Proceedings of the
46th IEEE Conference on Decision and Control
New Orleans, LA, USA, Dec. 12-14, 2007

Polynomial stochastic games via sum of squares optimization

Parikshit Shah and Pablo A. Parrilo
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract— **Stochastic games are an important class of games that generalize Markov decision processes to game theoretic scenarios. We consider finite state two-player zero-sum stochastic games over an infinite time horizon with discounted rewards. The players are assumed to have infinite strategy spaces and the payoffs are assumed to be polynomials. In this paper we restrict our attention to a very special class of games for which the *single-controller assumption* holds. It is shown that minimax equilibria and optimal strategies for such games may be obtained via semidefinite programming.**

I. INTRODUCTION

Markov decision processes (MDPs) are very widely used system modeling tools where a single agent attempts to make optimal decisions at each stage of a multi-stage process so as to optimize some reward or payoff [1]. Game theory is a system modeling paradigm that allows one to model a problem where several (possibly adversarial) decision makers make individual decisions to optimize their own payoff [2]. In this paper we study *stochastic games* [3, 4] that allow one to combine the modeling power of MDPs and games. Stochastic games may be viewed as *competitive MDPs* where several decision makers make decisions at each stage to maximize their own reward. Each state of a stochastic game is a simple game, but the decisions made by the players affect not only their payoff, but also the transition to the next state.

Notions of optimality in games have been extensively studied, and are very well understood. The most popular notion of optimality is the notion of a *Nash equilibrium*. While these equilibria are hard to compute in general, in certain cases they may be computed efficiently. Games involving two players and finite action spaces are known to have mixed strategy Nash equilibria. Moreover, Nash equilibria for such games may be computed efficiently via linear programming. Stochastic games were introduced by Shapley [4] in 1953. In his paper, he showed that the notion of a Nash equilibrium may be extended to stochastic games with finite state spaces and strategy sets. He also proposed a value iteration-like algorithm to compute the equilibria. Shapley showed that associated to this notion of equilibrium, there is a notion of value associated to the stochastic game. This value is a vector indexed by the state, and corresponds to the optimal payoff for the players given the initial state of the game. It was shown in 1981 by Parthasarathy and Raghavan [3, 5] that the value and optimal strategies for stochastic games satisfying the *single controller* assumption could be computed efficiently via linear programming (thus proving that such problems with rational data could be computed in a finite number of steps).

This research was funded in part by AFOSR MURI subawards 2003-07688-1 and 102-1080673.

While computational techniques for finite games are reasonably well understood, there has been some recent interest in the class of *infinite games* [6, 7]. In this important class, players have access to an infinite number of pure strategies, and the players are allowed to randomize over these choices. In a recent paper [6], Parrilo describes a technique to solve two player, zero-sum infinite games with polynomial payoffs via semidefinite programming. It is natural to wonder whether the techniques from finite stochastic games can be extended to infinite stochastic games (i.e. finite state stochastic games where players have access to infinitely many pure strategies). In particular, since finite, single-controller, zero-sum games can be solved via linear programming, can similar infinite stochastic games be solved via semidefinite programming? The answer is affirmative, and this paper focuses on establishing this result. The linear program that solves the finite action stochastic game satisfying (SC) (a condition defined below) can be extended to an infinite dimensional optimization problem when the actions are uncountably infinite. The main contribution of this paper is the establishment of the following properties of the infinite dimensional optimization problem:

1) Its optimal solutions correspond to minimax equilibria.
2) The problem can be solved efficiently by semidefinite programming.

Section II of this paper provides a formal description of the problem and introduces the basic notation used in the paper. It also briefly describes some elegant results about polynomial nonnegativity, moment sequences of nonnegative measures, and their connection to semidefinite programming. Section III states and proves the main result of this paper. Finally, we state some natural extensions of this problem, conclusions, and directions of future research.

II. PRELIMINARIES

A. Problem formulation

We consider the problem of solving two-player zero-sum stochastic games via mathematical programming. The game consists of finitely many states with two adversarial players making simultaneous decisions. Each player receives a payoff that depends on the actions of both players and the state (i.e. each state can be thought of as a particular zero-sum game). The transitions between the states are random (as

1-4244-1498-9/07/$25.00 ©2007 IEEE.

in a finite state Markov decision process), and the transition probabilities in general depend on the actions of the players and the current state. The process runs over an infinite horizon. Player 1 attempts to maximize his reward over the horizon (via a discounted accumulation of the rewards at each stage) while player 2 tries to minimize his payoff to player 1. If $(a_1^1, a_1^2, \ldots)$ and $(a_2^1, a_2^2, \ldots)$ are sequences of actions chosen by players 1 and 2 resulting in a sequence of states $(s_1, s_2, \ldots)$ respectively, then the reward of player 1 is given by:

$$\sum_{k=1}^{\infty} \beta^k r(s_k, a_1^k, a_2^k).$$

The game is completely defined via the specification of the following data:

1) The (finite) state space $\mathcal{S} = \{1, \ldots, S\}$.
2) The sets of actions for players 1 and 2 given by A_1 and A_2.
3) The payoff function, denoted by $r(s, a_1, a_2)$, for a given set of state s and actions a_1 and a_2 (of players 1 and 2).
4) The probability transition matrix $p(s'; s, a_1, a_2)$ which provides the conditional probability of transition from state s to s' given players' actions.
5) The discount factor β ($\beta < 1$).

Throughout this paper we make the following important assumption about the probability transition matrix:

Assumption SC
The probability transition to state s' conditioned upon the current state being s depends only on s, s', and the action a_1 of player 1 for every s and s'. This probability is *independent of the action of player* 2. Thus, $p(s'; s, a_1, a_2) = p(s'; s, a_1)$. This is known as the *single-controller assumption*.

In this paper we will be concerned with the case where the action spaces of the two players A_1 and A_2 are uncountably infinite sets. For the sake of simplicity we will often consider the case where $A_1 = A_2 = [0, 1] \subset \mathbb{R}$. The results easily generalize to the case where the strategy sets are finite unions of arbitrary intervals of the real line. For the sake of simplicity, we also assume that the action sets are the same for each state, though this assumption may also be relaxed. We will denote by a_1 and a_2, the actual actions chosen by players 1 and 2 from their respective action spaces. The payoff function is assumed to be a polynomial in the variables a_1 and a_2 with real coefficients:

$$r(s, a_1, a_2) = \sum_{i=0}^{n_s} \sum_{j=0}^{m_s} r_{ij}(s) a_1^i a_2^j.$$

Finally, we assume that the transition probability $p(s'; s, a_1)$ is a polynomial in the action a_1.

The decision process runs over an infinite horizon, thus it is natural to restrict one's attention to stationary strategies for each player, i.e. strategies that depend only on the state of the process and not on time. Moreover, since the process involves two adversarial decision makers, it is also natural to look for randomized strategies (or mixed strategies) rather than pure strategies so as to recover the notion of a minimax equilibrium. A *mixed* strategy for player 1 is a finite set of probability measures $\mu = [\mu(1), \ldots, \mu(S)]$ supported on the action set A_1. Each probability measure corresponds to a randomized strategy for player 1 in some particular state, for example $\mu(k)$ corresponds to the randomized strategy that player 1 would use when in state k. Similarly, player 2's strategy will be represented by $\nu = [\nu(1), \ldots, \nu(N)]$. (A word on notation: Throughout the paper, indices in parentheses will be used to denote the state. Bold letters will be used indicate vectorization with respect to the state, i.e. collection of objects corresponding to different states into a vector with the i^{th} entry corresponding to state i. The Greek letters ξ, μ, ν will be used to denote measures. Subscripts on these Greek letters will be used to denote moments of the measures. For example $\xi_j(i)$ denotes the j^{th} moment of the measure ξ corresponding to state i.)

A strategy μ leads to a probability matrix $P(\mu)$ such that $P_{ij}(\mu) = \int_{A_1} p(j; i, a_1) d\mu(i)$. Thus, once player 1 fixes a strategy μ, the probability transition matrix is fixed, and can be obtained by integrating each entry in the matrix with respect to the measure μ. (Since the entries are polynomials, upon integration, these entries depend affinely on the moments $\mu_k(i)$). Given strategies μ and ν, the reward collected by player 1 in some stage s is given by:

$$r(s, \mu(s), \nu(s)) = \int_{A_1} \int_{A_2} r(s, a_1, a_2) d\mu(s) d\nu(s).$$

The reward collected over the infinite horizon (for fixed strategies $\mu(s)$ and $\nu(s)$) starting at state s, $v_\beta(s, \mu(s), \nu(s))$, is given by the system of equations:

$$v_\beta(s, \mu(s), \nu(s)) = r(s, \mu(s), \nu(s)) + \beta \sum_{s' \in \mathcal{S}} \left(\int_{A_1} p(s'; s, a_1) d\mu(s) \right) v_\beta(s', \mu(s'), \nu(s')) \quad \forall s.$$

Vectorizing $v_\beta(s, \mu(s), \nu(s))$, we obtain

$$\mathbf{v}_\beta(\mu, \nu) = (I - \beta P(\mu))^{-1} \mathbf{r}(\mu, \nu),$$

where $\mathbf{r}(\mu, \nu) = [r(1, \mu(1), \nu(1)), \ldots, r(S, \mu(S), \nu(S))] \in \mathbb{R}^S$.

The problem is to find equilibrium strategies μ^0 and ν^0 which satisfy the saddle point property:

$$\mathbf{v}_\beta(\mu, \nu^0) \leq \mathbf{v}_\beta(\mu^0, \nu^0) \leq \mathbf{v}_\beta(\mu^0, \nu)$$

for all mixed strategies μ, ν. (Mixed strategies that satisfy this saddle point property achieve the Nash equilibrium.) One may note that $\mathbf{v}_\beta(\mu, \nu)$ is a vector in $\mathbb{R}^S$ indexed by the initial state of the Markov process. Hence the above inequality is a vector inequality and is to be interpreted componentwise.

B. SDP Characterization of Nonnegativity and Moments

Let A be some interval on the real line. The set of univariate polynomials that are nonnegative on A has an exact semidefinite description. The set of (finite) vectors

in $\mathbb{R}^n$ which correspond to moment sequences of measures supported on A also has an exact semidefinite description. We briefly review these notions here and introduce some related notation [6].

Let $\mathbb{R}[x]$ denote the set of univariate polynomials with real coefficients. Let $p(x) = \sum_{k=0}^{n} p_k x^k \in \mathbb{R}[x]$. We say that $p(x)$ is nonnegative on A if $p(x) \geq 0$ for every $x \in A$. We denote the set of nonnegative polynomials of degree n which are nonnegative on A by $\mathcal{P}(A)$. (To avoid cumbersome notation, we exclude the degree information in the notation. Moreover the degree will usually be clear from the context.) The polynomial $p(x)$ is said to be a *sum of squares* if there exist polynomials $q_1(x), \ldots, q_k(x)$ such that $p(x) = \sum_{i=1}^{k} q(x)^2$. It is well known that a univariate polynomial is a sum of squares if and only if $p(x) \in \mathcal{P}(\mathbb{R})$.

Let μ denote a measure supported on the set A. The i^{th} moment of the measure μ is denoted by

$$\mu_i = \int_A x^i d\mu.$$

Let $\bar{\mu} = [\mu_0, \ldots, \mu_n]$ be a vector in $\mathbb{R}^{n+1}$. We say that $\bar{\mu}$ is a *moment sequence* of length $n+1$ if it corresponds to the first $n+1$ moments of some nonnegative measure μ supported on the set A. The *moment space*, denoted by $\mathcal{M}(A)$ is the subset of $\mathbb{R}^{n+1}$ which corresponds to moments of nonnegative measures supported on the set A. We say that a nonnegative measure μ is a *probability measure* if its zeroth order moment $\mu_0 = 1$. The set of moment sequences of length $n+1$ corresponding to probability measures is denoted by $\mathcal{M}_P(A)$.

Let $\mathcal{S}^n$ denote the set of $n \times n$ symmetric matrices and define the linear operator $\mathcal{H} : \mathbb{R}^{2n-1} \to \mathcal{S}^n$ as:

$$\mathcal{H} : \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{2n-1} \end{bmatrix} \mapsto \begin{bmatrix} a_1 & a_2 & \ldots & a_n \\ a_2 & a_3 & \ldots & a_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_n & a_{n+1} & \ldots & a_{2n-1} \end{bmatrix}.$$

Thus $\mathcal{H}$ is simply the linear operator that takes a vector and constructs the associated Hankel matrix which is constant along the antidiagonals. We will also frequently use the adjoint of this operator, the linear map $\mathcal{H}^* : \mathcal{S}^n \to \mathbb{R}^{2n-1}$:

$$\mathcal{H}^* : \begin{bmatrix} m_{11} & m_{12} & \ldots & m_{1n} \\ m_{12} & m_{22} & \ldots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{1n} & m_{2n} & \ldots & m_{nn} \end{bmatrix} \mapsto \begin{bmatrix} m_{11} \\ 2m_{12} \\ m_{22} + 2m_{13} \\ \vdots \\ m_{nn} \end{bmatrix}.$$

This map flattens a matrix into a vector by adding all the entries along antidiagonals. One can give a semidefinite characterization of polynomials that are nonnegative on an interval. Since in this paper we are typically considering the interval to be $[0, 1]$ we give an explicit semidefinite characterization of $\mathcal{P}([0, 1])$. We define the following matrices:

$$L_1 = \begin{bmatrix} I_{n\times n} \\ 0_{1\times n} \end{bmatrix}, \quad L_2 = \begin{bmatrix} 0_{1\times n} \\ I_{n\times n} \end{bmatrix},$$

where $I_{n\times n}$ stands for the $n \times n$ identity matrix.

Lemma 1: The polynomial $p(x) = \sum_{k=0}^{2n} p_k x^k$ is nonnegative on [0,1] if and only if there exist matrices $Z \in \mathcal{S}^{n+1}$ and $W \in \mathcal{S}^n$, $Z \succeq 0, W \succeq 0$ such that

$$\begin{bmatrix} p_0 \\ \vdots \\ p_{2n} \end{bmatrix} = \mathcal{H}^*(Z + \frac{1}{2}(L_1 W L_2^T + L_2 W L_1^T) - L_2 W L_2^T).$$

Proof: See [6]. ■

In this paper, we will also be using a very important classical result about the semidefinite representation of moment spaces [10, 11]. We give an explicit characterization of $\mathcal{M}([0, 1])$ and $\mathcal{M}_P([0, 1])$.

Lemma 2: The vector $\bar{\mu} = [\mu_0, \mu_1, \ldots, \mu_{2n}]^T$ is a valid set of moments for a nonnegative measure supported on $[0, 1]$ if and only if

$$\begin{aligned} \mathcal{H}(\bar{\mu}) &\succeq 0 \\ \tfrac{1}{2}(L_1^T \mathcal{H}(\bar{\mu}) L_2 + L_2^T \mathcal{H}(\bar{\mu}) L_1) - L_2^T \mathcal{H}(\bar{\mu}) L_2 &\succeq 0. \end{aligned} \quad (1)$$

Moreover, it is a moment sequence corresponding to a probability measure if and only if in addition to (1) it satisfies $\mu_0 = 1$.

Proof: A proof may be found in [10]. ■

III. Infinite Strategy Games

A. Problem Setup

In this paper we consider stochastic games in which each player can choose from uncountably many different actions. In particular, each player can choose actions from the set $[0, 1]$. The number of states $|\mathcal{S}| = S$ is assumed to be finite. The payoff function $r(s, a_1, a_2)$ is a polynomial in a_1 and a_2 for each $s \in \mathcal{S}$. In finite action stochastic games one tries to determine probability vectors $\mathbf{f}$ and $\mathbf{g}$ (for players 1 and 2 respectively) representing probability distributions (mixed strategies) over the finite sets A_1 and A_2 (see [3]). In this paper these are replaced by probability measures $\mu(s)$ and $\nu(s)$. These measures represent mixed strategies over the uncountable action spaces. (We remind the reader that for each player there are S measures, each measure corresponding to a mixed strategy in a particular state. For example $\mu(s)$ corresponds to the mixed strategy player 1 would adopt when the game is in state s.)

B. Preliminary Results

In this section we establish that the problems (P') and (D') (stated below) are exact duals and that they are efficiently computable via their SDP counterparts (SP) and (SD). In subsection C we prove that the solution to the problem (P') provides the optimal cost to go and an optimal strategy for player 2 and that the optimal solution to (D') provides an optimal strategy for player 1. Consider the

optimization problem:

$$(P) \quad \begin{array}{l}\text{Minimize} \sum_{s\in\mathcal{S}} v(s) \\ \nu(s), v(s)\end{array}$$

(a) $v(s) \geq \int_{a_2\in A_2} r(s,a_1,a_2)d\nu(s) + \beta\sum_{s'\in\mathcal{S}} p(s';s,a_1)v(s')$ for all $s\in\mathcal{S}, a_1\in A_1$

(b) $\nu(s)$ is a measure supported on A_2 for all $s\in\mathcal{S}$

It is infinite dimensional since it involves optimization over the space of probability measures. A feasible set of measures $\nu(s)$ correspond to a security strategy where player 2 would have to pay no more than $v(s)$ to player 1 for a given initial state s. Note that the constraints (a) are a system of polynomial inequalities with coefficients that depend on the measure ν only via finitely many moments. More concretely, let $r(s,a_1,a_2) = \sum_{i,j}^{n_s,m_s} r_{ij}(s)a_1^i a_2^j$ be the payoff polynomial. Then $\int r(s,a_1,a_2)d\nu(s) = \sum_{i,j} r_{ij}(s)a_1^i\nu_j(s)$. Using this observation, this problem may be rewritten as:

$$(P') \quad \begin{array}{l}\text{Minimize} \sum_{s\in\mathcal{S}} v(s) \\ \bar{\nu}(s), v(s)\end{array}$$

$$(c) \quad v(s) - \sum_{i,j} r_{ij}(s)a_1^i\nu_j(s) - \beta\sum_{s'\in\mathcal{S}} p(s';s,a_1)v(s') \in \mathcal{P}(A_1) \text{ for all } s\in\mathcal{S}$$

$$(d) \quad \bar{\nu}(s)\in\mathcal{M}(A_2), \text{ and } \nu_0(s)=1 \text{ for all } s\in\mathcal{S}.$$

Consider also the optimization problem (D') stated below. We will establish that (P') and (D') form a primal-dual pair of polynomial optimization problems.

$$(D') \quad \begin{array}{l}\text{Maximize} \sum_{s\in\mathcal{S}} \alpha(s) \\ \alpha(s), \bar{\xi}(s)\end{array}$$

$$(e) \quad \sum_{i,j} r_{ij}(s)\xi_i(s)a_2^j - \alpha(s) \geq 0 \quad \forall a_2\in A_2, s\in\mathcal{S}$$

$$(f) \quad \bar{\xi}(s)\in\mathcal{M}(A_2) \quad \forall s\in\mathcal{S}$$

$$(g) \quad \sum_s \int_{A_1} (\delta(s,s') - \beta p(s',s,a_1))d\xi(s) = 1 \quad \forall s'\in\mathcal{S}.$$

The constraints (c) give a system of polynomial inequalities in a_1, one inequality per state. Fix some state s. Let the degree of the inequality for that state by d_s. Let $[a_1]_{d_s} = [1, a_1, a_1^2, \ldots a_1^{d_s}]$. The first term in constraint (c) can be rewritten in vector form as:

$$\sum_{i,j} r_{ij}(s)a_1^i\nu_j(s) = \bar{\nu}(s)^T R(s)^T [a_1]_{d_s},$$

where $R(s)$ is a matrix that contains the coefficients of the polynomial $r(s,a_1,a_2)$. We define a value vector for the game by $\mathbf{v}^* = [v^*(1),\ldots,v^*(S)]^T$ which will turn out to be the discounted value of the stochastic game (which is dependent on the initial state). The second term in the constraint (c) which depends on the probability transition $p(s';s,a_1)$ is also a polynomial in a_1 whose coefficients depend on the coefficients of $p(s';s,a_1)$ and $\mathbf{v}$. Specifically

$$\sum_{s'=1}^{S} p(s';s,a_1)v(s') = \mathbf{v}^T Q(s)^T [a_1]_{d_s},$$

for some matrix $Q(s)$ which contains the coefficients of $p(s';s,a_1)$.

Lemma 3: Let $A_1 = A_2 = [0,1]$. Let $E_s \in \mathbb{R}^{d_s\times S}$ be the matrix which has a 1 in the $(1,s)$ position. Then the semidefinite program (SP) given by:

$$(SP) \quad \begin{array}{l}\text{Minimize} \sum_{s\in\mathcal{S}} v(s) \\ \bar{\nu}(s), v(s)\end{array}$$

$$(h) \quad \begin{array}{l}\mathcal{H}^*(Z_s + \frac{1}{2}(L_1 W_s L_2^T + L_2 W_s L_1^T) - L_2 W_s L_2^T) \\ = E_s\mathbf{v} - \beta Q(s)\mathbf{v} - R(s)\bar{\nu}(s) \quad \forall s\in\mathcal{S}\end{array}$$

$$(i) \quad \mathcal{H}(\bar{\nu}(s)) \succeq 0 \quad \forall s\in\mathcal{S}$$

$$(j) \quad \begin{array}{l}\frac{1}{2}\left({L_1}^T\mathcal{H}(\bar{\nu})(s)L_2 + L_2^T\mathcal{H}(\bar{\nu})(s)L_1\right) \\ -{L_2}^T\mathcal{H}(\bar{\nu})(s)L_2 \succeq 0 \quad \forall s\in\mathcal{S}\end{array}$$

$$(k) \quad {e_1}^T\bar{\nu}(s) = 1 \quad \forall s\in\mathcal{S}$$

$$(l) \quad Z_s, W_s \succeq 0 \quad \forall s\in\mathcal{S}$$

exactly solves the polynomial optimization problem (P').

Proof: The polynomial in inequality (c) has the coefficient vector $E_s\mathbf{v} - \beta Q(s)\mathbf{v} - R(s)\bar{\nu}(s)$. The proof follows as a direct consequence of Lemma 1 concerning the semidefinite representation of polynomials nonnegative over $[0,1]$, and Lemma 2 concerning the semidefinite representation of moment sequences of nonnegative measures supported on $[0,1]$. ■

The dual of (SP) is given by the following semidefinite program:

$$(SD) \quad \begin{array}{l}\text{Maximize} \sum_{s\in\mathcal{S}} \alpha(s) \\ \alpha(s), \bar{\xi}(s)\end{array}$$

$$(m) \quad \begin{array}{l}\mathcal{H}^*(A_s + \frac{1}{2}(L_1 B_s L_2^T + L_2 B_s L_1^T) - L_2 B_s L_2^T) = \\ R_s^T\bar{\xi}(s) - \alpha(s)e_1 \quad \forall s\in\mathcal{S}\end{array}$$

$$(n) \quad \mathcal{H}(\bar{\xi}(s)) \succeq 0 \quad \forall s\in\mathcal{S}$$

$$(o) \quad \begin{array}{l}\frac{1}{2}\left({L_1}^T\mathcal{H}(\bar{\xi}(s))L_2 + L_2^T\mathcal{H}(\bar{\xi}(s))L_1\right) - \\ {L_2}^T\mathcal{H}(\bar{\xi}(s))L_2 \succeq 0 \quad \forall s\in\mathcal{S}\end{array}$$

$$(p) \quad \sum_s (E_s - \beta Q(s))^T\bar{\xi}(s) = 1$$

$$(q) \quad A_s, B_s \succeq 0 \quad \forall s\in\mathcal{S}.$$

Lemma 4: The dual SDP (SD) is equivalent to the polynomial optimization problem (D'):

Proof: This again follows as a consequence of lemmas 1 and 2. ■

Remarks 1. Note that in the dual problem, the moment sequences do not necessarily correspond to probability measures. Hence, to convert them to probability measures, one

needs to normalize the measure. Upon normalization, one obtains the optimal strategy for player 1.
2. The solution of the SDPs give moment sequences corresponding to the optimal measures. The optimal measures themselves can be chosen to be atomic and may be recovered by standard techniques that rely only on linear algebra (see [10], [6]).

Lemma 5: The polynomial optimization problems (P') and (D') are strong duals of each other.

Proof: We prove this by showing that the semidefinite program (SP) satisfies Slater's constraint qualification and that it is bounded from below. The result then follows from the strong duality of the equivalent semidefinite programs (SP) and (SD).
First pick $\mu(s)$ and $\nu(s)$ to be the uniform distribution on $[0,1]$ for each state $s \in \mathcal{S}$. One can show [10] that the moment sequence of μ is in the interior of the moment space of $[0,1]$. As a consequence, constraints (i) and (j) are strictly positive definite. Using the strategies μ and ν, evaluate the discounted value of this pair of strategies as:

$$\mathbf{v}_\beta(\mu,\nu) = [I - \beta P(\mu)]^{-1}\mathbf{r}(\mu,\nu).$$

Choose $\mathbf{v} > \mathbf{v}_\beta$. The polynomial inequalities given by (c) are all strictly positive and thus constraints (l) are strictly positive definite. The equality constraints are trivially satisfied.
To prove that the problem is bounded below, we note that $r(s,a_1,a_2)$ is a polynomial and that the strategy spaces for both players are bounded. Hence,

$$\inf_{a_1\in A_1, a_2 \in A_2} r(s,a_1,a_2)$$

is finite and provides a trivial lower bound for $v(s)$. ■

Lemma 6: Let $\bar{\nu}^*(s)$ and $\bar{\xi}^*(s)$ be optimal moment sequences for (P') and (D') respectively. Let $\nu^*(s)$ and $\xi^*(s)$ be the corresponding measures supported on A_1 and A_2 respectively. The following complementary slackness results hold for the optima of (P') and (D'):

$$v^*(s)\int_{A_1} d\xi^*(s) = \int_{A_2}\int_{A_1} r(s,a_1,a_2)d\xi^*(s)d\nu^*(s) + \beta\sum_{s'} v^*(s')\int_{A_1} p(s';s,a_1)d\xi^*(s) \quad \forall s \in \mathcal{S} \tag{2}$$

$$\alpha^*(s)\int_{A_2} d\nu^*(s) = \int_{A_2}\int_{A_1} r(s,a_1,a_2)d\xi^*(s)d\nu^*(s) \quad \forall s \in \mathcal{S}. \tag{3}$$

Proof: The result follows from the strong duality of the equivalent semidefinite representations of the primal-dual pair $(P')-(D')$. The Lagrangian function for (P') is given by:

$$\mathcal{L}(\xi,\alpha) = \inf_{\mathbf{v},\nu}\{\sum_{s=1}^{S} v(s) - \int_{A_1}[v(s) - \int_{A_2} r(s,a_1,a_2)d\nu(s) - \beta\sum_{s'} v(s')p(s';s,a_1)]d\xi(s) + \sum_s \alpha(s)(1-\nu_0(s))\}.$$

$\mathcal{L}(\xi,\alpha)$ must satisfy weak duality, i.e. $d^* \le p^*$. At optimality $p^* = \sum_s v^*(s)$ for some vector $\mathbf{v}^*$. However, strong duality holds, i.e. $p^* = d^*$. This forces the first complementary slackness relation. The second relation is obtained similarly by considering the Lagrangian of the dual problem. ■

C. Main Theorem

Let p^* be the optimal value of (P'), and d^* be the optimal value of (D'). Let $\nu^*(s)$ and $\xi^*(s)$ be the optimal measures recovered in (P') and (D'). Let

$$\mu^*(s) = \frac{\xi^*(s)}{\int_{A_1} d\xi^*(s)}.$$

so that μ^* is a normalized version of ξ^* (i.e. μ^* is a probability measure). Let $\mathbf{v}^*$ be the the vector of value functions obtained via the optimal solution of (P').

Theorem 1: The optimal solutions to the primal-dual pair (P'), (D') satisfy the following:

1) $p^* = d^*$.
2) $\mathbf{v}^* = v_\beta(\mu^*,\nu^*)$.
3) $\mathbf{v}_\beta(\mu^*,\nu^*)$ satisfies the saddle-point inequality:

$$\mathbf{v}_\beta(\mu,\nu^*) \le \mathbf{v}_\beta(\mu^*,\nu^*) \le \mathbf{v}_\beta(\mu^*,\nu) \tag{4}$$

for all mixed strategies μ,ν.

Proof:

1) Follows from the strong duality of the primal-dual pair $(P')-(D')$.
2) Using Lemma 6 equation (3) in normalized form (i.e. dividing throughout by $\xi_0^*(s)$) we obtain

$$v^*(s) = \int_{A_2}\int_{A_1} r(s,a_1,a_2)d\mu^*(s)d\nu^*(s) + \beta\sum_{s'} v^*(s')\int_{A_1} p(s';s,a_1)d\mu^*(s) \quad \forall s \in \mathcal{S}.$$

Upon simplification and vectorization of $v^*(s)$ one obtains

$$\mathbf{v}^* = r(\mu^*,\nu^*) + \beta P(\mu^*)\mathbf{v}^*.$$

Using a Bellman equation argument or by simply iterating this equation (i.e. substituting repeatedly for $\mathbf{v}^*$) it is easy to see that $\mathbf{v}^* = \mathbf{v}_\beta(\mu^*,\nu^*)$.
3) Consider inequality (c) it at its optimal value. We have for every state s:

$$v^*(s) \ge \int_{a_2\in A_2} r(s,a_1,a_2)d\nu^*(s) + \beta\sum_{s'\in\mathcal{S}} p(s';s,a_1)v^*(s').$$

Integrating with respect to some arbitrary probability measure $\mu(s)$ (with support on A_1), we get:

$$v^*(s) \ge \int_{A_2}\int_{A_1} r(s,a_1,a_2)d\mu(s)d\nu^*(s) + \beta\sum_{s'\in\mathcal{S}}\int_{A_1} p(s';s,a_1)v^*(s')d\mu(s).$$

Thus,

$$v^*(s) \ge r(s,\mu(s),\nu^*(s)) + \beta\sum_{s'\in\mathcal{S}}\int_{A_1} p(s';s,a_1)v^*(s')d\mu(s).$$

Iterating this equation, we obtain $\mathbf{v}_\beta(\mu^*,\nu^*) = \mathbf{v}^* \ge \mathbf{v}_\beta(\mu,\nu^*)$ for every strategy μ. This completes one side of the saddle point inequality.

Using the normalized version of equation (4), we get:

$$\frac{\alpha^*(s)}{\xi_0^*(s)} = \int_{A_2}\int_{A_1} r(s,a_1,a_2)d\mu^*(s)d\nu^*(s) = r(s,\mu^*(s),\nu^*(s)).$$

If we integrate inequality (e) in problem (D') with

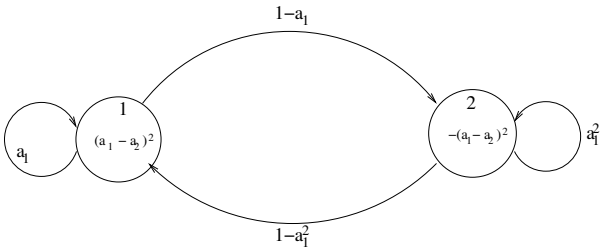

Fig. 1. A two state stochastic game with transition probabilities dependent only on the action of player 1. The payoffs associated to the states are indicated in the corresponding nodes. The edges are marked by the corresponding state transition probabilities.

respect to any arbitrary probability measure $\nu(s)$ with support on A_2 we obtain

$$\frac{\alpha^*(s)}{\xi_0^*(s)} \leq r(s, \mu^*(s), \nu(s)).$$

Thus $r(s, \mu^*(s), \nu^*(s)) \leq r(s, \mu^*(s), \nu(s))$ for every s. Multiplying throughout by $(I - \beta P(\mu^*))^{-1}$, we get $\mathbf{v}_\beta(\mu^*, \nu^*) \leq \mathbf{v}_\beta(\mu^*, \nu)$. This completes the other side of the saddle point inequality.

■

D. Example

Consider the two player discounted stochastic game with $\beta = 0.5$, $\mathcal{S} = \{1, 2\}$ with payoff function $r(1, a_1, a_2) = (a_1 - a_2)^2$ and $r(2, a_1, a_2) = -(a_1 - a_2)^2$. Let the probability transition matrix be given by:

$$P(a_1) = \begin{bmatrix} a_1 & 1 - a_1 \\ 1 - a_1^2 & a_1^2 \end{bmatrix}.$$

The polynomial optimization problem that computes the minimax strategies and the equilibrium values is the following:

$$\begin{aligned}
&\text{Minimize } v(1) + v(2) \\
&v(1) \geq a_1^2 - 2a_1\nu_1(1) + \nu_2(1) + \\
&\qquad \beta(a_1 v(1) + (1 - a_1)v(2)) \quad \forall a_1 \in [0, 1] \\
&v(2) \geq -a_1^2 + 2a_1\nu_1(2) - \nu_2(2) + \\
&\qquad \beta((1 - a_1^2)v(1) + a_1^2 v(2)) \quad \forall a_1 \in [0, 1] \\
&[1, \nu_1(1), \nu_2(1)]^T, [1, \nu_1(2), \nu_2(2)]^T \in \mathcal{M}([0, 1]).
\end{aligned}$$

Solving the SDP and its dual we obtain the following optimal cost-to-go and optimal moment sequences:

$$\mathbf{v}^* = [.298, -.158]^T$$

$$\begin{aligned}
\bar{\mu}^*(1) &= [1, .614, .614]^T \quad \bar{\mu}^*(2) = [1, .5, .25]^T \\
\bar{\nu}^*(1) &= [1, .614, .377]^T \quad \bar{\nu}^*(2) = [1, .614, .614]^T.
\end{aligned}$$

The corresponding measures obtained using standard techniques are supported at only finitely many points and are given by the following:

$$\begin{aligned}
\mu^*(1) &= .386\ \delta(a_1) + .614\ \delta(a_1 - 1) \\
\mu^*(2) &= \delta(a_1 - .5) \\
\nu^*(1) &= \delta(a_2 - .614) \\
\nu^*(2) &= .386\ \delta(a_2) + .614\ \delta(a_2 - 1).
\end{aligned}$$

IV. CONCLUSION AND FUTURE WORK

In this paper, we have presented a technique for solving two-player, zero-sum finite state stochastic games with infinite strategies and polynomial payoffs when the single-controller assumption holds. We show that the problem can be reduced to solving a system of univariate polynomial inequalities and moment constraints. We use techniques from the classical theory of moments and sum-of-squares to reduce the problem to a semidefinite programming problem. By solving a primal-dual pair of semidefinite programs, we obtain minimax equilibria and optimal strategies for the players.

It is known that finite-state, finite action, two-player zero-sum games which satisfy the *orderfield* property [12] may be solved via linear programming. The single-controller case, games with perfect information, switching controller stochastic games, separable reward-state independent transition (SER-SIT) games and additive games satisfy this property. We intend to extend these cases to the infinite strategy case with polynomial payoffs. General finite action stochastic games which do not satisfy the orderfield property are still amenable to computation via value iteration type techniques from dynamic programming. We plan to extend these results to the polynomial case as well.

ACKNOWLEDGEMENT

The authors would like to thank Ilan Lobel and Prof. Munther Dahleh for pointing out the linear programming solution to single controller finite stochastic games.

REFERENCES

[1] D. Bertsekas, *Dynamic programming and optimal control Vol. I.* Athena Scientific, Belmont, MA, 2005.
[2] D. Fudenberg and J. Tirole, *Game theory.* MIT Press, Cambridge, MA 1991.
[3] J. Filar and K. Vrieze, *Competitive Markov decision processes.* Springer, New York, 1997.
[4] L.S. Shapley. "Stochastic games", *Proceedings of the National Academy of Sciences*, USA, 39:1095-1100, 1953.
[5] Parthasarathy T. and Raghavan T. E. S., "An orderfield property for stochastic games when one player controls transition probabilities," *J. Optimization Theory,* Vol. 33, No. 3, pp. 375-392, Mar. 1981.
[6] P. A. Parrilo, "Polynomial games and sum of squares optimization," Proceedings of the 45th IEEE Conference on Decision and Control, Dec. 2006.
[7] N. D. Stein, A. Ozdaglar, and P. A. Parrilo, "Separable and low-rank continuous games," Proceedings of the 45th IEEE Conference on Decision and Control, Dec. 2006.
[8] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM Journal of Optimization*, Vol 11, No. 3, pp.796-817, 2000.
[9] P. A. Parrilo. *Structured semidefinite programs and semi-algebraic geometry methods in robustness and optimization.* PhD thesis, California Institute of Technology, May 2000.
[10] S. Karlin and L. Shapley. *Geometry of moment spaces,* vol. 12 of *Memoirs of the American Mathematical Society.* AMS 1953.
[11] J. A. Shohat and J. D. Tamarkin. *The problem of moments.* American Mathematical Society surveys, vol. II. AMS, New York, 1943.
[12] Raghavan T. E. S. and J. Filar, "Algorithms for stochastic games-a survey," *Methods and Models of Operations Research.* 35:437-472, 1991.