Constrained Markov games with transition probabilities controlled by a single player

Eitan Altman
INRIA BP93
2004 route des Lucioles
06902 Sophia Antipolis
FRANCE
altman@sophia.inria.fr

Saswati Sarkar
Dept. of Electrical and Systems Eng. University of Pennsylvania Philadelphia
19104, USA
swati@seas.upenn.edu

Eilon Solan
School of Mathematical Sciences
Tel Aviv University
ISRAEL
eilons@post.tau.ac.il

ABSTRACT

We consider a two-players zero-sum Markov game with side constraints where only one player controls the transition probabilities. We reduce the problem to that of solving an equivalent linear program. Our approach is different than the one previously used to derive such linear programs [4, 5, 9]. We introduce a new type of constraints: the "subsciption constraints" along with standard constraints which we call "realization constraints". We extend the results obtained in [4, 5, 9] to the case where both players have constraints.

1. INTRODUCTION

In Markov games (also known as stochastic games), there are finitely many matrix games, each identified with a "state". At each time one of these games is played; there are several players who choose simultaneously their actions (e.g. the row and the column of a matrix) and these actions determine a payoff to each player. The state evolution is given by a controlled transition probability matrix. It has the following Markovian property: the next state (i.e. the identity of the next game to be played) conditioned on the present state and on the present actions of the players does not depend on the past states and actions. In that sense, Markov games extend Markov chains.

This does not mean that the state process is Markovian, since some dependence of future states on all the past history given the present state can exist through the way actions are selected. A dependence can be brought in if the decision rule for choosing an action at a given time has a dependence on the whole history. One of the central research issues in Markov games has been to identify special structures of games that guarantee the existence of Markovian or of stationary equilibrium policies (a Markovian policy for a player chooses an action a according to a probability low that is a function of only the current state and time. The choice of a is independent on any previous action or state. A stationary policy for a player is a Markovian policy in which choices of actions do not depend on time). Under any Markov policy, the state process of the Markov game is a Markov chain. Under a stationary policy, this Markov chain is, moreover, time-homogeneous.

We study in this paper a zero-sum Markov game, in which two players have opposite objectives. We restrict to the case in which only one player controls the transition probabilities. We further introduce side constraints on both players. We propose a linear program approach for deriving optimal policies and for computing the value.

SMCtools '07, October 26, 2007, Nantes, France

2. THE MODEL

2.1 Probabilistic structure

We consider a Stochastic game characterized by the following objects:

- **State space.** $\mathbf{X}$ is a finite state space,
- **Initial distribution.** β is a probability distribution over $\mathbf{X}$, according to which the initial state is chosen.
- **Action spaces.** $\mathbf{A}^i$ stands for the finite action space of player i. At state $x \in \mathbf{X}$, the set of actions available to player k is $\mathbf{A}^k(x)$. Let $\mathbf{K}^k = (x, a),\ x \in \mathbf{X}, a \in \mathbf{A}^k(x)$.
- **Transition probabilities.** $\mathcal{P} = \{\mathcal{P}_{xay}\}$ stands for the transition probabilities; $\mathcal{P}_{xay}$ is the probability that the state moves from x to y if player 2 chose action a. We assume that player 1 has no influence on the transition probabilities.

A history h_n is a sequence

$$h_n = (x_1, a_1, x_2, a_2, ..., x_{n-1}, a_{n-1}, x_n),$$

where $x_\ell \in \mathbf{X}$, $a_\ell = (a_\ell^1, a_\ell^2)$, where $a_\ell^k \in \mathbf{A}^k(x_\ell)$, $\ell = 0, 1, 2, ...$ A player k (behavioral) strategy u is a sequence $(u_0, u_1, ...)$ where u_ℓ is a probability measure over $\mathbf{A}^k(x_\ell)$ conditioned on h_n.

Algorithms for stochastic games with a single controller were studied in [7].

Some classes of strategies. Denote by U^k the set of all strategies (also called policies) for player k. Let $U^k(S)$ and $U^k(M)$ be the set of stationary and of Markov policies, respectively, of player k. A stationary policy $u \in U^k(S)$ is identified with a set of probability functions denoted as $u(\cdot|x)$, over the actions $\mathbf{A}^k(x)$. For all $x \in \mathbf{X}$ $u(a|x)$ is then

the probability of choosing action a if the state is x. A Markov policy $u \in U^k(M)$ is identified with a set of probability functions denoted as $u(\cdot, n|x)$, over the actions $\mathbf{A}^k(x)$. For every $x \in \mathbf{X}$ and every positive integer n, $u(a, n|x)$ is the probability of choosing action a at time n if the state is x. We finally introduce the set $U^k(D)$ of pure stationary policies for player k.

An initial distribution β, together with a pair $u = (u^1, u^2)$ of policies $u^1 \in U^1, u^2 \in U^{2,}$ defines a unique probability measure P^u_β on the state-action trajectories. Let E^u_β be the corresponding expectation operator. We write by X_n, A^1_n, A^2_n the stochastic state and action processes.

2.2 Costs and Constraints

We consider a cost $C : U^1 \times U^2 \to \mathbb{R}$ which player 1 has to pay to player 2. Player 1 wishes to minimize this cost and player 2 wishes to maximize it.

Further introduce the cost functions $D^k_s : U^1 \times U^2 \to \mathbb{R}$, $k = 1, 2$, $s = 1, ..., m_k$. Player k has a set M_k of m_k *side constraints* of the form

$$D^k_s(u^1, u^2) \leq \xi^k_s, \ s \in M_k \tag{1}$$

where ξ^k_s are some constants.

We shall use two types of cost functions which we define in the next paragraphs. To that end we introduce first immediate costs functions; for each $x \in \mathbf{X}, a^1 \in \mathbf{A}^1(x), a^2 \in \mathbf{A}^2(x)$ we define the costs $c(x, a^1, a^2)$, and $d^k_s(x, a^1, a^2)$, $k = 1, 2$, $s \in M_k$.

1. **The expected discounted cost.** We define the expected discounted costs as

$$C_\alpha(\beta, u) = (1 - \alpha) \sum_{n=1}^{\infty} \alpha^{n-1} E^u_\beta c(X_n, A^1_n, A^2_n), \tag{2}$$

$$D^{k,s}_\alpha(\beta, u) = (1 - \alpha) \sum_{n=1}^{\infty} \alpha^{n-1} E^u_\beta d^k_s(X_n, A^1_n, A^2_n).$$

   where α is the discount factor. The above costs can be viewed as **realization-based costs**: they are the expected discounted sum of immediate costs of the realizations (X_n, A^1_n, A^2_n) over the time slots n.

2. We also define **subscription-type costs** of the form:

$$D^{j,s}_{subs}(u^j) = \sum_{(x,a^j) \in \mathbf{K}^j} d^j_s(x, a^j) u^j(a^j|x)$$

   where $u^j \in U^j(S)$. We call this type of costs "subscription-based" since it is based on the fraction of time during which a given action will be used at a given state, and is not based on how frequently the state will actually be visited (and the action used). As could be the case in subscription fees for services, the payment for planned use of a service can be done in advance and charging can be simplified by avoiding measuring the actual use of the resources.

For the case of a single player, and when the discounted cost is used, it is sufficient to consider stationary policies without loss of optimality. In the stochastic game framework this means that the best response against a stationary policy is a stationary policy. We will actually obtain stationary saddle-point policies for both players. This will allow us to restrict the stochastic game to stationary policies without loss of optimality.

3. CONSTRAINED ZERO-SUM GAMES: DEFINITIONS AND CLASSIFICATION

The definitions of constrained games that we present in this section are given with respect to abstract cost functions and they will later be used for either the discounted realization based costs or for the subscription type cost.

3.1 Orthogonal constraints

In this framework, one restricts the constraints of player k to depend only on the strategies of that player; that is $D^1_s(u^1, u^2)$ does not depend on u^2, and $D^2_s(u^1, u^2)$ does not depend on u^1. Let U^k_c be the set of strategies of player k that satisfy (1). Let $U_c := (U^1_c, U^2_c)$. We shall assume throughout

$$U^k_c \text{ is non empty, } k = 1, 2. \tag{3}$$

Set $U^k_c(S)$ and $U^k_c(M)$ to be the subset of $U^k(S)$ and $U^k(M)$, respectively, that satisfy (1).

Upper and lower values. The problem faced by player 1 is to find u^1 that achieves the upper value $\overline{V}$ defined as:

$$\overline{V} := \inf_{u^1 \in U^1_c} \sup_{u^2 \in U^2_c} C(u_1, u_2). \tag{4}$$

The problem faced by player 2 is to find u^2 that achieves the lower value $\underline{V}$ defined as:

$$\underline{V} := \sup_{u^2 \in U^2_c} \inf_{u^1 \in U^1_c} C(u_1, u_2). \tag{5}$$

Saddle point. Introduce the cost $C(u^1, u^2)$ 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

$$V := \inf_{u \in U^1_c} C(u, v^*) = C(u^*, v^*) = \sup_{v \in U^2_c} C(u^*, v) \tag{6}$$

(the policies u^* and v^* are also called optimal). If the saddle-point exists then we call V the value of the game which we denote by $V = \mathbf{val}_{u,v} C(u, v)$ where $u \in U^1_c$ and $v \in U^2_c$.

3.2 Non-orthogonal constraints

When the constraints of player k depend also on policies of player $j \neq k$, we define $U^k_c(u^j)$ to be the set of strategies of player k that satisfy (1) when player $j \neq k$ uses strategy u^j. Define U_c to be the set of pairs (u^1_c, u^2_c) such that $u^k \in U^k_c(u^j)$ for $k = 1, 2$ and $j \neq k$. We consider some special cases.

Valuation of the constraints of the adversary. Defining the lower value in the case that only player 2 has non-orthogonal constraints (or the upper valuer in the case it is player 1, or both the lower and upper value if both players have non-orthogonal constrained) requires additional specifications on the goals of the player. Indeed, in those cases, the game is not well defined unless one specifies further how a player values the constraints of the other player. In particular, we may define as in [2] the following frameworks related to this valuation.

- **The aggressive case.** Player k's primary objective is to prevent the other player $j \neq k$ from meeting her constraints. For example, if only player 1 has constraints then in the calculation of the lower value, player 2 can use only strategies that ensure that the constraints will be met, whatever player 1 plays.

- **The indifference case.** Player k does not care whether the constraints of player $j \neq k$ are met. If, e.g., $k = 1$, then in the calculation of the lower value player 2 may use strategies u^2 that do not guarantee that the constraints will be met, provided at strategy u^1 that, together with u^2, violates the constraints, yields low payoff to player 1.
- **The joint case [8].** Each constraint concerns both players, i.e. D_s^k and ξ^k do not depend on k. Then we have $u^1 \in U_c^1(u^2)$ if and only if $u^2 \in U_c^2(u^1)$. If, e.g., $k = 1$, then in the calculation of the lower value player 2 may use any strategy u^2 for which there exists a strategy u^1 such that $(u^1, u^2) \in U_c$, since player 1 will use only strategies u^1 that, together with u^2, meet player 1's constraints, which are identical to player 2's constraints.

We shall describe some special cases in more details.

3.3 The aggressive case

We need some definitions. Define U_g^k to be the set of policies for player k for which, no matter what strategy the other player j chooses among its feasible strategies, player 2 meets her constraints. In other words, $u^k \in U_g^k$ if $u^k \in U_c^k(u^j)$ for every $u^j \in U_c^j$.

If only player 2 has non-orthogonal constraints, then in the aggressive framework the lower value is

$$\underline{V} := \sup_{u^2 \in U_g^2} \inf_{u^1 \in U_c^1} C(u_1, u_2). \tag{7}$$

If only player 1 has non-orthogonal constraints, then in the aggressive framework the upper value is

$$\overline{V} := \inf_{u^1 \in U_g^1} \sup_{u^2 \in U_c^2} C(u_1, u_2). \tag{8}$$

3.4 The indifference case

Only player 2 has non-orthogonal constraints. In the case that only player 2 has non-orthogonal constraints, the upper value $\overline{V}$ is defined as:

$$\overline{V} := \inf_{u^1 \in U_c^1} \sup_{u^2 \in U_c^2(u^1)} C(u_1, u_2). \tag{9}$$

Only Player 1 has non-orthogonal constraints. In the case that only player 1 has non-orthogonal constraints, the lower value $\underline{V}$ defined as:

$$\underline{V} := \sup_{u^2 \in U_c^2} \inf_{u^1 \in U_c^1(u^2)} C(u_1, u_2). \tag{10}$$

Saddle point.

In the case of non-orthogonal constraints, we define a saddle-point couple $(u^*, v^*) \in U_c$ to be a policy for each player such that

$$\begin{aligned} V &:= \inf_{u \in U_c^1(v^*)} C(u, v^*) = C(u^*, v^*) \\ &= \sup_{v \in U_c^2(u^*)} C(u^*, v) \end{aligned} \tag{11}$$

If the saddle-point exists, then we call V the value of the game which we denote by $V = \mathbf{val}_{u,v} C(u, v)$ where $(u, v) \in U_c$.

Unless otherwise stated, we shall restrict to the indifference framework. We have shown in [2] that a saddle-point does not exist in general for the other frameworks.

4. MATHEMATICAL PROGRAMMING APPROACH

We derive below the upper value of the game (as defined in (7)) and an optimal stationary policy for player 1. We assume that only player 2 has constraints, which are realization-based. We reduce the problem to that of a mathematical programming. We then extend the result to the case where player 1 has subscription-type constraints.

We proceed as follows. We first fix a stationary policy u^1 for player 1, and formulate an LP to solve the problem of *best response for player 2* among stationary policies. We then derive an LP to solve the optimization problem (4) faced by player 1.

4.1 Best response for player 2

Fix a stationary policy u^1 for player 1. Then player 2 is faced with a Constrained Markov Decision Process (CMDP) whose sets of immediate costs, $c(u^1)$ and $d_s^2(u^1)$, are given by

$$c(u^1; x, b) := \sum_{a \in \mathbf{A}^1(x)} u^1(a|x) c(x, a, b), \tag{12}$$

$$d_s^2(u^1; x, b) := \sum_{a \in \mathbf{A}^1(x)} u^1(a|x) d_s^2(x, a, b), \tag{13}$$

We show how to compute its value

$$\sup_{u^2 \in U_c^2} C_\alpha(\beta, u^1, u^2). \tag{14}$$

Steps:

1. **No constraints: dynamic programming.** Assume first that there were no constraints on player 2. Then it is well known that the value of the MDP is the unique solution of the dynamic programming (DP):

   $$v_\alpha(x) = \max_{b \in \mathbf{B}(x)} \left((1-\alpha) c(u^1; x, b) + \alpha \sum_{y \in \mathbf{X}} \mathcal{P}_{xby} v_\alpha(y) \right)$$

   for all states x.

2. **No constraints: linear programming.** The above dynamic programming implies that the value of the MDP satisfies

   $$v_\alpha(x) \geq (1-\alpha) c(u^1; x, b) + \alpha \sum_{y \in \mathbf{X}} \mathcal{P}_{xby} v_\alpha(y)$$

   for all states x and actions $b \in \mathbf{A}^2(x)$. Functions with this properties are known as sub-harmonic functions for the discounted case, and the value of the MDP is known to be the **smallest sub-harmonic function** [1, 6]. This statement can be reformulated as an LP:

   $$\text{minimize} \sum_{x \in \mathbf{X}} \beta(x) \phi^2(x)$$

   over the real decision variables $\phi(y)$, $y \in \mathbf{X}$ s.t.

   $$\phi^2(x) \quad \geq \quad (1-\alpha) c(u^1; x, b) + \alpha \sum_{\ell \in \mathbf{X}^2} \mathcal{P}^2_{xb\ell} \phi^2(\ell),$$

   $\forall (x, b) \in \mathbf{K}^2$.

3. **Accounting for the constraints:** We now relax the constraints (1). Define non-positive Lagrange multipliers λ_s (one for each $s \in M_2$). Define the Lagrangian

$$L_\alpha(u^1;\beta,u,\lambda^2) =$$

$$C_\alpha(u^1;\beta,u)+ < \lambda^2, D^2_\alpha(u^1;\beta,u) - \xi^2 > .$$

(where the notation $< \cdot,\cdot >$ stands for the scalar product). u^* for player 2 maximizes $C_\alpha(u^1;\beta,u^2)$ over $u^2 \in U^2_c(S)$ if and only if it maximizes

$$\inf_{\lambda^2 \in R^{m_2}_-} L_\alpha(u^1;\beta,u^2,\lambda^2)$$

over $u^2 \in U^2(S)$. Throughout, R_- is the set of non-positive numbers.

4. **Change the order between max and inf.** The value remains the same under this change [1]. Hence the value of the constrained problem (14) equals that of

$$\inf_{\lambda^2 \in R^{m_2}_-} \max_{u \in U^2(S)} L_\alpha(u^1;\beta,u,\lambda^2)$$

5. Observe that $L_\alpha(u^1;\beta,u^2,\lambda^2)$ is the difference between (i) the expected discounted cost that corresponds to the immediate cost

$$j^\lambda(u^1;x,b) := c(u^1;x,b)+ < \lambda^2, d^2(u^1;x,b) >,$$

and (ii) $< \lambda^2, \xi^2 >$.

6. Hence, by step 2, its maximum over $U^2(S)$ (which equals that of $C_\alpha(u^1;\beta,u^2)$ over $u^2 \in U^2_c(S)$) is obtained by taking the minimum over $\lambda \in R^{m_2}_-$ of the difference between the value of the following LP and $< \lambda^2, \xi^2 >$:

$$\min_{\phi^2} \sum_{x \in \mathbf{X}} \beta(x)\phi^2(x) \text{ s.t.} \qquad (15)$$

$$\phi^2(x) \geq j^\lambda(u^1;x,b) + \sum_{y \in \mathbf{X}^2} \mathcal{P}^2_{xby}\phi^2(y),\ \forall x,b \qquad (16)$$

7. Hence the value of (9) when we hold u^1 fix equals that of the LP:

$$\min_{\phi^2,\lambda} \left(\sum_{x \in \mathbf{X}} \beta(x)\phi^2(x) - < \lambda^2, \xi^2 > \right)$$

s.t. (16) and $\lambda \leq 0$.

4.2 The value, and Player 1's optimal policy

We assume in this subsection that player 1 has no constraints. This is the situation studied in [4, 5] except that in our case the immediate costs involved in the constraints on player 2 are allowed to depend on the actions of both players, where as in these references they depend only on the actions of the player who has the constraints (player 2 in our case).

Player 1 has to choose a policy $u^1 \in U^1_c(S)$ that minimizes the value of LP (15). We note however that u^1 has an impact on the LP only through

$$j^\lambda(u^1;x,b) = \sum_{a \in \mathbf{A}^1(x)} u^1(a|x) j^\lambda(x,a,b),$$

which are linear in u^1. Thus the upper value (9) is obtained by solving the mathematical programming:

MP1 : $\min_{\phi^2,\lambda,u^1} \left(\sum_{x \in \mathbf{X}} \beta(x)\phi^2(x) - < \lambda^2, \xi^2 > \right)$

s.t. (16), $\lambda \leq 0$,

$u^1(a|x) \geq 0\ \forall x,a, \qquad \sum_{a \in \mathbf{A}^1} u^1(a|x) = 1,\ \forall x.$

The only non-linear term in the above MP is j^λ in (16), which is bi-linear in the terms u^1 and λ^2. In the special case that d^2 does not depend on a^1, however, the MP becomes a linear program, derived already in [4, 5] using a different approach. We note that the above MP provides not only the value of (4) but also the optimal u^1.

We now consider additional realization-based constraints on player 1. In other words, we aim at solving the following problem:

P2 : Obtain a saddle-point (u^*, v^*) for the game (6) where $C_\alpha(\beta,u,v)$ is the objective function that we minimize over the strategies U^2_c of player 2, and that we maximize over the strategies U^1_c of player 1, where
$U^1_c = \left\{ u^1 \in U^1(S) : D^{1,s}_{subs}(\beta,u^1) \leq \xi^1_s \right\}$

$U^2_c = \left\{ u^2 \in U^2(S) : D^{2,s}_\alpha(\beta,u^2) \leq \xi^2_s \right\}$

This problem corresponds to a situation in which player 1 is a customer, who is constrained by subscription-type costs, and player 2 is the service provider, who is constrained by realization-based costs. The game is equivalent to the following linear programming:

LP2 : $\min_{\phi^2,\lambda,u^1} \left(\sum_{x \in \mathbf{X}} \beta(x)\phi^2(x) - < \lambda^2, \xi^2 > \right)$

s.t. (16), $\lambda \leq 0$,

$u^1(a|x) \geq 0\ \forall x,a, \qquad \sum_{a \in \mathbf{A}^1} u^1(a|x) = 1,\ \forall x.$

$\sum_{(x,a) \in \mathbf{K}^1} u^1(a|x) d^{1,s}_{subs}(x,a) \leq \xi^1_s.$

This problem is more general than the one in [4, 5] as it has constraints on both players. LP2 directly provides an optimal stationary policy for player 1. As in [4, 5], the dual will provide an optimal stationary policy for player 2.

5. OCCUPATION MEASURE APPROACH

Preliminaries. Define for every $x \in \mathbf{X}$ and $u^2 \in U^2(S)$ the following discounted state occupation measure:

$$\pi_\alpha(\beta,u^2;y) = (1-\alpha)\sum_{t=0}^{\infty} \alpha^t \sum_{x \in \mathbf{X}} \beta(x) \Big([P(u^2)]^t \Big)_{xy}.$$

here $[P(u^2)]^0$ is the identity matrix. We further define the discounted state-action occupation measure to be

$$\pi_\alpha(\beta,u^2;y,b) = \pi_\alpha(\beta,u^2;y)u^2(b|y).$$

Let $u = (u^1, u^2)$, where $u^i \in U^i(S)$, and fix a distribution β over the initial state. The costs are related to the occupation measure through

$$C_\alpha(\beta, u) = \sum_{(x,b)\in\mathbf{K}^2} \pi_\alpha(\beta, u^2; x, b) \sum_{a\in\mathbf{A}^1} u^1(a|x)c(x, a, b)$$

$$D^k_{s,\alpha}(\beta, u) = \sum_{(x,b)\in\mathbf{K}^2} \pi_\alpha(\beta, u^2; x, b) \sum_{a\in\mathbf{A}^1} u^1(a|x)d^k_s(x, a, b)$$

Achievable occupation measures Let $\mathbf{Q}_\alpha$ be the set of vectors $\rho \in \mathbf{R}^{|\mathbf{K}^2|}$ satisfying

$$\begin{cases} \sum_{(y,b)\in\mathbf{K}^2} \rho(y, b)(\delta_x(y) - \alpha\mathcal{P}_{ybx}) = (1-\alpha)\beta(x), \quad \forall x, \\ \sum_{(y,b)\in\mathbf{K}^2} \rho(y, b) = 1, \\ \rho(y, b) \geq 0, \quad \forall (y, b) \in \mathbf{K}^2, \end{cases} \tag{17}$$

where $\delta_x(y)$ is the indicator which is equal to one if $x = y$ and is zero otherwise. The first equality is equivalent to

$$\sum_{b\in\mathbf{A}^2(x)} \rho(x, b) = (1-\alpha)\beta(x) + \alpha \sum_{(y,b)\in\mathbf{K}^2} \rho(y, b)\mathcal{P}_{ybx},$$

which means that the discounted frequency to visit the state x is the sum of the probability that x is the initial state (times $1-\alpha$) and the probability to move to x (times α).

Note that any ρ satisfying the above constraints is a probability measure.

Define

- $\prod^k_{pure}$ the finite set of occupation measures corresponding to all pure stationary policies of player 2.
- $\prod^2_S$ the set of occupation measures corresponding to all stationary policies of player 2.
- $M(\prod)$ the convex hull of a set $\prod$.

Define

$$\mathcal{C}(u^1, \rho) := \sum_{(x,b)\in\mathbf{K}^2} \rho(x, b) \sum_{a\in\mathbf{A}^1} u^1(a|x)c(x, a, b)$$

$$\mathcal{D}^k_{s,\alpha}(u^1, \rho) := \sum_{(x,b)\in\mathbf{K}^2} \rho(x, b) \sum_{a\in\mathbf{A}^1} u^1(a|x)d^k_s(x, a, b)$$

For a given probability measure ρ over $\mathbf{K}^2$, we define the stationary policy $w(\rho)$ for player 2 as

$$w_x(b, \rho) = \frac{\rho(x, b)}{\sum_{b'\in\mathbf{A}^2(x)} \rho(x, b')}, \qquad b \in \mathbf{A}^2(x), \tag{18}$$

whenever the denominator is non-zero (when it is zero, $w^2(\rho)$ is chosen arbitrarily). Here $w^2_i(a, \rho)$ is the probability that player 2 will choose action a at state i according to this stationary policy.

The next relations follow from [1].

LEMMA 1. *(i) $M(\prod^2_{pure}) = \prod^2_S = \mathbf{Q}_\alpha$.*
(ii) Then for any β, $u^1 \in U^1(S)$ and $u^2 \in U^2(S)$,

$$C_\alpha(\beta, (u^1, u^2)) = \mathcal{C}(u^1, \rho), \tag{19}$$

$$D^k_{s,\alpha}(\beta, (u^1, u^2)) = \mathcal{D}^s_k(u^1, \rho), \tag{20}$$

where $\rho(x, b) = \pi_\alpha(\beta, u^2; x, b)$.
(iii) Conversely, for every u^1 and every $\rho \in Q_\alpha$ Eqs. (19)-(20) hold, where $u_2(b|x) := w_x(b, \rho)$ is given in (18).

REMARK 1. *The set of occupation measures corresponding to the discounted cost (and hence also the set of achievable discounted costs) is thus a polytope. This is trivially true also for the subscription-type costs, where we can view directly $\{u^j(a|x)\}, (x, a) \in \mathbf{K}^j$ as the occupation measure for player j that corresponds to a policy u^j.*

6. A MATRIX GAME REPRESENTATION

We obtain from Lemma 1 (i) the following.

COROLLARY 1. *(i) for every $u \in U^2(S)$ there exists a probability measure γ^2 over the pure stationary policies $U(D)^2$ such that*

$$\pi_\alpha(\beta, u^2) = \sum_{u\in U(D)^2} \gamma^2(u)\pi_\alpha(\beta, u)$$

(ii) Conversely, for every probability measure γ^2 over the pure stationary policies $U(D)^2$, there is a stationary policy $\widehat{\gamma}^2 \in U^2(S)$ such that

$$\pi_\alpha(\beta, \widehat{\gamma}^2) = \sum_{u\in U(D)^2} \gamma^2(u)\pi_\alpha(\beta, u)$$

We have a similar representation for player 1, which concerns this time directly the policies. More precisely, we note that the set $U^1(S)$ is convex (an element of $U^1(S)$ is identified by a vector of $|\mathbf{X}|$ probabilities; the xth element is identified with a probability over $\mathbf{A}^1(x)$). Applying Krein-Milman's theorem, we obtain:

LEMMA 2. *For any stationary policy $u^1 \in U^1(S)$ there exists a probability measure γ^1 over the pure stationary policies $U(D)^1$ such that*

$$u^1 = \sum_{u\in U(D)^1} u\gamma^1(u)$$

(ii) Conversely, for every probability measure γ^1 over the pure stationary policies $U(D)^1$, there is a stationary policy $\widehat{\gamma}^1 \in U^1(S)$ such that

$$\widehat{\gamma}^1 = \sum_{u\in U(D)^1} u\gamma^1(u)$$

We shall call $\widehat{\gamma}^k$ the mixed representation of the stationary policy u^k. We obtain the following.

THEOREM 1. *Given any pair of stationary policies (u^1, u^2) with mixed representations $\widehat{\gamma}^1$ and $\widehat{\gamma}^2$, respectively, we have*

$$C_\beta(\alpha, \widehat{\gamma}^1, \widehat{\gamma}^2) = \sum_{u^1\in U^1(D)} \sum_{u^2\in U^2(D)} \gamma^1(u^1)\gamma^2(u^2)C_\beta(\alpha, u^1, u^2);$$

for all $s \in M_1$,

$$D^{1,s}_{subs}(\beta, \widehat{\gamma}^2) = \sum_{u^2\in U(D)^2} \gamma^2(u^2)D^{1,s}_{subs}(\beta, u^1)$$

and for all $s \in M_2$,

$$D^{2,s}_\alpha(\beta, \widehat{\gamma}^2) = \sum_{u^2\in U(D)^2} \gamma^2(u^2)D^{2,s}_\alpha(\beta, u^2)$$

In view of the Theorem, problem P2 can be viewed as a constrained matrix game, whose rows and columns correspond to the pure stationary policies of player 1 and 2, respectively. We can apply therefore directly the theory of constrained zero-sum matrix games from [3] to obtain a linear program and its dual that provide the saddle-point value and optimal policies.

Acknowledgement

The work of the first author was supported by the Bionets European project.

7. REFERENCES

[1] E. Altman, *Constrained Markov Decision Processes*, Chapman and Hall/CRC, 1999

[2] E. Altman and E. Solan, "Games with constraints with networking applications", submitted.

[3] A. Charnes, "Constrained games and linear programming", *Proceedings of the National Academy of Sciences of the USA*, Vol 39, 639–641, 1953.

[4] 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.

[5] 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.

[6] 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.

[7] T. E. S. Raghavan, "Finite-step algorithms for single-controller and perfect information stochastic games", In *Stochastic games and applications* (Stony Brook, NY, 1999), 227–251, NATO Sci. Ser. C Math. Phys. Sci., 570, Kluwer Acad. Publ., Dordrecht, 2003.

[8] J. B. Rosen. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica*, 33:153–163, 1965.

[9] O. J. Vrieze, "Linear programming and undiscounted stochastic games in which one player controls transitions", OR Spektrum 3, pp. 29–35, 1981.