = \prod_{j=1}^N \mathbf{A}^j$ is the global action space,
3. $c_i(x, a)$ is a K-dimentional vector of instantaneous costs for agent i when the global state is x and the global action is a,
4. Q_{xay} is the probability of moving from global state x to global state y when the actions are a,
5. **One-step delay sharing information:** The information available to agent i at time t is given by
$$h^i(t) = (x^i(t), \delta(t-1)), \text{ where}$$
$$\delta(t-1) = \{(x^j(s), a^j(s)) \mid s \leq t-1, 1 \leq j \leq N\}$$
Let $H^i(t)$ be the set of all possible informations of length t for agent i.
6. At time t, agent i chooses an action a according to some probability $u^i(t)$ where the choice is done independently of the choices of other agents. A **strategy** or *policy* u^i for agent i is defined to be a sequence $(u^i(1), u^i(2), ...)$ where $u^i(t)$ is a mapping from the local information set $H^i(t)$ to the set $\mathscr{P}(\mathbf{A}^i)$ of probabilities over $\mathbf{A}^i$. Let U^i denote the set of policies for agent i. Let $U := \prod_{j=1}^N U^i$ be the set of *multi-strategies*.
7. A policy u^i for agent i is said to be quasi stationary if $u^i(t)$ does not depend on t nor on $(x(s), a(s))$ for all $s < t-1$. It is thus only a function of $(x(t-1), a(t-1))$ and of $x^i(t)$. We denote by U^i_{qs} the set of such policies. A pure stationary multi-policy is a mapping from $\mathbf{X}$ to $\mathbf{A}$: it does not depend on time and on previous states and actions.
8. Each initial distribution β over $(x(1), a(1))$ and a multi-strategy u induce a unique probability measure over the space of histories. These define the distribution of the state and action stochastic processes $\{X(t), A(t)\}$.

Definition 2.1. *Define the* **full information version** *of the original problem as the one obtained by replacing the one-step delay sharing information by the full information: $h(t) = \delta(t)$, and then defining the initial probability distribution over the initial state $X(1)$.*

2.2 The team and the game problems

We associate with each agent i a performance (cost) vector

$$C^i(\beta, u) = (C^{i,0}(\beta, u), C^{i,1}(\beta, u), ..., C^{i,K}(\beta, u))$$

that depends on the initial distribution β over the state space and a multi-strategy u. $C^{i,k}(\beta, u)$ may stand for the expected average cost $C^{i,k}_{ea}(\beta, u)$ or for the discounted cost $C^{i,k}_{\alpha}(\beta, u)$ where

$$C^{i,k}_{ea}(\beta, u) = \limsup_{s \to \infty} \frac{1}{s} \sum_{t=1}^{s} \mathbb{E}^u_\beta[c_{i,k}(X(t), A(t))],$$

$$C_\alpha^{i,k}(\beta,u)=\limsup_{s\to\infty}\sum_{t=1}^{s}\alpha^{t-1}\mathbb{E}_\beta^u[c^{i,k}(X(t),A(t))]$$

for some discount factor α (which may depend on i).

Fix an initial distribution β over $X_1\times A_1$ and some constants $V^{i,k}, i=1,...,N, k=1,...,K$. Then agent i is faced with the problem

$$\min_{u^i} C^{i,0}(\beta,u)\ \text{ s. t. }\ C^{i,k}(\beta,u)\le V^{i,k},\ k=1,...,K$$

A policy u^i that satisfies the above constraints is said to be feasible for agent i.

Definition 2.2. *We say that the Slater condition holds if for each agent i there exists a policy u^i such that for every policy v^{-i} of the other players, $C^{i,k}(\beta,[u^i,v^{-i}]) < V^{i,k},\ k=1,...,K$, i.e. all constraints are satisfied with strict inequality.*

We shall discuss the following cases:

- **Non-Cooperative Game:** We wish to find a multistrategy u such that each of its components u^i is feasible, and for any agent i and any policy v^i that is feasible for agent i, $C^{i,0}(\beta,u)\le C^{i,0}(\beta,[v^i,u^{-i}])$,
- **The Team Problem:** $C^{i,0}$ is the same for all i; we then eliminate i from the notation. We seek for a policy u that minimizes $C^0(\beta,u)$ over all policies that are feasible for all i.

3 Transforming into a full information problem

We introduce the following Markov Decision Process:

- $\widehat{\mathbf{X}}=\mathbf{X}\times\mathbf{A}$: the global state $\widehat{x}(t)$ of this MDP at time t equals to the pair $x(t-1)\times a(t-1)$ of the initial Markov Decision Process.
- $\widehat{\mathbf{A}}^i$ is given by the set of mappings from $\mathbf{X}^i$ to $\mathbf{A}^i$. We call such a map a "micro pure stationary policy". We call a map from $\mathbf{X}^i$ to the set $\mathscr{P}(\mathbf{A}^i)$ of probability distributions over $\mathbf{A}^i$ a "micro randomized stationary policy".
- Let $r=(x,a)$, $s=(y,b)$ be elements of $\widehat{\mathbf{X}}$ and g an element of $\widehat{\mathbf{A}}$. Define

$$\widehat{Q}_{rgs}=Q_{xay}1\{g^i(y^i)=b^i, i=1,...,N\}$$

- We define the immediate expected costs as:

$$\hat{c}_i(\hat{x},g)=\hat{c}_i(x,a,g)=\sum_{y\in\mathbf{X}}Q_{xay}c_i(x,g^i(y^i))$$

- Let $\widehat{H}^i(t)$ be the set of all possible informations of length t for agent i.
- At time t, agent i chooses an action $\widehat{a}$ according to some probability $\widehat{u}^i(t)$ where the choice is done independently of the choices of other agents. A **strategy** or *policy* $\widehat{u}^i$ for agent i is defined to be a sequence $(\widehat{u}_1^i,\widehat{u}_2^i,...)$ where $\widehat{u}^i(t)$ is a mapping from the information set $\widehat{H}(t)$ to the set $\mathscr{P}(\widehat{\mathbf{A}}^i)$ of probabilities over $\widehat{\mathbf{A}}^i$. Let $\widehat{U}^i$ denote the set of policies for agent i. Let $\widehat{U}:=\prod_{j=1}^N\widehat{U}^i$ be the set of *multi-strategies*.
- Each initial distribution $\widehat{\beta}$ over the state space and a multi-strategy $\widehat{u}$ induce a unique probability measure over the space of histories. These define the distribution of the state and action stochastic processes $\{\widehat{X}(t),\widehat{A}(t)\}$.

- A policy $\widehat{u}^i$ for agent i is said to be stationary if $u^i(t)$ depend only on the state $\hat{x}$ at time t, and is the same for all t. A multistrategy $\widehat{u}$ is stationary if each of its N components $\widehat{u}^i$ is stationary.

Let $\hat{q}^i$ be a stationary multi-strategy in the transformed MDP. It chooses at a state $\widehat{x} = (x, a)$ a mapping g^i with probability $\hat{q}^i(g^i)$. We show how to transform this into a quasi-stationary policy u^i in our original problem so as to obtain the same distribution on the processes $\{X(t), A(t)\}$.

In the original problem we need to specify for every player i and every action a the probability of choosing action a given the available information.

Lemma 3.1. *Set for each i and each $(x(t-1), a(t-1), x^i(t))$,*

$$u^i(x(t-1), a(t-1), x^i(t))(a) = \sum_{k=1}^{\widehat{K}} p_k 1\{g_k^i(x^i(t)) = a^i\} \tag{1}$$

where

$$p_k = [\hat{q}^i(x(t-1), a(t-1))]_k$$

is the probability under $\hat{q}^i(x(t-1), a(t-1))$ of choosing g_k. Then the state and action processes $(X(t), A(t))$ in the transformed model have the same distribution as $\widehat{X}(t+1)$.

Proof.- We establish the claim by induction on t: we show that (1) implies A1(t) for all t where

- **A1(t):** $(X(s), A(s)\ ; s \leq t)$ in the PO-MDP model has the same distribution as $(\widehat{X}(s+1)\ ; s \leq t)$ in the transformed model

It holds for $t = 1$. Assume it holds for some $(t-1)$. In the transformed model we have

$$\mathbb{P}\{A^i(t) = a^i | x(t-1), a(t-1), x^i(t)\} = \sum_{k=1}^{\widehat{K}} p_k 1\{g_k^i(x^i(t)) = a^i\}.$$

for all $(x(t-1), a(t-1), x^i(t))$. By choosing $u(t)$ according to (1) it follows that A1(t) holds. This establishes the proof. $\diamond$

Eq. (1) can be used also in the opposite direction.

Lemma 3.2. *For any agent i and stationary policy u^i, there exists a policy $\hat{q}^i$ in the transformed model that satisfies* (1)*. With this choice we then have A1(t) for all t.*

Proof.- Choose an agent i and a pair $(x(t-1), a(t-1))$. We have to show that there exists a measure q over $\widehat{\mathbf{A}}$ such that (1) holds for all $x^i(t)$ and $a^i(t)$. We first note that the set of micro randomized stationary policies is clearly a compact convex set. Since it is compact, by the Krein-Milman theorem it is the convex hull of its extreme points. A micro stationary policy that is not pure is obviously not an extreme point. The extreme points are therefore the micro pure stationary policies. Therefore there exists p such that (1) holds. $\diamond$

Definition 3.1. *An MDP is said to be ergodic if under any pure stationary policy the state process is an ergodic Markov chain.*

4 Applications: Markov Games and Team Problems

4.1 Markov Games

Theorem 4.1. *Consider the Markov game where either*

- *all costs $C^{i,k}$ are discounted, or*
- *where (i) for some players all the costs discounted, (ii) for the others all costs are expected average costs, and where (iii) the full information version of the original MDP (see Definition 2.1) is ergodic.*

Assume that the Slater condition holds (Definition 2.2) for the original PO-MDP. Then (i) there exists a stationary equilibrium in the transformed MDP, (ii) there exists a quasi-stationary equilibrium in the original PO-MDP which can be computed by applying the transformation (1) to any stationary equilibrium in the transformed problem.

Proof. It is easily seen that if the full information version of the original MDP is ergodic then so is the transformed MDP. Then (i) follows from [1] and (ii) from combining this with Lemma 3.1, 3.2. ◇

Remark 4.1. *The equivalence for the team problem without the constraints was known long ago. It is the randomization that are needed in the case of stochastic game (with or without constraints) and in the case of team problem with constraints that make the equivalence result a non trivial extension of the tean problem without constraints.*

4.2 Comments on the Team Problem

The team problem with the one-step delayed sharing information has been well studied [5, 2, 3, 4] in the absence of constraints. It is tempting to think that the case with constraints is a special case of the game problem with constraints, obtained by taking the cost of all players to be the same. We call this the corresponding game problem.

Relation to the game problem
An equilibrium of this game need not be, however, a solution for the team problem. The reason is that in the team problem we seek for a cooperative solution which means that not only it cannot be improved by a deviation of a single agent (as is the case in the equilibrium notion in the game); it cannot be improved by any simultaneous deviation of any number of players. If we just consider the team problem as a game with equal costs to minimize, then a Nash equilibrium for that game need not to be a solution to the initial team problem. However any solution to the team problem is a Nash equilibrium to that game.

However, assume that we restrict the team problem to some class of policies $\widetilde{U}$, and (i) the corresponding game has a unique equilibrium and (ii) there exists an optimal solution to the team problem restricted to $\widetilde{U}$. Then the equilibrium of the corresponding game is an optimal solution to the team problem.

The correlation problem
Next we introduce another problem that arises in the team problem and not in the game problem and is due to the constraints; it is a generic problem that is not directly related to the delay sharing information.

Consider two agents: A and B, each having two pure actions: a and b, and the cost functions given by Table 1:

		B: a	B: b
A	a	0	0
	b	0	-1

Table 1: The cost C^0 to be minimized

		B: a	B: b
A	a	-1	100
	b	100	1

Table 2: Payoffs of the players

The goal is to minimize the cost or equivalently to maximize the probability of both choosing action b. We introduce constraints expressed as cost associated to combination of actions. The cost for each combination of actions is given by Table 2. Assume that each agent has an upper-bound of 0 on the expected cost C^1. If agent A uses action a with probability p and agent B uses action a with probability q then the constraint has the form

$$C^{A,1}(p,q) = C^{B,1}(p,q) = 100(p\bar{q} + \bar{p}q) - pq + \bar{p}\bar{q}.$$

We are thus faced with the problem of maximizing $\bar{p}\bar{q}$ subject to $C^{i,1}(p,q) \leq 0$. The value is $0,000025$ and the optimal policies are $p = 0,995$ and $q = 0,995$. This value is much smaller than $1/2$ which would be obtained if the agents could correlate their actions; the optimal solution would then consists of playing (a,a) with probability half and (b,b) with probability half.

The solution of the team problem depends on whether correlation is possible or not between the agents. By transforming a Markov team problem with the delayed sharing (or with other) information structure into an equivalent full information problem we can use the theory of a single user Markov Decision Processes in the absence of constraints since in that case it is known that optimal pure stationary policies exist and therefore no correlation is required. In presence of constraints optimal stationary policies of the equivalent single controller problem with full information need randomization. When going back to the original PO-MDP, it may not be possible to perform this randomization as it may require correlation between the agents. If the case of delayed sharing framework, it is natural to assume that it is impossible to perform joint randomization and have the result available instantaneously.

Note: since one can observe the actions of the other, we can use the delayed knowledge of the outcome of a randomization as a correlating tool.

5 Power Control Problem with Two Players

We consider a decentralized Markovian control problem in the context of wireless communications, namely the uplink power control problem over interference channels with infinite horizon.

The information available to a mobile at any time t follows the one-step delay sharing information pattern. Such kind of decentralized Markovian team problems were considered by Hsu and Marcus [2] and they gave a policy iteration algorithm to solve the team problem.

We formulate the problem as a stochastic game problem with one-step delay sharing information pattern.

For simplicity, we consider two mobiles and we denote $x(t) = (x^1(t), x^2(t))$ to the channel state configuration of the mobiles at time t, where $x^i(t)$ may have a good state channel G or a bad state channel B. The joint state configuration of both mobiles at each state is given by Table 3.

Mobile 1 \ Mobile 2	state G	state B
state G	(G,G)	(G,B)
state B	(B,G)	(B,B)

Table 3: The joint state configuration

Mobile 1 \ Mobile 2	action L	action U
action L	(L,L)	(L,U)
action U	(U,L)	(U,U)

Table 4: The joint action configuration

We assume that the channel states of each mobile follow a Markov chain with transition probabilities given by the matrix:

$$\mathbf{P}_t^i = \left[P_{ab}^i = \mathbb{P}\{x^i(t+1) = b) | x^i(t) = a\}\right]$$

where $a, b \in \{G, B\}$. We assume that the transition probability for each mobile is stationary.

Consider that the possible actions for each mobile at each stage is to transmit to the base station with high power transmission U, or with low power transmission L. The joint action configuration for each scenario is given by Table 4

The configuration of states of both mobiles is an element of the set:

$$\begin{aligned}\widehat{\mathbf{X}} =& \{GGUU, GGUL, GGLU, GGLL, GBUU, GBUL, GBLU, GBLL,\\ & BGUU, BGUL, BGLU, BGLL, BBUU, BBUL, BBLU, BBLL\}.\end{aligned}$$

Let SINR^i denote the Signal to Interference plus Noise Ratio corresponding to the signal received from mobile i at the base station. Each mobile has two pure strategies: transmit with a high power transmission U or with a low power transmission L.

The SINR^i can be computed on each scenario according to the actions of each mobile. For example if mobile 1 has a good channel state G and chooses to transmit with a high power transmission U and mobile 2 has a bad channel state B and chooses to transmit with a low power transmission L, then

$$\text{SINR}^1 = \frac{h_G^1 P_U^1}{N_0 + h_B^2 P_L^2} \quad \text{and} \quad \text{SINR}^2 = \frac{h_B^2 P_L^2}{N_0 + h_G^1 P_U^1}$$

Under many modulation schemes the probability of a successful transmission of a packet is known to be a monotone increasing function of the SINR [6]. We thus assume that mobile i has a successful probability given by $f_i(\text{SINR}_i)$.

As examples of success probability as a function of the SINR, we have the following expressions for the bit error probability as a function of the SINR [7] for the GFSK modulation:

$$p_e(\text{SINR}) = \frac{1}{2}\exp(-\frac{1}{2}\text{SINR})$$

In the absence of redundancy this gives the following expression for f of a packet of N bits provided that the bit loss process is independent

$$f(\text{SINR}) = (1 - p_e(\text{SINR}))^N$$

6 Numerical Results

We next study a numerical example and examine the game problem. The solutions were obtained by applying the Lemke-Howson algorithm [8] to the stochastic equivalent full information stochastic game.

We use the following values of the parameters: $h_G^1 = h_G^2 = 20$ dB, $h_B^1 = h_B^2 = 10$ dB $P_U^1 = P_U^2 = 0.6$ Watt, $P_L^1 = P_L^2 = 0.3$ Watt, where the transition probability matrix of the channel states for each player is given by

$$P_1 = P_2 = \begin{bmatrix} 2/3 & 1/3 \\ 2/5 & 3/5 \end{bmatrix}$$

Note that even if the state transitions are not controlled by the actions of the players in the original game, for the transformed game with the global state defined as $\hat{x}(t) = (x(t-1), a(t-1)$ they are controlled. We find that for a game lasting 100 stages, both the discounted payoff criteria with a discounted factor of $\alpha = 0.5$, and the average payoff criteria permit the same equilibrium strategy: to always transmit using the highest power no matter what the state of the transformed game. The values for the discounted value problem and the average value problem are

$$C_\alpha^1 = \begin{pmatrix} 28.4278 \\ 27.3728 \\ 25.4023 \\ 21.7732 \end{pmatrix} \quad C_\alpha^2 = \begin{pmatrix} 28.4278 \\ 25.4023 \\ 24.3728 \\ 21.7732 \end{pmatrix} \quad C_{ea}^1 = \begin{pmatrix} 0.5686 \\ 0.4875 \\ 0.5080 \\ 0.4355 \end{pmatrix} \quad C_{ea}^2 = \begin{pmatrix} 0.5686 \\ 0.5080 \\ 0.4875 \\ 0.4355 \end{pmatrix}$$

6.1 An equivalent game with no delay

We solve the power control game with the same parameters without the one step delayed sharing information pattern. Its a standard stochastic game with the state space $\mathbf{X} = \{GG, GB, BG, BB\}$ Using the individual channel state transition probability matrices P_1 and P_2 of the two players, we can find the joint transition probability matrix of the states of this stochastic game. Notice that in this setting, the transitions are not controlled by the actions of any player. Thus we only need to consider the immediate rewards at each state to compute the equilibrium strategies at that state. The game lasts 100 stages. Similar to the delayed sharing case, both the discounted payoff criteria with a discounted factor of $\alpha = 0.5$, and the average payoff criteria

permit the same equilibrium strategy for the players: to always transmit using the highest power regardless of the state of the game. The values for the discounted value problem and the average value problem are

$$C_\alpha^1 = \begin{pmatrix} 28.3902 \\ 23.0129 \\ 26.8735 \\ 21.7041 \end{pmatrix} \quad C_\alpha^2 = \begin{pmatrix} 28.3902 \\ 26.8735 \\ 23.0129 \\ 21.7041 \end{pmatrix} \quad C_{ea}^1 = \begin{pmatrix} 0.5678 \\ 0.4603 \\ 0.5375 \\ 0.4341 \end{pmatrix} \quad C_{ea}^2 = \begin{pmatrix} 0.5678 \\ 0.5375 \\ 0.4603 \\ 0.4341 \end{pmatrix}$$

Discussion: We can have non-trivial equilibrium strategies for the players for both the problems if we include a cost for power transmission in the payoffs which is proportional to the level of power chosen. For finding the mixed strategies of the players at each state we again use the Lemke-Howson algorithm used for solving bimatrix games. Since in the problem with delay and that with no delay, the equilibrium strategy is the same i.e. always transmit with high power, both the problems will give the same *realized* value at the end. But since in the problem without delay the players have more information, they have a better estimate of the expected value.

7 Conclusions

In this paper we studied constrained Markov games in which each agent has some local information on its own environment and it is assumed that the information of the others agents is available to that agent after one stage.

We solve the team problem as well as the stochastic non-cooperative game on this setting and solve a power control problem in a stochastic non-cooperative game context.

References

[1] E. Altman and A. Shwartz, "Constrained Markov Games: Nash Equilibria", *Annals of Dynamic Games*, vol. 5, Birkhäuser, V. Gaitsgory, J. Filar and K. Mizukami, editors, pp. 213-221, 2000.

[2] K. Hsu and S. I. Marcus, "Decentralized Control of Finite State Markov Processes," 19th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes Vol. 19, pp. 143–148, Dec. 1980.

[3] S. M. Ross, "Introduction to Stochastic Dynammic Programming," Academia Press, New York, 1983.

[4] P. Varaiya and J. Walrand, "On delayed sharing patterns," IEEE Transactions on Automatic Control, Vol. 23, pp. 443–445, June 1978.

[5] J. W. Grizzle, S. I. Marcus, K. Hsu, "Decentralized control of a multiaccess broadcast network," 20th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes, Vol. 20, pp. 390–391, Dec. 1981.

[6] P. Soldati, "Cross-Layer Optimization of Wireless Multi-hop Networks," Royal Institute of Technology (KTH), 2007 (ISBN 978-7178-711-8).

[7] J. G. Proakis, "Communication System Engineering", Prentice Hall International Editions 1994.

[8] Lemke, C. E., Howson, Jr., J. T.: Equilibrium Points of Bimatrix Games, Journal of the Society of Industrial and Applied Mathematics, 12, pp. 413-423, 1964.