# Polynomial stochastic games via sum of squares optimization

Parikshit Shah
University of Wisconsin, Madison, WI
email: pshah@discovery.wisc.edu

Pablo A. Parrilo
Massachusetts Institute of Technology, Cambridge, MA
email: parrilo@mit.edu

Stochastic games are an important class of problems 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 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.

*Key words*: dynamic games; polynomial optimization

*MSC2000 Subject Classification*: Primary: 90C40 ; Secondary: 90C22

*OR/MS subject classification*: Primary: stochastic games ; Secondary: nonlinear programming, applications

---

**1. 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 problems where several (possibly adversarial) decision makers make individual decisions to optimize their own payoff [2]. In this paper we study *stochastic games* [3], a framework that combines 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 current payoff, but also the transition to the next state.

Notions of solutions in games have been extensively studied, and are very well understood. The most popular notion of a solution in game theory is that of a *Nash equilibrium*. While these equilibria are hard to compute in general, in certain cases they may be computed efficiently. For games involving two players and finite action spaces, mixed strategy minimax equilibria always exist (see, e.g., [2]). These minimax saddle points correspond to the well-known notion of a Nash equilibrium. From a computational standpoint such games are considered tractable because Nash equilibria 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 minimax 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. In 1981 Parthasarathy and Raghavan [5, 3] studied single controller games. Single controller games are games where the probabilities of transitions are controlled by the action of only one player. They showed that stochastic games satisfying this property could be solved efficiently via linear programming (thus proving that such problems with rational data could be computed in a finite number of steps).

While computational techniques for finite games are reasonably well understood, there has been some recent interest in the class of *infinite games*; see [6, 7] and the references therein. 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 main contribution of this paper is to provide a computationally efficient, finite dimensional characterization of the solution of single-controller polynomial stochastic games. For this, we extend the

linear programming formulation that solves the finite action single-controller stochastic game (i.e., under assumption (SC) below), to an infinite dimensional optimization problem when the actions are uncountably infinite. We furthermore establish the following properties of this infinite dimensional optimization problem:

(i) Its optimal solutions correspond to minimax equilibria.

(ii) The problem can be solved efficiently by semidefinite programming.

Section 2 of this paper provides a formal description of the problem and introduces the basic notation used in the paper. We show that for two-player zero-sum polynomial stochastic games, equilibria exist and that the corresponding equilibrium value vector is unique. (This proof is essentially an adaptation of the original proof by Shapley in [4] for finite stochastic games). In Section 2 we also briefly review some elegant results about polynomial nonnegativity, moment sequences of nonnegative measures, and their connection to semidefinite programming. Section 3 states and proves the main result of this paper. In Section 4 we present an example of a two-player, two-state stochastic game, and compute the equilibria via semidefinite programming. Finally, in Section 5 we state some natural extensions of this problem, conclusions, and directions of future research.

## 2. Problem description

**2.1 Stochastic games** 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 that make 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 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:

(i) The (finite) state space $\mathcal{S} = \{1, \ldots, S\}$.

(ii) The sets of actions for players 1 and 2 given by $A_1$ and $A_2$.

(iii) 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).

(iv) 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.

(v) The discount factor $\beta$, where $0 \leq \beta < 1$.

To fix ideas, consider the following example of a two-state stochastic game (i.e. $\mathcal{S} = \{1, 2\}$). The action spaces of the two players are $A_1 = A_2 = [0, 1]$. The payoff function in state 1 is $r(1, a_1, a_2) = r_1(a_1, a_2)$ and the payoff function in state 2 is given by $r(2, a_1, a_2) = r_2(a_1, a_2)$. Both are assumed to be polynomials in $a_1$ and $a_2$. The probability transition matrix is:

$$P = \begin{bmatrix} p_{11}(a_1, a_2) & p_{12}(a_1, a_2) \\ p_{21}(a_1, a_2) & p_{22}(a_1, a_2) \end{bmatrix}.$$

Every entry in this matrix is assumed to be a polynomial in $a_1$ and $a_2$. An example of a stochastic game is depicted graphically as shown in Fig. 1. We will return to this specific example in Section 4, where we explicitly solve for the equilibrium strategies of the two players. Through most of 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. This probability is *independent of the action of player* 2, and $p(s'; s, a_1, a_2) = p(s'; s, a_1)$. This is known as the *single-controller assumption.*

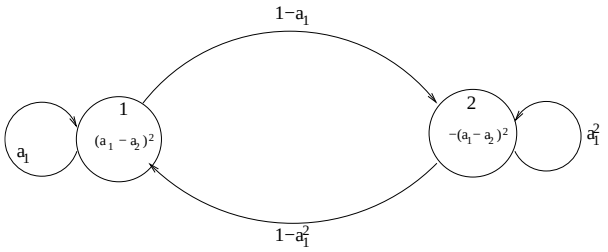


Figure 1: A two state stochastic game. The payoff functions associated to the states are denoted by $r_1$ and $r_2$. The edges are marked by the corresponding state transition probabilities.

In this paper we will be concerned with the case where the action spaces $A_1$ and $A_2$ of the two players 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 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: $r(s, a_1, a_2) = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} r_{ij}(s) a_1^i a_2^j$. 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 $\nu = [\nu(1), \ldots, \nu(S)]$. (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 $\xi$, $\mu$, $\nu$ will be used to denote measures. Subscripts on these Greek letters will be used to denote moments of the measures. A bar over a greek letter indicates a (finite) moment sequence (the length of the sequence being clear from the context). For example $\xi_j(i)$ denotes the $j^{th}$ moment of the measure $\xi$ corresponding to state $i$, and $\bar{\xi}(i) = [\xi_0(i), \ldots, \xi_n(i)]$).

A strategy $\mu$ leads to a probability transition 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 $\mu$, the probability transition matrix is fixed, and can be obtained by integrating each entry in the matrix with respect to the measure $\mu$. (Since the entries are polynomials, upon integration, these entries depend affinely on the moments $\mu(i)$). Given strategies $\mu$ and $\nu$, the expected 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 $\mathbf{r}(\mu, \nu) = [r(1, \mu(1), \nu(1)), \ldots, r(S, \mu(S), \nu(S))] \in \mathbb{R}^S$.

**2.2 Solution Concept and Existence of Equilibria** We now briefly discuss the question: "What is a reasonable solution concept for stochastic games?" Recall that for zero-sum normal form games, a Nash equilibrium is a widely used notion of equilibrium in competitive scenarios. It is also well-known that Nash equilibria (or equivalently saddle points) correspond to the minimax notion of an equilibrium,

i.e. points that satisfy the following equality:

$$\min_{\mu}\max_{\nu} v(\mu,\nu) = \max_{\nu}\min_{\mu} v(\mu,\nu).$$

While there may exist no pure strategies that satisfy this equality, it may be achieved by allowing randomization over the allowable strategies. In his seminal paper [4], Shapley generalized the notion of Nash equilibria to stochastic games. He defined the notion of a "stationary equilibrium" to be a pair of randomized strategies (over the action space) that depended only on the state of the game. (Of course, to be an equilibrium, these mixed strategies must also satisfy the no-deviation principle). For stochastic games, once one restricts attention to stationary equilibria, instead of having unique "values" (as in normal form games), one has a unique "value vector". This vector is indexed by the state and the $i^{th}$ component is interpreted as the equilibrium value Player 1 can expect to receive (over the infinite discounted process) conditioned on the fact that the game starts in state $i$. Note that different states of the game may be favorable to different players. Since the actions affect both payoffs and state transitions, players must balance their strategies so that they receive good payoffs in a particular state along with favorable state transitions. The "no unilateral deviation" principle, saddle point inequality (interpreted row-wise, i.e., conditioned upon a particular state) and the equivalence of the minmax and maxmin over randomized strategies all extend to the stochastic game case, and when we restrict attention to games with just one state, we recover the classical notions of equilibrium.

DEFINITION 2.1 *A pair of vector of mixed strategies (indexed by the state) $\mu^0$ and $\nu^0$ which satisfy the saddle point property:*

$$\mathbf{v}_\beta(\mu,\nu^0) \le \mathbf{v}_\beta(\mu^0,\nu^0) \le \mathbf{v}_\beta(\mu^0,\nu) \tag{1}$$

*for all (vectors of) mixed strategies $\mu,\nu$ are called* equilibrium strategies. *The corresponding vector $\mathbf{v}_\beta(\mu^0,\nu^0)$ is called the* value vector *of the game.*

One should note that $\mathbf{v}_\beta(\mu,\nu)$ is a vector in $\mathbb{R}^S$ indexed by the initial state of the Markov process.

In his original paper, Shapley [4] showed that stationary equilibria always exist (and that the corresponding value-vectors are unique) for two-player, zero-sum, finite state, finite action stochastic games. The issue of existence and uniqueness of value vectors is well-studied, indeed, under fairly weak conditions it has been shown that zero-sum stochastic games have a saddle point solution [8]. Throughout the paper, we assume that the transition probabilities are polynomial functions of the actions of the players. It is important to note that the results of this subsection *do not depend upon the single-controller assumption.* As a by-product of the proof of existence and uniqueness, one obtains an algorithm for computing equilibria for such games [9]. This algorithm is analogous to *value-iteration* in dynamic programming, and consists of solving a sequence of simple (non-stochastic) games whose value-vectors converge to the true value vector.

For a two-player zero-sum polynomial game with payoff function $p(x,y)$ and strategy space (of both players) $A=[0,1]$ it can be shown that a mixed-strategy Nash equilibrium always exists, the game has a unique value and that the optimal strategies can be computed by semidefinite programming [6], [10].

Given a polynomial stochastic game with payoff functions $r(s,a_1,a_2)$ and transition probabilities $p(t;s,a_1,a_2)$ (sometimes we will hide the state indices and write the entire matrix as $P(a_1,a_2)$), fix a state $s$ and define the polynomial $G^s(\alpha) = r(s,a_1,a_2) + \beta\sum_{t\in\mathcal{S}} p(t;s,a_1,a_2)\alpha_t$. One may perform iterations using this vector $\alpha\in\mathbb{R}^S$. We call the iterates of these vectors $\alpha^k\in\mathbb{R}^S$ ($k$ is the iteration index), and denote the $s^{th}$ component of this vector by $\alpha^k_s$. Define the operator $T_s:\mathbb{R}^S\to\mathbb{R}$ to be $T_s\alpha = \mathrm{val}(G^s(\alpha))$. Let $T\alpha = [T_1\alpha,\dots T_S\alpha]^T$. Starting with an arbitrary vector $\alpha^0\in\mathbb{R}^S$ we execute a recursion which consists of computing the vectors $T^k(\alpha)$. Note that each step of the recursion consists of solving a zero-sum polynomial game, which may be accomplished by solving a single semidefinite program. The following theorem establishes the existence of a value vector for the class of games under consideration. The proof is a straightforward adaptation of the original proof of Shapley for finite stochastic games.

THEOREM 2.1 *Consider a zero sum stochastic game with polynomial payoff functions $r(s,a_1,a_2)$, polynomial transition probabilities $p(s';s,a_1,a_2)$ and discount factor $\beta\in(0,1)$. Let $A_1$ and $A_2$, the action spaces of players* 1 *and* 2 *respectively, be compact.*

(*i*) *The operator $T$ has a unique fixed point $\phi$ such that $T\phi = \phi$.*

(*ii*) *$\phi$ is the unique value vector of the stochastic game.*

(*iii*) *The optimal strategies of the stochastic game correspond to the optimal strategies of the auxiliary (one-shot) game $G^s(\phi)$.*

PROOF. The theorem holds under more general conditions, a proof may be found in [8, Prop. 5.2] and also [9]. □

Note that to algorithmically compute approximate equilibria, it is sufficient to iterate the operator $T_s(\alpha)$ (each step involves solving a polynomial game in normal form using SDP), and by solving a sequence of such problems, one can compute $T^k(\alpha)$ within a provable accuracy level. The rate of convergence of this iteration is not very attractive and in the rest of this paper we focus attention on *single-controller games*, for which equilibria can be computed by solving a single semidefinite program.

**2.3 SDP Characterization of Nonnegativity and Moments** To be able to solve the class of games under consideration using SDP, we state the semidefinite representations of nonnegative of polynomials on intervals and also moment spaces. Let $A = [0, 1]$. 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 $\mathcal{P}(A)$. (We exclude the degree information in the notation for brevity.) 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 $p(x) = \sum_{i=1}^{k} q_i(x)^2$. 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 $\mu$ denote a measure supported on the set $A$. The $i^{th}$ moment of the measure $\mu$ 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 $\mu$ 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. A nonnegative measure $\mu$ is a *probability measure* if $\mu_0 = 1$. 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. We define two linear operators $\mathcal{H} : \mathbb{R}^{2n-1} \to \mathcal{S}^n$ and $\mathcal{H}^* : \mathcal{S}^n \to \mathbb{R}^{2n-1}$:

$$\mathcal{H} : \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{2n-1} \end{bmatrix} \mapsto \begin{bmatrix} a_1 & a_2 & \ldots & a_n \\ a_2 & a_3 & \ldots & a_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_n & a_{n+1} & \ldots & a_{2n-1} \end{bmatrix} \qquad \mathcal{H}^* : \begin{bmatrix} m_{11} & m_{12} & \ldots & m_{1n} \\ m_{12} & m_{22} & \ldots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{1n} & m_{2n} & \ldots & m_{nn} \end{bmatrix} \mapsto \begin{bmatrix} m_{11} \\ 2m_{12} \\ m_{22} + 2m_{13} \\ \vdots \\ m_{nn} \end{bmatrix}.$$

Thus $\mathcal{H}$ is simply the linear operator that takes a vector and constructs the associated Hankel matrix which is constant along the antidiagonals and the adjoint $\mathcal{H}^*$ 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 the interval $[0, 1]$. We define the following matrices: $L_1 = \begin{bmatrix} I_{n\times n} & 0_{1\times n} \end{bmatrix}^T$ and $L_2 = \begin{bmatrix} 0_{1\times n} & I_{n\times n} \end{bmatrix}^T$, where $I_{n\times n}$ stands for the $n \times n$ identity matrix.

LEMMA 2.1 *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. The proof follows from the characterization of nonnegative polynomials on intervals. It is well known that

$$p(x) \geq 0 \quad \forall x \in [0, 1] \Leftrightarrow p(x) = z(x) + x(1 - x)w(x),$$

where $z(x)$ and $w(x)$ are sums of squares. For further details, please see [9, 11]. □

In this paper, we will also be using a very important classical result about the semidefinite representation of moment spaces [12, 13]. We give an explicit characterization of $\mathcal{M}([0,1])$ and $\mathcal{M}_P([0,1])$.

LEMMA 2.2 *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 $\mathcal{H}(\bar{\mu}) \succeq 0$ and $\frac{1}{2}(L_1^T\mathcal{H}(\bar{\mu})L_2 + L_2^T\mathcal{H}(\bar{\mu})L_1) - L_2^T\mathcal{H}(\bar{\mu})L_2 \succeq 0$. Moreover, it is a moment sequence corresponding to a probability measure if and only if in addition to it satisfies $\mu_0 = 1$.*

PROOF. The proof follows by dualizing Lemma 2.1. Alternatively, a direct proof may be found in [12]. □

For example, for $2n = 2$ the sequence $[\mu_0, \mu_1, \mu_2]$ is a moment sequence corresponding to a measure supported on $[0,1]$ if and only if the following inequalities are true:

$$\begin{bmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{bmatrix} \succeq 0 \text{ and } \quad \mu_1 - \mu_2 \geq 0.$$

## 3. Infinite Strategy Stochastic Games

**3.1 Preliminary Results** The zero-sum finite strategy single player stochastic game can be solved by linear programming. The following is the generalization of the linear program $(P)$ mentioned above:

$$\begin{array}{ll} & \text{minimize } \sum_{s=1}^{S} v(s) \\ & \nu(s), v(s) \\ (a) & v(s) \geq \int_{a_2 \in A_2} r(s,a_1,a_2)d\nu(s) + \beta \sum_{s'=1}^{S} p(s';s,a_1)v(s') \text{ for all } s \in \mathcal{S}, a_1 \in A_1 \\ (b) & \nu(s) \text{ is a measure supported on } A_2 \text{ for all } s \in \mathcal{S}. \end{array}$$

Since $\int r(s,a_1,a_2)d\nu(s) = q_\nu(s,a_1)$, a univariate polynomial in $a_1$ for each $s \in \mathcal{S}$, for a fixed vector $v(s)$, the constraints (a) are a system of polynomial inequalities. Note that the coefficients of $q$ will depend on the measure $\nu$ 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 the following problem.

$$\begin{array}{ll} & \text{minimize } \sum_{s=1}^{S} v(s) \\ & \bar{\nu}(s), v(s) \\ (c) & v(s) - \sum_{i,j} r_{ij}(s)a_1^i \nu_j(s) - \beta \sum_{s'=1}^{S} p(s';s,a_1)v(s') \in \mathcal{P}(A_1) \text{ for all } s \in \mathcal{S} \\ (d) & \bar{\nu}(s) \in \mathcal{M}(A_2), \text{ and } \nu_0(s) = 1 \text{ for all } s \in \mathcal{S}. \end{array} \qquad (P')$$

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 $r(s,a_1,a_2)$. Similar to the finite strategy case we define a vector by $\mathbf{v}^* = [v^*(1), \ldots, v^*(S)]^T$ which will turn out to be the value vector of the stochastic game (which is indexed by the 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 $p(s';s,a_1)$. Note that using the characterization of nonnegative polynomials and moments of measures, the problem $(P')$ has a semidefinite representation. We call this representation $(SP)$. We call its dual $(SD)$. For brevity, we do not write the explicit form of the SDPs, the interested reader may refer to [9].

LEMMA 3.1 *The dual of* $(P')$ *is equivalent to the following polynomial optimization problem:*

$$\begin{array}{ll} & \text{maximize } \sum_{s=1}^{S} \alpha(s) \\ & \alpha(s), \bar{\xi}(s) \\ (e) & \sum_{i,j} r_{ij}(s)\xi_i(s)a_2^j - \alpha(s) \geq 0 \quad \forall a_2 \in A_2, s \in \mathcal{S} \\ (f) & \bar{\xi}(s) \in \mathcal{M}(A_2) \quad \forall s \in \mathcal{S} \\ (g) & \sum_s \int_{A_1} (\delta(s,s') - \beta p(s',s,a_1)) d\xi(s) = 1 \quad \forall s' \in \mathcal{S}. \end{array} \tag{$D'$}$$

PROOF. This again follows as a consequence of Lemmas 2.1 and 2.2. □

REMARK 3.1 *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*.*

LEMMA 3.2 *The polynomial optimization problems* $(P')$ *and* $(D')$ *satisfy strong duality*

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 [12] that the moment sequence of $\mu$ is in the interior of the moment space of $[0,1]$. As a consequence, corresponding constraints in the SDP representations are strictly positive definite. Using the strategies $\mu$ and $\nu$, 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 corresponding SDP constraints 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 3.3 *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:

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

$\mathcal{L}(\xi,\alpha)$ must satisfy weak duality, i.e. $d^* \leq p^*$. 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. □

Having established the necessary machinery, we next show that the solution to problem $(P')$ is in fact the desired equilibrium solution.

**3.2 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 $\mu^*$ is a normalized version of $\xi^*$ (i.e. $\mu^*$ is a probability measure). Let $\mathbf{v}^*$ be the vector obtained as the optimal solution of $(P')$.

THEOREM 3.1 *The optimal solutions to the primal-dual pair* $(P')$, $(D')$ *satisfy the following:*

(*i*) $p^* = d^*$.

(*ii*) $\mathbf{v}^* = v_\beta(\mu^*, \nu^*)$.

(*iii*) $\mathbf{v}_\beta(\mu^*, \nu^*)$ *satisfies the saddle-point inequality:*

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

*for all mixed strategies* $\mu, \nu$.

PROOF. 1) Follows from the strong duality of the primal-dual pair $(P') - (D')$.

(ii) Using Lemma 3.3 equation (2) in normalized form (i.e. dividing throughout by $\xi_0^*(s)$, which is the zeroth order moment of the measure $\xi(s)$) 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 $\mathbf{v}^* = \mathbf{v}_\beta(\mu^*, \nu^*)$.

(iii) Consider inequality (c) at its optimal value. We have for every state $s$:

$$v^*(s) \geq \int_{a_2 \in A_2} r(s, a_1, a_2) d\nu^*(s) + \beta \sum_{s'=1}^{S} p(s'; s, a_1) v^*(s').$$

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

$$v^*(s) \geq \int_{A_2} \int_{A_1} r(s, a_1, a_2) d\mu(s) d\nu^*(s) + \beta \sum_{s'=1}^{S} \int_{A_1} p(s'; s, a_1) v^*(s') d\mu(s).$$

Thus, $v^*(s) \geq r(s, \mu(s), \nu^*(s)) + \beta \sum_{s'=1}^{S} \int_{A_1} p(s'; s, a_1) v^*(s') d\mu(s)$. Iterating this equation, we obtain $\mathbf{v}_\beta(\mu^*, \nu^*) = \mathbf{v}^* \geq \mathbf{v}_\beta(\mu, \nu^*)$ for every strategy $\mu$. This completes one side of the saddle point inequality.

Using the normalized version of equation (5), 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 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.

□

**3.3 Obtaining the measures** Solutions to the semidefinite programs $(SP)$ and $(SD)$ provide the moment sequences corresponding to optimal strategies. Additional computation involving some linear algebraic operations are required to recover the actual measures [13], [14], [9]. We briefly outline this standard computational method.

Let $\bar{\mu} \in \mathbb{R}^{2n}$ be a given moment sequence. We wish to find a nonnegative measure $\mu$ supported on the real line with these moments. The resulting measure will be composed of finitely many atoms (i.e. a discrete measure) of the form $\sum w_i \delta(x - a_i)$ where

$$\mathbf{Prob}(x = a_i) = w_i \quad \forall i.$$

Construct the following linear system:

$$\begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_{n-1} \\ \mu_1 & \mu_2 & \cdots & \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-2} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{bmatrix} = - \begin{bmatrix} \mu_n \\ \mu_{n+1} \\ \vdots \\ \mu_{2n-1} \end{bmatrix}.$$

Note that the Hankel matrix that appears on the left hand side is a sub-matrix of $\mathcal{H}(\bar{\mu})$. We assume without loss of generality that the above matrix is strictly positive definite. (Suppose the above matrix is not full rank, construct a smaller $k \times k$ linear system of equations by eliminating the last $n-k$ rows and columns of the matrix so that the $k \times k$ submatrix is full rank, and therefore strictly positive definite.) By inverting this matrix we solve for $[c_0, \ldots, c_{n-1}]^T$. Let $x_i$ be the roots of the polynomial equation

$$x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0 = 0.$$

It can be shown that the $x_i$ are all real and distinct, and that they are the support points of the discrete measure. Once the supports are obtained, the weights $w_i$ may be obtained by solving the nonsingular Vandermonde system given by:

$$\sum_{i=1}^{n} w_i x_i^j = \mu_j \quad (0 \leq j \leq n-1).$$

**4. 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 $r(2, a_1, a_2) = -(a_1 - a_2)^2$. Figure 1 graphically illustrates this stochastic game, consisting of two states (the nodes) with polynomial transition probabilities dependent on $a_1$ (as marked on the edges of the graph). Within the nodes, the payoffs associated to the corresponding states are indicated. Let the probability transition matrix be given by:

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

To understand this game, consider first the zero-sum (non-stochastic normal form game) with payoff function $p(a_1, a_2) = (a_1 - a_2)^2$ over the strategy space $[0, 1]$. This game (called the "guessing game") was studied by Parrilo in [6]. If Player 2 is able to guess the action of Player 1, he can simply imitate his action (i.e. set $a_2 = a_1$ and his payoff to player 1 would be zero (this is the minimum possible since $(a_1 - a_2)^2 \geq 0$). Player 1 would try to confuse player 2 as much as possible and thus randomize between the extreme actions $a_1 = 0$ and $a_1 = 1$ with a probability of $\frac{1}{2}$. Player 2's best response would be to play $a_2 = \frac{1}{2}$ with probability 1.

In the game described in Fig. 1, in State 1 Player 1 plays the role of confuser and Player 2 plays the role of guesser. In state 2, the roles of the players are reversed, Player 1 is the guesser and Player 2 the confuser. However, the problem is complicated a bit by the fact that State 1 is advantageous to Player 1 so that at every stage he has incentive to play a strategy that gives him a good payoff as well as maximize the chances of transitioning to State 1.

Solving the SDP and its dual corresponding to this example, we obtain the following the value vector to be $\mathbf{v}^* = [.298, -.158]^T$ and optimal moment sequences:

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

The corresponding measures obtained as explained in subsection 3.3 are supported at only finitely many points, and are given by the following:

$$\mu^*(1) = .386\ \delta(a_1) + .614\ \delta(a_1 - 1) \quad \mu^*(2) = \delta(a_1 - .5) \quad \nu^*(1) = \delta(a_2 - .614) \quad \nu^*(2) = .386\ \delta(a_2) + .614\ \delta(a_2 - 1).$$

Consider, for example, play in State 1. If Player 1 were playing obliviously with respect to the state transitions, he would play actions $a_1 = 0$ and $a_1 = 1$ with one half probability each. However, to increase the probability of staying in State 1 he plays action 1 with a higher probability. Player 2 cannot affect the state transition probabilities directly, thus he must play a myopic best response. (A myopic best response is one that is a best response for the game in the current state). Note that in state 1, once Player 1's strategy is fixed, the (only) best response for Player 2 is to play the action $a_2 = 0.614$ with probability 1. In state 2, player 1's best strategy is to play $a_1 = 0.5$. Player 2 picks an action from his myopic best response set (in this case, all probability distributions that are supported on the points 0 and 1).

**5. Conclusions 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. We established the existence of equilibria for such games. As a by-product we got an algorithm that converged to unique value vector of the game (however this algorithm does not seem to have very attractive convergence rates). We focused mainly on the case where the single-controller assumption holds. We showed

that the problem can be reduced to solving a system of univariate polynomial inequalities and moment constraints and be solved by semidefinite programming problem. By solving a primal-dual pair of semidefinite programs, we computed minimax equilibria and the optimal strategies.

**Acknowledgement:** The authors would like to thank Ilan Lobel and Prof. Munther Dahleh for bringing to their attention the linear programming solution to single controller finite stochastic games.

**References**

[1] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, 2005, vol. I.

[2] D. Fudenberg and J. Tirole, *Game theory*. Cambridge, MA: MIT Press, 1991.

[3] J. A. Filar and K. Vrieze, *Competitive Markov decision processes*. New York: Springer, 1997.

[4] L. S. Shapley, "Stochastic games," *Proc. Nat. Acad. Sci. U. S. A.*, vol. 39, pp. 1095–1100, 1953.

[5] T. Parthasarathy and T. E. S. Raghavan, "An orderfield property for stochastic games when one player controls transition probabilities," *J. Optim. Theory Appl.*, vol. 33, no. 3, pp. 375–392, 1981.

[6] P. A. Parrilo, "Polynomial games and sum of squares optimization," in *Proceedings of the 45$^{th}$ IEEE Conference on Decision and Control*, 2006.

[7] N. Stein, A. Ozdaglar, and P. A. Parrilo, "Separable and low-rank continuous games," in *Proceedings of the 45$^{th}$ IEEE Conference on Decision and Control*, 2006.

[8] S. Sorin, *A first course on zero-sum repeated games*. Berlin: Springer-Verlag, 2002.

[9] P. Shah and P. A. Parrilo, "Polynomial stochastic games via sum of squares optimization," *arXiv:0806.2469v1*, 2008.

[10] M. Dresher, S. Karlin, and L. S. Shapley, "Polynomial games," in *Contributions to the Theory of Games*, ser. Annals of Mathematics Studies, no. 24. Princeton, N. J.: Princeton University Press, 1950, pp. 161–180.

[11] P. A. Parrilo, "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization," Ph.D. dissertation, California Institute of Technology, May 2000.

[12] S. Karlin and L. Shapley, *Geometry of moment spaces*, ser. Memoirs of the American Mathematical Society. AMS, 1953, vol. 12.

[13] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, ser. American Mathematical Society Mathematical surveys, vol. II. New York: American Mathematical Society, 1943.

[14] L. Devroye, *Nonuniform random variate generation*. New York: Springer-Verlag, 1986.