Mathematical Methods in Operations Research 2005
Special issue in honor of *Arie Hordijk*

Zero-sum constrained stochastic games with independent state processes

Eitan Altman[1⋆], **Konstantin Avrachenkov**[1], **Richard Marquez**[2⋆], **Gregory Miller**[3⋆⋆]

[1] INRIA, Centre Sophia-Antipolis, 2004 Route des Lucioles, B.P.93, 06902 Sophia-Antipolis Cedex, France
[2] Postgrado en Ingeniería de Control, Facultad de Ingeniería, Universidad de Los Andes, Mérida 5101, Venezuela
[3] Probability Theory Department, Applied Mathematics and Physics Faculty of Moscow Aviation Institute: 4, Volokolamskoe sh., GSP-3, Moscow A-80, Russia 125993 .

Received: January 2005 / Revised version: March 2005

Abstract We consider a zero-sum stochastic game with side constraints for both players with a special structure. There are two independent controlled Markov chains, one for each player. The transition probabilities of the chain associated with a player as well as the related side constraints depend only on the actions of the corresponding player; the side constraints also depend on the player's controlled chain. The global cost that player 1 wishes to minimize and that player 2 wishes to maximize, depend however on the actions and Markov chains of both players. We obtain a linear programming (LP) formulations that allows to compute the value and saddle point policies for this problem. We illustrate the theoretical results through a zero-sum stochastic game in wireless networks in which each player has power constraints.

1 Introduction

Zero-sum stochastic games have been an active area of research and a useful tool in many applications. Yet, it is well known that identifying saddle point policies even in zero-sum stochastic games with finite state and action spaces is hard. Unlike the situation in Markov Decision Processes (MDPs)

Send offprint requests to: Eitan Altman

⋆ *The work of these authors has been partially supported by an ECOS-Nord grant for French-Venezuelan collaboration*

⋆⋆ *The work of this author has been supported by INTAS (YSF 04-83-3623)*

in which stationary optimal policies are known to exist (under suitable conditions), and unlike the situation in constrained MDPs (CMDPs) with a multichain structure, in which optimal Markov policies exist [7,10], we know that saddle point policies in stochastic games need in general to depend on the whole history [5]. This difficulty motivated researchers to search for various possible structures of stochastic games in which saddle point policies exist among stationary or Markov strategies and are easier to compute.

In this paper we consider two CMDPs, where the transition probabilities of each one are controlled by just one of the players who has information only on the history of the CMDP it controls. The cost is determined jointly by the states and actions of both CMDPs. For the expected average cost, we obtain linear programs (LPs) that allow us to compute the value and saddle point policies for both the unichain as well as the multi-chain ergodic structure. We illustrate the theoretical results through a zero-sum stochastic game in wireless networks in which each player has power constraints.

Related work Several papers have already dealt with constrained stochastic games. An important class of zero-sum stochastic games that can be solved using LPs has been introduced in parallel in [8–10,12]. In those games only one player controls the transition probabilities (but both players determine the cost through their actions). The existence of a stationary Nash equilibrium in non zero-sum constrained stochastic games has been established in [3] under a Slater-type condition. A highly non-stationary saddle-point was obtained in [11] for constrained stochastic games with expected average costs. Our work as well as the class of games we study are based on [6], who introduced LPs for obtaining the saddle point policies and the value of stochastic games with sample average costs and a unichain structure.

2 The model

We consider two MDPs characterized by the triplet $(\mathbf{I}^k, \mathbf{A}^k, \mathbf{P}^k)$, $k = 1, 2$, where $\mathbf{I}^k, \mathbf{A}^k$ stand for the finite state and action spaces, respectively, and where $\mathbf{P}^k = \{P^k_{iaj}\}$ stands for the corresponding transition probabilities; P^k_{iaj} is the probability that player k's state moves from i to j if the player chose action a. At state $i \in \mathbf{I}^k$, the set of actions available to player k is $\mathbf{A}^k(i)$. Let $\mathbf{K}^k$ stand for the set of $(i, a), i \in \mathbf{I}^k, a \in \mathbf{A}^k(i)$.

Define a history h_n in MDP k as $h_n = (i_0, a_0, i_1, a_1, ..., i_{n-1}, a_{n-1}, i_n)$ where $i_\ell \in \mathbf{I}^k$, $a_\ell \in \mathbf{A}^k(i_\ell)$, $\ell = 0, 1, 2, ...$ A player k strategy u is a sequence $(u_0, u_1, ...)$ where u_ℓ is a probability measure over $\mathbf{A}^k(i_\ell)$ conditioned on h_n. Note that player k strategies do not depend on the realizations of the cost. If they were allowed to depend on these then a player could use the costs to estimate the state and actions of the other player.

Player k has m_k *side constraints* of the form $D^k_s(\beta^k, u^k) \leq \xi^k_s$, $s = 1, 2, ..., m_k$, where β^k is a probability distribution of the initial state of player k, and where ξ^k_s are some constants. Let $\beta = (\beta^1, \beta^2)$. We shall write the side constraints in the vector form

$$D^k(\beta^k, u^k) \leq \xi^k, \qquad k = 1, 2. \tag{1}$$

Denote by U^k the set of all strategies (also called policies) for player k, and let U_c^k be the set of strategies of player k that satisfy (1). Let $U_c := (U_c^1, U_c^2)$. We shall assume throughout

$$U_c^k \text{ is non empty, } k = 1, 2. \tag{2}$$

Let $U^k(S)$ and $U^k(M)$ be the set of stationary and of Markov policies, respectively, of player k, and set $U_c^k(S)$ and $U_c^k(M)$ to be the corresponding subsets that satisfy (1). A stationary policy $u \in U^k(S)$ is identified with a set of probability functions denoted (with some abuse of notation) as $u(\cdot|i)$, over the actions $\mathbf{A}^k(i)$. For all $i \in \mathbf{I}^k$ $u(a|i)$ is then the probability of choosing action a if the state is i. A Markov policy $u \in U^k(M)$ is identified with a set of probability functions denoted (with some abuse of notation) as $u(\cdot, n|i)$, over the actions $\mathbf{A}^k(i)$. For all $i \in \mathbf{I}^k$ and integer n $u(a, n|i)$ is the probability of choosing action a at time n if the state is i.

We further introduce the cost $C(\beta^1, \beta^2, u^1, u^2)$ where $u^k \in U^k$ which player 1 wishes to minimize and which player 2 wishes to maximize. We seek a saddle point couple $(u^*, v^*) \in U_c$, i.e. a policy for each player such that

$$V := \inf_{u \in U_c^1} C(\beta, u, v^*) = C(\beta, u^*, v^*) = \sup_{v \in U_c^2} C(\beta, u^*, v) \tag{3}$$

Next we specify what C and D will stand for.
Let $c(i, j, a, b)$ correspond to the immediate cost for player 1 when she is at state i and chooses action a, and when player 2 is at state j and chooses action b.
Let $d_s^k(i, a)$ be an immediate cost related to the sth side constraint of player k, when she is at state i and chooses action a.

The expected average cost. We define the expected average costs as

$$C_{ea}(\beta, u^1, u^2) = \limsup_{t \to \infty} \frac{1}{t} \sum_{n=0}^{t-1} E_\beta^u c(I_n^1, I_n^2, A_n^1, A_n^2), \tag{4}$$

$$D_{ea}^{k,s}(\beta^k, u^k) = \limsup_{t \to \infty} \frac{1}{t} \sum_{n=0}^{t-1} E_{\beta^k}^{u^k} d_s^k(I_n^k, A_n^k).$$

Remark 1 It follows from the proof of [2, Theorem 2.8] that our results are unchanged if we replace the lim sup in (4) by lim inf.

3 The unichain case

We consider the expected average cost with a unichain structure: under any pure stationary policy[1] u^k for player k, the corresponding Markov chain has a single ergodic class.

[1] A pure policy is one that does not use any randomizations.

We solve the problem

$$\inf_{u^1 \in U_c^1} \sup_{u^2 \in U_c^2} C_{ea}(\beta, u_1, u_2). \tag{5}$$

To do that, we first fix a stationary policy u^1 for player 1, then the player 2 is faced with a CMDP for which we know that an optimal policy exists within the stationary policies, see [2, Theorem 2.8][2]. So for player 2 we find the optimal value of the cost (4) $\Theta_{ea}^*(u^1)$ for a fixed stationary u^1. We then solve the optimization problem $\inf_{u^1 \in U_c^1(S)} \Theta_{ea}^*(u^1)$. Later we shall show that indeed one can restrict to stationary policies for player 1 without loss of optimality.

3.1 Some definitions

Define for a fixed $u \in U^k$, $x_{ea}^{k,t}(\beta^k, u) = \{x_{ea}^{k,t}(\beta^k, u; i, a)_{(i,a)\in \mathbf{K}^k}\}$ where

$$x_{ea}^{k,t}(\beta^k, u; i, a) := \frac{1}{t} \sum_{n=0}^{t-1} P_{\beta^k}^u(I_n^k = i, A_n^k = a), \quad (i,a) \in \mathbf{K}^k$$

($P_{\beta^k}^u$ is the unique probability measure corresponding to a policy $u \in U^k$ for an initial distribution β^k over the states). The set $X_{ea}^k(\beta^k, u)$ defined as the set of accumulation points of $x_{ea}^{k,t}(\beta^k, u)$ is known as a set of occupation measures corresponding to a strategy u^k and an initial distribution β^k.

Let $\mathbf{Q}_{ea}^k$ be the set of vectors $\rho \in \mathbf{R}^{|\mathbf{K}^k|}$ satisfying

$$\mathbf{Q}_{ea}^k = \begin{cases} \displaystyle\sum_{(j,a)\in\mathbf{K}^k} \rho(j,a)(\delta_i(j) - \mathbf{P}_{jai}^k) = 0, \quad \forall i \in \mathbf{I}^k, \\ \displaystyle\sum_{(j,a)\in\mathbf{K}^k} \rho(j,a) = 1, \\ \rho(j,a) \geq 0, \quad \forall (j,a) \in \mathbf{K}^k, \end{cases} \tag{6}$$

where $\delta_i(j)$ is the indicator which is equal to one if $i = j$ and is zero otherwise. It should be noted that any ρ satisfying the above constraints is a probability measure.

Define further

$$\mathbf{Q}_{ea,c}^k := \left\{ \rho \in \mathbf{Q}_{ea}^k : \sum_{(j,a)\in\mathbf{K}^k} \rho^k(j,a) d_s^k(j,a) \leq \xi_s^k, s = 1, ..., m_k \right\}$$

Note that $\mathbf{Q}_{ea,c}^k$ is non-empty due to Assumption (2), see [7].

It is shown in [1,7] that the set of achievable occupation measures achieved by all feasible strategies $u^k \in U_c^k$ equals to those achieved by stationary policies and further equals to the set $\mathbf{Q}_{ea,c}^k$.

[2] This reference implies the sufficiency of stationary policies for both the cases of maximizing as well as for minimizing C_{ea} subject to side constraints.

For a given probability measure ρ over $\mathbf{K}^k$, we define the stationary policy $w^k(\rho^k)$ as

$$w_i^k(a,\rho) = \frac{\rho(i,a)}{\sum_{a'\in\mathbf{A}(i)} \rho(i,a')}, \qquad a \in \mathbf{A}^k(i), \tag{7}$$

whenever the denominator is non-zero (when it is zero, $w^k(\rho)$ is chosen arbitrarily). Here $w_i^k(a,\rho)$ is the probability that player k will choose action a at state i according to this stationary policy.

Let u be a policy for which for all i and a, $P_\beta^u(I_n^k = i, A_n^k = a)$ has a limit which we denote by $\pi_{ea}^k(\beta^k, u; i, a)$. For policies with this property we have

$$\pi_{ea}^k(\beta^k, u; i, a) = \lim_{t\to\infty} x_{ea}^{k,t}(\beta^k, u; i, a).$$

These include in particular the stationary policies. If $\pi^k(u) = \{\pi^k(u;i)\}_{i\in\mathbf{I^k}}$ is the unique steady state distribution of the Markov chain induced by a stationary policy $u \in U^k$, then

$$\pi_{ea}^k(\beta^k, u; i, a) = \pi^k(u; i)u(a|i),$$

which is independent of the initial distribution β^k.

3.2 Player 2

We fix a stationary policy u^1 for player 1. Then player 2 is faced with a standard CMDP. It follows from [1] that the optimal value for player 2 *among all policies* U_c^2 is given by the value of the following LP:

$$\text{Find} \quad \Theta_{ea}^*(u^1) := \max_{\rho^2\in\mathbf{Q}_{ea,c}^2} \sum_{(j,a)\in\mathbf{K}^2} \rho^2(j,a)c(u^1; j, a) \tag{8}$$

$$\text{where} \quad c(u^1; j, b) := \sum_{(i,a)\in\mathbf{K}^1} \pi_{ea}^1(\beta^1, u^1; i, a)c(i, j, a, b), \quad u^1 \in \mathbf{U}^1. \tag{9}$$

Moreover, $w^2(\rho^2)$ is an optimal stationary policy for player 2 in this CMDP [1]. Hence the above LP allows us to obtain a best response of player 2 against a stationary policy of player 1.

We shall also use the dual LP. Its decision variables are ψ^2, $\phi^2(i), i \in \mathbf{I}^2$ as well as the m-dimensional non-positive vector $\lambda^2 \in \mathbf{R}_-^{m_2}$ (ψ^2 will correspond to the value of the expected average problem for fixed stationary u^1 and for an immediate reward of $c(u^1; j, b)+ < \lambda^2, d^2(j, b) >$, and λ^2 will correspond to Lagrange multipliers related to the side constraints of player 2). With $< \cdot, \cdot >$ denoting the scalar product, we have:

$$\Theta_{ea}^*(u^1) := \min_{\psi^2,\phi^2,\lambda^2} \psi^2 - \sum_{s=1}^{m_2} \lambda_s^2 \xi_s^2 \quad \text{subject to} \tag{10}$$

$$\phi^2(j) + \psi^2 \ \geq \ c(u^1; j, b)+ < \lambda^2, d^2(j, b) >$$

$$+\sum_{\ell\in\mathbf{I}^2}\mathbf{P}^2_{jb\ell}\phi^2(\ell), \quad \forall(j,b)\in\mathbf{K}^2 \tag{11}$$

$$\lambda^2_s \le 0, \qquad s=1,\dots,m_2.$$

3.3 Player 1

It follows from the previous subsection that Player 1 is faced with the optimization problem: $\inf_{u^1\in U^1_c(S)}\Theta^*_{ea}(u^1)$ where $\Theta^*_{ea}(u^1)$ is given in (10). It is seen from (9), however, that the dependence on u^1 is only through occupation measure $X^1_{ea}(\beta^1,u^1)$. We know from [1] that

$$\left\{X^1_{ea}(\beta^1,u^1): u^1\in U^1_c(S)\right\}=\mathbf{Q}^1_{ea,c} \tag{12}$$

Moreover, for any $\rho^k\in\mathbf{Q}^k_{ea}$, the stationary policy defined in (7) provides

$$\pi^k_{ea}(\beta^k,w(a,\rho);i,a)=\rho^k(i,a),$$

see [1]. Hence, the following LP provides the value for problem (5), when player 1 restricts to stationary policies (we shall show later that such restriction is without loss of optimality).

$$\mathbf{LP_{ea}}: \qquad \text{Find } \mathcal{C}^*_{ea} := \min_{\psi^2,\phi^2,\lambda^2,\rho^1\in\mathbf{Q}^1_{ea,c}} \psi^2-\sum_{s=1}^{m_2}\lambda^2_s\xi^2_s \quad \text{subject to}$$

$$\phi^2(j)+\psi^2 \ \ge \sum_{(i,a)\in K^1}\rho^1(i,a)c(i,j,a,b)+<\lambda^2,d^2(j,b)>$$

$$+\sum_{\ell\in\mathbf{I}^2}\mathbf{P}^2_{jb\ell}\phi^2(\ell), \quad \forall(j,b)\in\mathbf{K}^2$$

$$\lambda^2_s \le 0, \qquad s=1,\dots,m_2.$$

Moreover, for any $\rho^*_1\in\mathbf{Q}^1_{ea,c}$ for which $(\psi^2,\phi^2,\lambda^2,\rho^1)$ achieves the above minimization, the corresponding $w^1(\rho^*_1)$ (defined in (7)) provides a policy for Player 1 which is the best among stationary policies.

Next we consider the problem

$$\sup_{u^2\in U^2_c(S)}\inf_{u^1\in U^1_c} C_{ea}(\beta,u_1,u_2). \tag{13}$$

Introduce the following LP:

$$\mathbf{DP_{ea}}: \text{Find } \mathbf{C}^*_{ea} := \max_{\psi^1,\phi^1,\lambda^1,\rho^2\in\mathbf{Q}^2_{ea,c}} \psi^1-\sum_{s=1}^{m_1}\lambda^1_s\xi^1_s \quad \text{subject to}$$

$$\phi^1(i)+\psi^1 \ \le \sum_{(j,b)\in K^2}\rho^2(j,b)c(i,j,a,b)+<\lambda^1,d^1(i,a)>$$

$$+\sum_{\ell\in\mathbf{I}^1}\mathbf{P}^1_{ia\ell}\phi^1(\ell), \quad \forall(i,a)\in\mathbf{K}^1$$

$$\lambda^1_s \ge 0, \qquad s=1,\dots,m_1.$$

By the same arguments as before, for any $\rho_2^* \in \mathbf{Q}^2_{ea,c}$ for which $(\psi^1, \phi^1, \lambda^1, \rho^2)$ achieves the maximal value of $\mathbf{C}^*_{ea}$, the corresponding $w^2(\rho_2^*)$ (defined in (7)) provides a policy for Player 2 which is the best among stationary policies for problem (13).

Due to the duality of the $\mathbf{LP_{ea}}$ and $\mathbf{DP_{ea}}$, we conclude that $\mathcal{C}^*_{ea} = \mathbf{C}^*_{ea}$ and thus that $(w^1(\rho_1^*), w^2(\rho_2^*))$ are a saddle point for (3). Indeed,

$$\begin{aligned} C^*_{ea} &= \sup_{u^2 \in U_c^2(S)} \inf_{u^1 \in U_c^1} C_{ea}(\beta, u_1, u_2) \leq \sup_{u^2 \in U_c^2} \inf_{u^1 \in U_c^1} C_{ea}(\beta, u_1, u_2) \\ &\leq \inf_{u^1 \in U_c^1} \sup_{u^2 \in U_c^2} C_{ea}(\beta, u_1, u_2) \leq \inf_{u^1 \in U_c^1(S)} \sup_{u^2 \in U_c^2} C_{ea}(\beta, u_1, u_2) = \mathcal{C}_{ea} \end{aligned}$$

which implies that all inequalities hold with equality.

4 The expected average cost: multichain case

Following [7,10], we introduce the class of policies $U^k(1)$ which are all policies u of player k for which the set $X^k_{ea}(\beta^k, u)$ is a singleton. Define $\tilde{U}^k(M^*) = U^k(1) \cap U^k(M)$. We further define $U^k(M^*)$ as the subset of $\tilde{U}^k(M^*)$ of policies for which $P^u_{\beta^k}(I^k_n = i, A^k_n = a)$ has a single limit. It follows from [7, Theorem 2] that the set of all occupation measures achieved by strategies $u^k \in U^k(M^*)$ is equal to the set of all occupation measures achieved by all policies.

Define $\mathbf{Q}^k_{eam,c}(\beta^k)$ as the set of couples (ρ, z) satisfying

$$\begin{cases} \displaystyle\sum_{(j,a) \in \mathbf{K}^k} \rho(j,a)(\delta_i(j) - \mathbf{P}^k_{jai}) = 0, & \forall i \in \mathbf{I}^k \\ \displaystyle\sum_{a \in \mathbf{A}^k(i)} \rho(i,a) + \sum_{(j,a) \in \mathbf{K}^k} z(j,a)(\delta_i(j) - \mathbf{P}^k_{jai}) = \beta^k_i, & \forall i \in \mathbf{I}^k \\ \rho(j,a) \geq 0, \quad z(j,a) \geq 0 & \forall (j,a) \in \mathbf{K}^k \\ \displaystyle\sum_{(j,a) \in \mathbf{K}^k} \rho(j,a) d^k_s(j,a) \leq \xi^k_s, & s = 1, ..., m_k, \end{cases}$$

For the meaning of the new decision variable z, see [4]. For any policy $u \in U^k(M^*)$ of player k, the other player is faced with a CMDP for which there exists an optimal policy within $U^l(M^*)$ $(l \neq k)$ that can be computed as in [7]. This follows from the same arguments as in the proof of [2, Theorem 2.8]. Thus, for a fixed $u^1 \in U^1(M^*)$, the value of this CMDP is given by that of the LP (8) with $\mathbf{Q}^k_{eam,c}(\beta^k)$ replacing $\mathbf{Q}^k_{ea,c}(\beta^k)$. Its dual is

$$\Theta^*_{ea}(u^1) := \min_{\psi^2, \phi^2, \lambda^2} < \beta^2, \psi^2 > - \sum_{s=1}^{m_2} \lambda^2_s \xi^2_s \quad \text{subject to}$$

$$\sum_{\ell \in \mathbf{I}^2} (\delta_j(\ell) - \mathbf{P}^2_{jb\ell}) \psi^2(\ell) \geq 0, \qquad \forall (j,b) \in \mathbf{K}^2$$

$$\phi^2(j) + \psi^2(j) \geq c(u^1; j, b) + < \lambda^2, d^2(j,b) > + \sum_{\ell \in \mathbf{I}^2} \mathbf{P}^2_{jb\ell} \phi^2(\ell), \quad \forall (j,b) \in \mathbf{K}^2$$

$$\lambda^2_s \leq 0, \qquad s = 1, \ldots, m_2,$$

where $c(u^1; j, b)$ is given in (9). To minimize $\Theta^*_{ea}(u^1)$ over $u^1 \in U^1_c(M^*)$, we have to solve

$$\begin{aligned}
\mathbf{LP_{eam}}(\beta) : \mathcal{C}^*_{ea} &:= \min_{\psi^2,\phi^2,\lambda^2,(\rho^1,z^1)\in\mathbf{Q}^1_{ea,c}(\beta^1)} < \beta^2, \psi^2 > - \sum_{s=1}^{m_2} \lambda^2_s \xi^2_s \quad \text{s.t.} \\
&\sum_{\ell\in\mathbf{I}^2} (\delta_j(\ell) - \mathbf{P}^2_{jb\ell})\psi^2(\ell) \geq 0, \qquad \forall (j,b) \in \mathbf{K}^2 \\
\phi^2(j) + \psi^2(j) \geq &\sum_{(i,a)\in\mathbf{K}^1} \rho^1(i,a)c(i,j,a,b) + < \lambda^2, d^2(j,b) > + \\
&\sum_{\ell\in\mathbf{I}^2} \mathbf{P}^2_{jb\ell}\phi^2(\ell), \quad \forall (j,b) \in \mathbf{K}^2 \\
\lambda^2_s \leq &\ 0, \qquad s = 1, \ldots, m_2.
\end{aligned}$$

For any optimal solution of the above LP, one can obtain from the variables (ρ^1, z^1) an optimal policy $u^1 \in U^1(M^*)$ for player 1, as it is done in [7]. The dual of the above LP then provides an optimal policy for player 2.

5 Examples in wireless communications

5.1 Example 1

We consider two mobile terminals and one base station. Mobile 1 seeks to transmit information to the base station. Mobile 2 has an antagonistic objective: to prevent or to jam the transmissions of mobile 1 to the base station. We consider a discrete time model. At each slot n, mobile k transmits a packet with power level p^k_n. The radio channel between mobile k and the base station is characterized by a Markov chain $\mathbf{M}^k$. The channel state of both mobiles are independent. The channel state of a mobile determines the power attenuation between the mobile and the base station. Denote by $h^k(\zeta)$ the attenuation of mobile k's power when at state $\zeta \in \mathbf{M}^k$. The throughput (the amount of bits per second) that mobile 1 can send to the base station at a given slot n is given by

$$T(\zeta^1, \zeta^2, p^1, p^2) = B \log_2 \left(1 + \frac{p^1 h^1(\zeta^1)}{N_0 + p^2 h^2(\zeta^2)}\right) \tag{14}$$

where B is a channel bandwidth, ζ^1, ζ^2 are the channel states and p^1, p^2 are the power levels. N_0 is a constant that stands for the thermal noise power at the receiver. The term $p^k h^k(\zeta^k)$ determines the power level received at the base station from mobile k. The term $\frac{p^1 h^1(\zeta^1)}{N_0 + p^2 h^2(\zeta^2)}$ is the ratio between the power received at the base station from mobile 1 and the total power of noise and interference. Eq. (14) is known as the Shannon capacity. It gives the least upper bound of the transmission rate that can be achieved with an error probability less than any $\epsilon > 0$, if we assume that the interference of player 2 at a slot is presented as a Gaussian white noise (this excludes the possibility of the receiver to decode the signal of Player 2 which, if successful, would have allowed to subtract it from the noise experienced by player 1).

Mobile k's action set is given by a discrete set $\mathbf{Pow}^k$, where $\mathbf{Pow}^k$ stands for the transmission power and is given by a finite ordered set $\mathbf{Pow}^k = (pow_1^k, ..., pow_\nu^k)$.

We assume that each mobile has a constraint on the power that it can use. We see that our formalism of independent state processes can indeed be used to model and solve this problem. In particular, the expected average cost seems to be appropriate if the mobiles have constraints on the expected average power consumption.

In example 1, no player controls the transitions. It might at first seem to be a special case of the framework of [9,12] where only one player controls the transitions and the other doesn't. But in fact, the framework of [9, 12] is different from ours since in the former, both players have full state information whereas in our framework, each player has its own information.

5.2 Numerical calculations for example 1

Let the radio channel between mobile k and the base station be characterized by a Markov chain $\mathbf{M}^k$ with states $\zeta_i = 0, \dots, N$, $N = 10$ and the following transition probabilities:

$$\begin{array}{ll} P_{i,i}^k = P_{i,i+1}^k = \frac{1}{2}, & i = 0; \\ P_{i,i}^k = P_{i,i-1}^k = P_{i,i+1}^k = \frac{1}{3}, & i = 2, \dots, N-1; \\ P_{i,i}^k = P_{i,i-1}^k = \frac{1}{2} & i = N. \end{array} \tag{15}$$

The transition probabilities (15) imply that at each slot the Markov chain with the same probability does one of the following: preserves its state, changes it to the next one or changes it to the previous one.

Each state of the Markov chain radio channel correspond to some level of the power attenuation:

ζ_i	0	1	2	...	10
$h^k(\zeta_i)$	0.0	0.1	0.2	...	1.0

Let mobile k's action set $\mathbf{Pow}^k$ be given by $\mathbf{Pow}^k = (0, ..., 10)$. The exact power of the signal of the mobile k is $\mathrm{P}_j^k = \mathrm{P}_0 \mathbf{Pow}_j^k$, where P_0 is some base value of the power, and $\mathbf{Pow}_j^k$ is one of the elements of $\mathbf{Pow}^k$. For the noise power at the receiver we will assume that $N_0 = \mathrm{P}_0 n_0$, where we take $n_0 = 1$. As the cost function depends only on the ratio between the power received from the first mobile and the total power of noise and interference of the second, we do not need to specify the exact value of the base power P_0.

Let the expected average power consumption of both mobiles be constrained by the following bound:

$$D^k_{ea}(u^k) \leq 5\mathrm{P}_0.$$

The immediate cost related to this constraint is

$$d^k(\xi^k_i, \mathrm{P}^k_j) = \mathrm{P}^k_j = \mathrm{P}_0 \mathbf{Pow}^k_j.$$

As the transition probabilities of both players do not depend on their strategies, the problem is of unichain case and thus has a solution within stationary policies.

On fig. 1 one can see the supports of the optimal policies of both players.

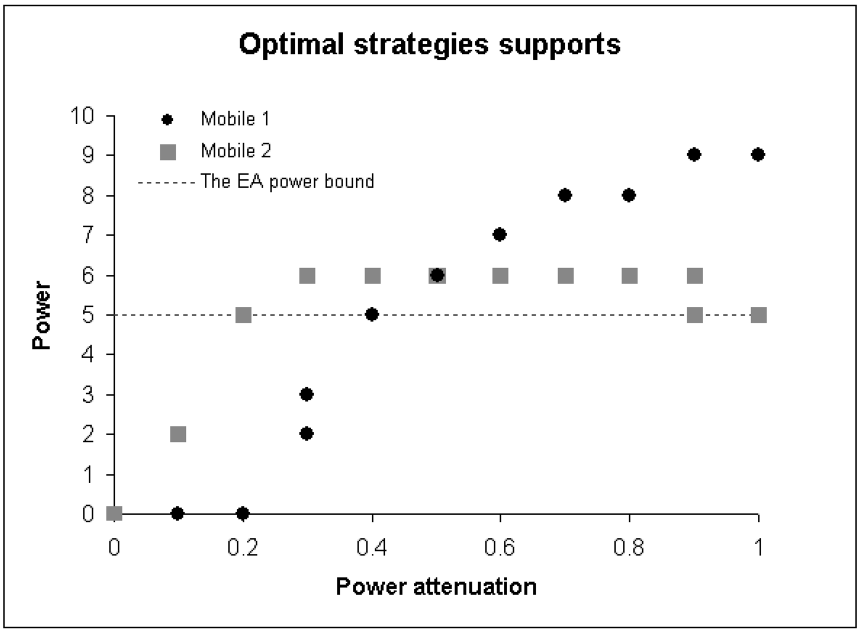

Fig. 1 Supports of the optimal policies.

The exact values of the stationary policies $w^k(h^k(\zeta_i), pow_j)$ are the following:

$$w^1(0,0) = w^1(0.1,0) = w^1(0.2,0) = w^1(0.4,5) = w^1(0.5,6) =$$
$$= w^1(0.6,7) = w^1(0.7,8) = w^1(0.8,8) = w^1(0.9,9) = w^1(1.0,9) = 1,$$
$$w^1(0.3,2) = \frac{1}{3}, \quad w^1(0.3,3) = \frac{2}{3};$$

$$w^2(0,0) = w^2(0.1,2) = w^2(0.2,5) = w^2(0.3,6) = w^2(0.4,6) =$$
$$= w^2(0.5,6) = w^2(0.6,6) = w^2(0.7,6) = w^2(0.8,6) = w^2(1.0,5) = 1,$$
$$w^2(0.9,5) = \frac{2}{3}, \quad w^2(0.9,6) = \frac{1}{3}.$$

The value of the expected average cost in this problem is $\mathcal{C}^*_{ea} = \mathbf{C}^*_{ea} = 0.9207$.

5.3 Example 2

Let us consider the same statement as in example 1, but now we will presume that if at time t, mobile k uses some power level then at time $t+1$ it can only move to the neighboring states (increasing or decreasing the level by 1) or stay at the same power level. This is compatible with the UMTS standard for the 3rd generation cellular phones in Europe.

It then follows that the Mobile k's state is thus given by a set $\mathbf{I}^k = (\mathbf{M}^k \times \mathbf{Pow}^k)$, where $\mathbf{M}^k$ stands for the channel state and $\mathbf{Pow}^k$ stands for the present transmission power. The action set of mobile k at state $i = (\zeta^k, pow_j)$ is $\mathbf{A}^k(i) = \{-1, 0, 1\}$ for $1 < j < \nu$ where $a = -1$ results in a decrease of the power level to pow_{j-1}, $a = 0$ means remaining at the same power level, and $a = 1$ means increasing the power level to pow_{j+1}. Moreover, for $j = 1$, $\mathbf{A}^k(i) = \{0, 1\}$ and for $j = \nu^k$, $\mathbf{A}^k(i) = \{-1, 0\}$.

In this case the CMDPs are not unichain anymore. Moreover, unlike the previous example, here each mobile indeed controls also the state transitions of his MDP.

Acknowledgements

We would like to thank the reviewer for several very useful suggestions that significantly improved the presentation and the content of the paper.

References

1. E. Altman, *Constrained Markov Decision Processes*, Chapman and Hall/CRC, 1999
2. E. Altman and A. Shwartz, "Markov decision problems and state-action frequencies", *SIAM J. Control and Optimization*, **29**, No. 4, pp. 786-809, 1991.
3. E. Altman and A. Shwartz, "Constrained Markov Games: Nash Equilibria", Annals of the International Society of Dynamic Games, vol. 5, Birkhauser, V. Gaitsgory, J. Filar and K. Mizukami, editors, pp. 303-323, 2000.
4. E. Altman and F. Spieksma, The Linear Program approach in Markov Decision Problems revisited, ZOR - Methods and Models in Operations Research, Vol. 42, Issue 2, pp. 169-188, 1995.
5. J. F. Mertens and A. Neyman, "Stochastic Games", *Int. Journal of Game Theory* Vol. 10, Issue 2, page 53-66, 1981.
6. E. Gómez-Ramírez, K. Najim and A.S. Poznyak, "Saddle-point calculation for constrained finite Markov chains". Journal of Economic Dynamics and Control, **27**, pp. 1833-1853, 2003.
7. A. Hordijk and L. C. M. Kallenberg, "Constrained undiscounted stochastic dynamic programming", *Mathematics of Operations Research,* **9**, No. 2, May 1984.

8. A. Hordijk and L. C. M. Kallenberg, “Linear programming and Markov games I”, in *Game Theory and Mathematical Economics*, O. Moeschlin and D. Pallschke (eds.), North Holland, pp. 291–305, 1981.
9. A. Hordijk and L. C. M. Kallenberg, “Linear programming and Markov games II”, in *Game Theory and Mathematical Economics*, O. Moeschlin and D. Pallschke (eds.), North Holland, pp. 307–320, 1981.
10. L. C. M. Kallenberg (1994), “Survey of linear programming for standard and nonstandard Markovian control problems, Part I: Theory”, *ZOR – Methods and Models in Operations Research*, **40**, pp. 1-42.
11. N. Shimkin, “Stochastic games with average cost constraints”, *Annals of the International Society of Dynamic Games, Vol. 1: Advances in Dynamic Games and Applications*, Eds. T. Basar and A. Haurie, Birkhauser, 1994.
12. O. J. Vrieze, “Linear programming and undiscounted stochastic games in which one player controls transitions”, OR Spektrum 3, pp. 29–35, 1981.