Mathematical Programming 50 (1991) 227-237
North-Holland

Nonlinear programming and stationary equilibria in stochastic games

J.A. Filar*

Department of Mathematics, University of Maryland, Baltimore County, MD 21228, USA

T.A. Schultz

School of Business Administration, Augusta College, Rock Island, IL 61201, USA

F. Thuijsman** and O.J. Vrieze

Department of Mathematics, University of Limburg, P.O. Box 616, 6200 MD Maastricht, Netherlands

Received 23 November 1987
Revised manuscript received 25 May 1989

Stationary equilibria in discounted and limiting average finite state/action space stochastic games are shown to be equivalent to global optima of certain nonlinear programs. For zero sum limiting average games, this formulation reduces to a linear objective, nonlinear constraints program, which finds the "best" stationary strategies, even when ε-optimal stationary strategies do not exist, for arbitrarily small ε.

Key words: Stochastic game theory.

Introduction

In 1964 Mangasarian and Stone [12] showed an interesting connection between quadratic programming and bimatrix games. In particular, they constructed a quadratic program which has a global maximum of zero, and showed that the optima of this program form Nash equilibria of the bimatrix game in question. Despite encouraging numerical results reported in [12], this mathematical programming approach to the computation of Nash equilibria of non-cooperative games appears to be all but forgotten. Of course, in the context of bimatrix games, the results of [12] were superseded by the finite algorithm of Lemke and Howson [11]. The latter algorithm, however, is not easily extended to general N-person games (see [14, 20]

* The work of the first author was supported in part by the Air Force Office of Scientific Research, and by the National Science Foundation under Grant No ECS-8704954.

** The work of the third author was supported by The Netherlands Organization for Scientific Research NWO, project 10-64-10.

for such an extension), whereas the method of Mangasarian and Stone [12] can be extended to the N-person case, at the cost of only increasing the order of the polynomials in the objective function and the constraints.

In this paper we demonstrate that the basic conceptual approach of [12] can further be extended to infinite horizon stochastic games with either the discounted or the limiting average payoff criteria. In particular, by combining appropriate properties of Markov chains with a nonlinear program in the spirit of Mangasarian and Stone [12], we are able to characterize all stationary Nash equilibria of these games (when they exist) as global minima of that program. In doing so we solve one of the main open problems of stochastic games, as stated by Bewley and Kohlberg [1]. In the process, we clarify a characterization due to Sobel [17] of stationary equilibria in the average reward stochastic games.

Our results strengthen and supersede earlier attempts at characterization of stationary equilibria (see, e.g., [2, 6, 16, 19]).

For the zero sum case we also demonstrate that even when stationary optimal strategies fail to exist, the global optima of our nonlinear programs supply the "best" stationary strategies.

Finally, it ought to be mentioned that, perhaps, the first nonlinear programming formulation of stochastic games was due to Rothblum [15].

1. Preliminaries and notation

A *stochastic game* Γ is defined by the 5-tuple $\langle S, K, A, R, Q\rangle$, where S is a finite set of states, K is a finite set of players, $A=\{A_s^k \mid k\in K,\ s\in S\}$ consists of finite sets of actions for each player k in each state s, and R and Q, described below, respectively denote a reward and transition law. Given a starting state $s\in S$, play proceeds through stages $t=0, 1, 2, \ldots$, and based on the choices of actions by the players, rewards (according to R) are received at each stage, and transitions (according to Q) among the states take place. The players control the course of the game through strategies, and a relatively simple class of strategies are the *stationary strategies*. A *stationary strategy* for player $k\in K$, denoted $f^k\in F_S^k=\{f_s^k(a)\mid s\in S,\ a\in A_s^k\}$, describes the probability with which player $k\in K$ chooses action $a\in A_s^k$ whenever the game is in state $s\in S$. When each player $k\in K$ fixes a stationary strategy $f^k\in F_S^k$, a *(joint) stationary strategy* $f\in F_S^K=\{f^k\mid k\in K, f^k\in F_S^k\}$ is defined. For a joint stationary strategy $\hat{f}\in F_S^K$, and for some player $k\in K$ and stationary strategy $f\in F_S^k$, the joint stationary strategy $\hat{f}^{\langle k,s,f\rangle}$ denotes the strategy where all players choose actions dictated by $\hat{f}$, except that player k, when the game is in state $s\in S$, chooses actions dictated by f. Similarly, $\hat{f}^{\langle k,f\rangle}$ denotes the strategy where player k plays according to f in all states, and the other players according to $\hat{f}$. For a player $k\in K$, state $s\in S$, and action $a\in A_s^k$, the joint stationary strategy $\hat{f}^{\langle k,s,a\rangle}$ indicates that player k chooses action a with probability 1 when the game is in state s, and that otherwise all players choose actions dictated by $\hat{f}$.

The reward law $R = \{r_s^k(f) \mid f \in F_S^K, k \in K, s \in S\}$ denotes the (immediate or one-stage) expected payoff to player $k \in K$ if the game is in state $s \in S$. Similarly, the transition law $Q = \{q(s' \mid s, f) \mid f \in F_S^K, s, s' \in S\}$ denotes the probability that a transition occurs from state $s \in S$ to state $s' \in S$ when the strategy $f \in F_S^K$ is used. It should be clear from these definitions that

$$r_s^k(f) = \sum_{a \in A_s^k} f_s^k(a) r_s^k(f^{\langle k,s,a \rangle}) \quad \forall k \in K, s \in S, f \in F_S^K, \tag{1.1}$$

and

$$q(s' \mid s, f) = \sum_{a \in A_s^k} f_s^k(a) q(s' \mid s, f^{\langle k,s,a \rangle}) \quad \forall k \in K, s \in S, f \in F_S^K. \tag{1.2}$$

Further, it should be clear that in order to define the reward (or transition) law, it is sufficient to define rewards to each player (or transitions to each state) in a state $s \in S$ for "joint actions" in the finite set $\bigtimes_{k \in K} A_s^k$; the terms $r_s^k(f)$ (or $q(s' \mid s, f)$) are weighted averages of these pure rewards (or transitions).

In order to evaluate the game's total return, the players aggregate the stream of stage-by-stage rewards. Specifically, given a fixed joint stationary strategy $f \in F_S^K$ and a starting state $s \in S$, let $E_s^k(f, t)$ denote the expected payoff to player $k \in K$ in the tth stage when the starting state is s and the joint strategy f is followed. The *β-discounted* (*game*) *payoff* to player $k \in K$ when the game starts in state $s \in S$ is defined as (for $\beta \in [0, 1)$)

$$\phi_s^k(\beta; f) = \sum_{t=0}^{\infty} \beta^t E_s^k(f, t) \tag{1.3}$$

and the *limiting average* (*game*) *payoff* is defined as

$$\phi_s^k(f) = \lim_{T \uparrow \infty} \frac{1}{T+1} \sum_{t=0}^{T} E_s^k(f, t). \tag{1.4}$$

For non-stationary strategies, the limit in (1.4) may not exist; in this case it is customary to take the lim inf.

A joint strategy $\hat{f} \in F_S^K$ forms a *β-discounted* (*stationary*) *equilibrium* with *equilibrium payoffs* $\hat{v} \in \{v_s^k \mid k \in K, s \in S\}$ if

$$\hat{v}_s^k = \phi_s^k(\beta; \hat{f}) \geq \phi_s^k(\beta; \hat{f}^{\langle k,f \rangle}) \quad \forall k \in K, s \in S, f \in F_S^k,$$

and forms a *limiting average* (*stationary*) *equilibrium* if

$$\hat{v}_s^k = \phi_s^k(\hat{f}) \geq \phi_s^k(\hat{f}^{\langle k,f \rangle}) \quad \forall k \in K, s \in S, f \in F_S^k.$$

A one player game is equivalent to the well-known Markov Decision Process (MDP), and we make use of the considerable literature for MDPs. It can be easily verified that for an equilibrium $\hat{f} \in F_S^K$, each $\hat{f}^k \in F_S^k$ for $k \in K$ forms an optimal maximizing strategy in the MDP obtained from the game where all the remaining players $l \in K$, $l \neq k$, fix their strategies at $\hat{f}^1$ (see, e.g., [5, 10]). For MDPs it is known that $\hat{f}$ is an optimal stationary strategy (usually called a policy) for the single player (usually called the decision maker) in the β-discounted MDP, and $\hat{v}_s = \phi_s(\beta; \hat{f})$ $\forall s \in S$ (dropping the superscript k) is the optimal β-discounted payoff starting in state s, if and only if

$$\hat{v}_s = r_s(\hat{f}) + \beta \sum_{s' \in S} \hat{v}_{s'} q(s' | s, \hat{f}) \quad \forall s \in S, \tag{1.5}$$

$$\hat{v}_s \geqslant r_s(\hat{f}^{\langle s,a \rangle}) + \beta \sum_{s' \in S} \hat{v}_{s'} q(s' | s, \hat{f}^{\langle s,a \rangle}) \quad \forall s \in S,\, a \in A_s, \tag{1.6}$$

where $\hat{f}^{\langle s,a \rangle}$ indicates that whenever in state $s \in S$, the decision maker chooses action $a \in A_s$ with probability 1. A strategy $\hat{f}$ is an optimal stationary policy in the limiting average MDP, and $\hat{v}_s = \phi_s(\hat{f})$ $\forall s \in S$ is the optimal limiting average payoff starting in state s if and only if (i) and (ii) hold:

(i) $$\hat{v}_s = \sum_{s' \in S} \hat{v}_{s'} q(s' | s, \hat{f}) \quad \forall s \in S \tag{1.7}$$

$$\geqslant \sum_{s' \in S} \hat{v}_{s'} q(s' | s, \hat{f}^{\langle s,a \rangle}) \quad \forall s \in S,\, a \in A_s. \tag{1.8}$$

(ii) There exist vectors $\hat{w}$, $\hat{t} \in R^S$ such that (cf. [3] and [9])

$$\hat{v}_s + \hat{w}_s = r_s(\hat{f}) + \sum_{s' \in S} \hat{w}_{s'} q(s' | s, \hat{f}) \quad \forall s \in S \tag{1.9}$$

and

$$\hat{v}_s + \hat{t}_s \geqslant r_s(\hat{f}^{\langle s,a \rangle}) + \sum_{s' \in S} \hat{t}_{s'} q(s' | s, \hat{f}^{\langle s,a \rangle}) \quad \forall s \in S,\, a \in A_s. \tag{1.10}$$

2. Discounted games

For non-zero sum β-discounted stochastic games, it is known that stationary equilibria exist. This result was independently discovered by Fink [7], Takahashi [18], Rogers [13] and Sobel [17].

The theorem presented below characterizes all stationary equilibria as global optima of a suitably constructed mathematical program.

Theorem 2.1. *Let $\Gamma = \langle S, K, A, R, Q \rangle$ be a stochastic game, and let $\hat{v} \in \{v_s^k \mid k \in K, s \in S\}$, and $\hat{f} \in F_S^K$ be given. The joint strategy $\hat{f}$ forms a β-discounted (Nash) equilibrium with equilibrium payoffs $\hat{v}$ if and only if the variables $(\hat{v}, \hat{f})$ are a global minimum (objective value will be zero) in the following nonlinear program NLP 2.2.*

NLP 2.2. Variables $v \in \{v_s^k \mid k \in K, s \in S\}$, $f \in \{f_s^k(a) \mid k \in K, s \in S, a \in A_s^k\}$;

$$\text{minimize} \quad \sum_{k \in K} \sum_{s \in S} \left[v_s^k - r_s^k(f) - \beta \sum_{s' \in S} v_{s'}^k q(s' | s, f) \right]$$

subject to

(i) $v_s^k \geqslant r_s^k(f^{\langle k,s,a\rangle}) + \beta \sum_{s' \in S} v_{s'}^k q(s' | s, f^{\langle k,s,a\rangle}) \quad \forall k \in K, s \in S, a \in A_s^k,$

(ii) $\sum_{a \in A_s^k} f_s^k(a) = 1 \quad \forall k \in K, s \in S,$

(iii) $f_s^k(a) \geqslant 0 \quad \forall k \in K, s \in S, a \in A_s^k.$

Proof of Theorem 2.1. First notice that by constraints (i), the global optimum of NLP 2.2 is at least zero. Let $\hat{f}$ be β-discounted stationary equilibrium with equilibrium payoffs $\hat{v}$. Since $\hat{f}$ is a stationary strategy, constraints (ii) and (iii) of NLP 2.2 are satisfied. At equilibrium, each player $k \in K$ solves a β-discounted MDP, so relations (1.5) and (1.6) hold for each player $k \in K$, which implies that $(\hat{v}, \hat{f})$ is feasible with objective value zero. Hence $(\hat{v}, \hat{f})$ is optimal.

Now, assume $(\hat{v}, \hat{f})$ is a global minimum to NLP 2.2. The zero objective (by the first part of this proof) and the constraints (i), (ii) and (iii) imply that (1.5) and (1.6) hold for each player $k \in K$, $\hat{v}^k$, and $\hat{f}^k$. Hence, $\hat{f}^k$ is optimal for player k in the MDP obtained when all the other players fix $\hat{f}^l$, $l \in K$, $l \neq k$. Therefore, $\hat{f}$ is an equilibrium. □

From a practical standpoint, an interesting feature of NLP 2.2 is that reduction of the objective to near zero implies that a near equilibrium has been found. To make this precise, for an $\varepsilon > 0$, a stationary strategy $\hat{f} \in F_S^K$ forms a *β-discounted ε-equilibrium* if

$$\phi_s^k(\beta; \hat{f}) + \varepsilon \geqslant \phi_s^k(\beta; \hat{f}^{\langle k,f\rangle}) \quad \forall k \in K, s \in S, f \in F_S^K.$$

The notion here is that each player $k \in K$ gains no more than ε by a unilateral deviation from the strategy $\hat{f}^k$.

Corollary 2.3. *If variables $(\hat{v}, \hat{f})$ are feasible in NLP 2.2, and the NLP 2.2 objective has value $\gamma > 0$, then $\hat{f}$ forms an ε-equilibrium with some ε no greater than $\gamma/(1-\beta)$.*

Proof. Since the NLP 2.2 objective is γ, and since the constraints (i) force the NLP 2.2 objective to be non-negative term-by-term,

$$\hat{v}_s^k - r_s^k(\hat{f}) - \beta \sum_{s' \in S} \hat{v}_{s'}^k q(s' | s, \hat{f}) \leqslant \gamma \quad \forall k \in K, s \in S,$$

which by applying a standard iteration of inequalities argument, implies

$$\hat{v}_s^k - \phi_s^k(\beta; \hat{f}) \leqslant \gamma/(1-\beta) \quad \forall k \in K, s \in S,$$

or, equivalently

$$\hat{v}_s^k \leqslant \phi_s^k(\beta; \hat{f}) + \varepsilon \quad \forall k \in K, s \in S, \quad \text{where } \varepsilon = \gamma/(1-\beta). \tag{2.1}$$

The constraints (i) of NLP 2.2 provide the relations

$$\hat{v}_s^k \geqslant \phi_s^k(\beta; \hat{f}^{\langle k,f\rangle}) \quad \forall k \in K, s \in S, f \in F_S^K, \tag{2.2}$$

and (2.1) combined with (2.2) provide the result. □

3. Limiting average games

The issue of existence of equilibria in limiting average games is of greater order of difficulty than in discounted games. The examples of [4] and [8] demonstrate that stationary limiting average equilibria do not always exist. Sobel [17] attempted the first explicit characterization of limiting average games which possess stationary equilibria, and in Theorem 4 of [17] claims that the existence of such equilibria is equivalent to feasibility of a set of linear and nonlinear constraints. While this theorem holds in the important irreducible case, the following example demonstrates that it fails even in the "unichain" case where transient states occur.

Example 3.1. Consider the following 1-player game (MDP):

$$K=\{1\},\quad S=\{1,2\},\quad A_1^1=\{1,2\},\quad A_2^1=\{1\}.$$

Recall that a stationary strategy $f\in\{f_s^k(a)\,|\,k\in K,\ s\in S,\ a\in A_s^k\}=F_S^K$ has the form $f=((f_1^1(1),f_1^1(2)),(f_2^1(1)))$. Let $\hat f$ and $\bar f\in F_S^K$ be defined by $\hat f=((1,0),(1))$ and $\bar f=((0,1),(1))$. The reward and transition laws are defined as

$$r_1^1(\hat f)=1,\qquad r_1^1(\bar f)=r_2^1(\hat f)=r_2^1(\bar f)=0,$$

and, for all $f\in F_S^K$ and $s\in S$,

$$q(s'\,|\,s,f)=\begin{cases}0 & \text{if } s'=1,\\ 1 & \text{if } s'=2.\end{cases}$$

Of course, this means that state 2 is absorbing. Clearly, $\bar f=((0.5,0.5),(1))$ forms a limiting average equilibrium with equilibrium payoffs $\bar v_1^1=\bar v_2^1=0$, but this equilibrium strategy and payoff do not satisfy conditions 4.6 and 4.7 simultaneously of Theorem 4 in [17].

The following theorem gives a complete characterization of all stationary equilibria for limiting average stochastic games, without any restrictions on the ergodic structure of the process.

Theorem 3.2. *Let $\Gamma=\langle S,K,A,R,Q\rangle$ be a stochastic game, and let $\hat v\in\{v_s^k\,|\,k\in K,\ s\in S\}$, and $\hat f\in F_S^K$ be given. The strategy $\hat f$ forms a limiting average (Nash) equilibrium with equilibrium payoffs $\hat v$ if and only if there exist $\hat t\in\{t_s^k\,|\,k\in K,\ s\in S\}$ and $\hat w\in\{w_s^k\,|\,k\in K,\ s\in S\}$ such that the variables $(\hat v,\hat f,\hat t,\hat w)$ are a global minimum with objective of zero in the following NLP* 3.3.

NLP 3.3. Variables

$$v\in\{v_s^k\,|\,k\in K,\ s\in S\},\quad f\in\{f_s^k(a)\,|\,k\in K,\ s\in S,\ a\in A_s^k\},$$

$$t\in\{t_s^k\,|\,k\in K,\ s\in S\},\quad w\in\{w_s^k\,|\,k\in K,\ s\in S\};$$

$$\text{minimize}\quad \sum_{k\in K}\sum_{s\in S}\left[v_s^k-\sum_{s'\in S}v_{s'}^k q(s'\,|\,s,f)\right]$$

subject to

(i) $v_s^k \geqslant \sum_{s' \in S} v_{s'}^k q(s' | s, f^{\langle k,s,a \rangle}) \quad \forall k \in K, s \in S, a \in A_s^k,$

(ii) $v_s^k + t_s^k \geqslant r_s^k(f^{\langle k,s,a \rangle}) + \sum_{s' \in S} t_{s'}^k q(s' | s, f^{\langle k,s,a \rangle}) \quad \forall k \in K, s \in S, a \in A_s^k,$

(iii) $v_s^k + w_s^k = r_s^k(f) + \sum_{s' \in S} w_{s'}^k q(s' | s, f) \quad \forall s \in S,$

(iv) $\sum_{a \in A_s^k} f_s^k(a) = 1 \quad \forall k \in K, s \in S,$

(v) $f_s^k(a) \geqslant 0 \quad \forall k \in K, s \in S, a \in A_s^k.$

Proof of Theorem 3.2. First notice that by constraint (i), for a global optimum (if it exists) of NLP 3.3, the objective is at least zero. Let $\hat{f}$ be a limiting average stationary equilibrium strategy with equilibrium payoff $\hat{v}$. Since $\hat{f}$ is a stationary strategy, constraints (iv) and (v) of NLP 3.3 are satisfied. At equilibrium, each player $k \in K$ solves a limiting average MDP, so relations (1.7) to (1.10) hold for each player $k \in K$, and suitable $\hat{t}^k$ and $\hat{w}^k$. This implies that $(\hat{v}, \hat{f}, \hat{t}, \hat{w})$ is feasible with objective zero, and hence that we have an optimal solution.

Now, assume $(\hat{v}, \hat{f}, \hat{t}, \hat{w})$ are a global minimum to NLP 3.3 with objective value of zero. The zero objective along with constraint (iii) implies that $\hat{v}_s^k$ is exactly the limiting average payoff to player k when the game starts in state s and stationary strategies $\hat{f}$ are followed. Then the constraints (i) and (ii) of NLP 3.3 provide the conclusion that $\hat{f}$ is a limiting average equilibrium with equilibrium payoffs $\hat{v}$. □

If a global minimum is found in a formulation of NLP 3.3 without a constraint similar to (iii) above, it is possible to derive the relations

$$\hat{v}_s^k \geqslant \phi_s^k(\hat{f}^{\langle k,s,f \rangle}) \quad \forall k \in K, s \in S, f \in F_S^K, \tag{3.1}$$

but $|\hat{v}_s^k - \phi_s^k(\hat{f})|$ may not be zero or even small $\forall k \in K$, $s \in S$. Both (3.1) and equality (closeness) of $\hat{v}$ and $\phi(\hat{f})$ are needed to conclude that a limiting average equilibrium (near-equilibrium) has been found. Also, a feasible point in NLP 3.3 with a small objective does not necessarily imply a near-equilibrium has been found, and a result analogous to Corollary 2.3 cannot be presented here.

4. Limiting average two-person, zero sum games

The characterization developed in Section 3 also includes the important two person, zero sum games (where $K = \{1, 2\}$, and $r^1(f) = -r^2(f)\ \forall f \in F_S^K$), but for these games a simpler, more powerful formulation can be derived. The zero sum reward law implies that $\phi_s^1(f) = -\phi_s^2(f)\ \forall s \in S, f \in F_S^K$, and $\hat{f} = (\hat{f}^1, \hat{f}^2) \in F_S^K$ forms a limiting average equilibrium (called an *optimum or optimal strategies*) with *optimal payoffs* $\hat{v} \in \{v_s^k \mid k \in \{1,2\}, s \in S\}$ if the following condition is satisfied:

$$\phi_s^1(f^1, \hat{f}^2) \leq \phi_s^1(\hat{f}) = \hat{v}_s^1 = -\hat{v}_s^2 = -\phi_s^2(\hat{f}) \leq -\phi_s^2(\hat{f}^1, f^2)$$

for all $s \in S$, $f^1 \in F_S^1$, $f^2 \in F_S^2$, and forms an *ε-optimum* if

$$\phi_s^1(f^1, \hat{f}^2) \leq \phi_s^1(\hat{f}) + \varepsilon \tag{4.1}$$

and

$$-\phi_s^2(\hat{f}) - \varepsilon \leq -\phi_s^2(\hat{f}^1, f^2) \tag{4.2}$$

for all $s \in S$, $f^1 \in F_S^1$, $f^2 \in F_S^2$. Of course, ε-optimality with $\varepsilon = 0$ is equivalent to optimality, since $\phi_s^1(f) = -\phi_s^2(f)$ $\forall f \in F_S^K$.

Given the fact that stationary optimal strategies need not exist, it is useful to develop a measure of the "distance" from optimality for an arbitrary pair of stationary strategies. We propose the following natural measure of such distance:

Let $\hat{f} = (\hat{f}^1, \hat{f}^2) \in F_S^K$ be arbitrary and fixed, and define

$$\delta(\hat{f}) = \sum_{s \in S} \left[\max_{f^1} \phi_s^1(f^1, \hat{f}^2) - \min_{f^2} \phi_s^1(\hat{f}^1, f^2) \right]. \tag{4.3}$$

Note that every term in (4.3) is non-negative, and $\delta(\hat{f}) = 0$ if and only if $\hat{f}$ is optimal. Further, when $0 < \delta(\hat{f})$, then $\hat{f}$ is ε-optimal for an ε no greater than $\delta(\hat{f})$.

The following theorem shows that the nonlinear program NLP 4.2 below characterizes games with stationary optimal and ε-optimal strategies, and finds the "best" stationary strategies with respect to the measure $\delta(f)$ defined above. As such, it represents a generalization and an improvement of Theorem 2.1 in [6]. It should be noted that for $\hat{f} \in F_S^K$ where $K = \{1, 2\}$, the terms $r(\hat{f}^{\langle k,s,a \rangle})$ and $q(s' \mid s, \hat{f}^{\langle k,s,a \rangle})$ are linear in $\hat{f}$.

Theorem 4.1. *Let $\Gamma = \langle S, K, A, R, Q \rangle$ be a two person, zero sum, stochastic game, and let $\hat{f} \in F_S^K$ be given. If there exist $\hat{v} \in \{v_s^k \mid k \in K, s \in S\}$ and $\hat{t} \in \{t_s^k \mid k \in K, s \in S\}$ such that the variables $(\hat{v}, \hat{f}, \hat{t})$ are feasible in the following NLP 4.2 and the NLP objective has value of ε or less, then strategy $\hat{f}$ is ε-optimal. Conversely, if $\hat{f}$ is ε-optimal, then there exist $\hat{v}$ and $\hat{t}$ such that $(\hat{v}, \hat{f}, \hat{t})$ are feasible in NLP 4.2, and the NLP 4.2 objective value is $2|S|\varepsilon$ or less.*

NLP 4.2. Variables

$$v \in \{v_s^k \mid k \in K, s \in S\}, \quad f \in \{f_s^k(a) \mid k \in K, s \in S, a \in A_s^k\},$$

$$t \in \{t_s^k \mid k \in K, s \in S\};$$

$$\text{minimize} \quad \sum_{s \in S} [v_s^1 + v_s^2]$$

subject to

(i) $v_s^1 \geqslant \sum_{s' \in S} v_{s'}^1 q(s' | s, f^{\langle 1,s,a \rangle}) \quad \forall s \in S,\ a \in A_s^1,$

(ii) $v_s^1 + t_s^1 \geqslant r_s^1(f^{\langle 1,s,a \rangle}) + \sum_{s' \in S} t_{s'}^1 q(s' | s, f^{\langle 1,s,a \rangle}) \quad \forall s \in S,\ a \in A_s^1,$

(iii) $v_s^2 \geqslant \sum_{s' \in S} v_{s'}^2 q(s' | s, f^{\langle 2,s,a \rangle}) \quad \forall s \in S,\ a \in A_s^2,$

(iv) $v_s^2 + t_s^2 \geqslant r_s^2(f^{\langle 2,s,a \rangle}) + \sum_{s' \in S} t_{s'}^2 q(s' | s, f^{\langle 2,s,a \rangle}) \quad \forall s \in S,\ a \in A_s^2,$

(v) $\sum_{a \in A_s^k} f_s^k(a) = 1 \quad \forall k = 1, 2,\ s \in S,$

(vi) $f_s^k(a) \geqslant 0 \quad \forall k = 1, 2,\ s \in S,\ a \in A_s^k.$

Proof of Theorem 4.1. Let $(\hat{v}, \hat{f}, \hat{t})$ be feasible in NLP 4.2, and let $\sum_{s \in S} [\hat{v}_s^1 + \hat{v}_s^2] = \varepsilon$. The constraints (v) and (vi) imply that $\hat{f} \in F_S^K$.

Along similar lines as in the proof of Theorem 3.2, it can be shown from the constraints (i) to (iv) that $\hat{f}$ forms an ε-optimal pair.

Now assume $\hat{f} = (\hat{f}^1, \hat{f}^2) \in F_S^K$ is ε-optimal, so $\hat{f}$ satisfies constraints (v) and (vi). By solving an MDP for player 1 with $\hat{f}^2$ fixed, variables $\hat{v}_s^1$ and $\hat{t}_s^1$ for $s \in S$ can be found such that constraints (i) and (ii) of NLP 4.2 are satisfied. Moreover,

$$\hat{v}_s^1 = \max_{f^1 \in F_S^1} \phi_s^1(f^1, \hat{f}^2) \leqslant \phi_s^1(\hat{f}) + \varepsilon, \tag{4.4}$$

since $\hat{f}$ is ε-optimal. By solving an MDP for player 2 with $\hat{f}^1$ fixed, variables $\hat{v}_s^2$, $\hat{t}_s^2$ for $s \in S$ can be found such that constraints (iii) and (iv) of NLP 4.2 are satisfied, and such that

$$\hat{v}_s^2 = \max_{f^2 \in F_S^2} \phi_s^2(\hat{f}^1, f^2) \leqslant \phi_s^2(\hat{f}) + \varepsilon. \tag{4.5}$$

Thus the variables $(\hat{v}, \hat{t}, \hat{f})$ are feasible, and by summing (4.4) and (4.5) over $s \in S$, the NLP 4.2 objective is less than or equal to $2|S|\varepsilon$. □

Theorem 4.1 implies that by solving NLP 4.2, the "best" stationary strategies can be found, as can be seen from the following results.

Corollary 4.3. *If variables $\hat{z} = (\hat{v}, \hat{f}, \hat{t})$ are a global minimum in NLP* 4.2 *with objective value $\varphi(\hat{z})$, then $\varphi(\hat{z}) = \delta(\hat{f}) \leqslant \delta(f)$ $\forall f \in F_S^K$ (see* (4.3)).

Proof. Let $\hat{z} = (\hat{v}, \hat{f}, \hat{t})$ be a global minimum in NLP 4.2 with objective value $\varphi(\hat{z})$. The constraints (i) and (ii) of NLP 4.2 imply

$$\hat{v}_s^1 \geqslant \max_{f^1 \in F_S^1} \phi_s^1(f^1, \hat{f}^2) \quad \forall s \in S, \tag{4.6}$$

while the constraints (iii) and (iv) of NLP 4.2 imply

$$\hat{v}_s^2 \geqslant \max_{f^2 \in F_S^2} \phi_s^2(\hat{f}^1, f^2) = -\min_{f^2 \in F_S^2} \phi_s^1(\hat{f}^1, f^2) \quad \forall s \in S. \tag{4.7}$$

Since $\hat{z}$ is a minimum in NLP 4.2, equality must hold in both (4.6) and (4.7); otherwise, by solving MDPs with $\hat{f}^1$ and $\hat{f}^2$ fixed, variables $\bar{v}$ and $\bar{t}$ could be found such that equality holds in (4.6) and (4.7), and the new variables could be used to contradict the minimality of $\hat{z}$.

Summing (4.6) and (4.7) (with equality holding) over $s \in S$ yields $\varphi(\hat{z}) = \delta(\hat{f})$.

For any $\bar{f} \in F_S^K$, by solving appropriate MDPs, variables $\bar{v}$ and $\bar{t}$ can be found such that $\bar{z} = (\bar{v}, \bar{f}, \bar{t})$ is feasible in NLP 4.2 and

$$\bar{v}_s^1 = \max_{f^1 \in F_S^1} \phi_s^1(f^1, \bar{f}^2) \quad \forall s \in S$$

and

$$\bar{v}_s^2 = = -\min_{f^2 \in F_S^2} \phi_s^2(\bar{f}^1, f^2) \quad \forall s \in S.$$

Thus, $\delta(\bar{f}) = \varphi(\bar{z})$ by construction and $\varphi(\bar{z}) \geqslant \varphi(\hat{z}) = \delta(\hat{f})$ by minimality of $\hat{z}$ and derivation above. □

The next corollary can be shown along similar lines.

Corollary 4.4. *Suppose that the minimum in NLP* 4.2 *does not exist, but that the infimum equals* η (*necessarily non-negative*), *then for every* $\varepsilon > 0$ *there exists* $\hat{f} \in F_S^K$ *that is* $(\eta + \varepsilon)$*-optimal.* □

Remark. Note that by Theorem 4.1 and Corollary 4.4, the existence of ε-optimal stationary strategies for every $\varepsilon > 0$ is equivalent to the infimum of NLP 4.2 being equal to zero.

References

[1] T. Bewley and E. Kohlberg, "On stochastic games with stationary optimal strategies," *Mathematics of Operations Research* 1 (1978) 104-125.

[2] M. Breton, A. Haurie, J.A. Filar and T. Schultz, "On the computation of equilibria in discounted stochastic dynamic games," in: T. Basar, M. Beckmann and W. Krelle, eds., *Lecture notes in Economics and Mathematical Systems No. 265* (Springer, Berlin, 1986) pp. 64-87.

[3] D. Blackwell, "Discrete dynamic programming," *Annals of Mathematical Statistics* 33 (1962) 719-726.

[4] D. Blackwell and T.S. Ferguson, "The big match," *Annals of Mathematical Statistics* 39 (1968) 159-163.

[5] C. Derman, *Finite State Markovian Decision Processes* (Academic Press, New York, 1970).

[6] J.A. Filar and T. Schultz, "Nonlinear programming and stationary strategies in stochastic games," *Mathematical Programming* 35 (1986) 243-247.

[7] A.M. Fink, "Equilibrium in a stochastic n-person game," *Journal of Science of Hiroshima University, Series A-I* 28 (1964) 89-93.

[8] D. Gillete, "Stochastic games with zero stop probabilities," in: M. Dresher, A.W. Tucker and P. Wolfe, eds., *Contributions to the Theory of Games III* (Princeton University Press, Frinceton, NJ, 1957) pp. 179-187.

[9] A. Hordijk and L.C.M. Kallenberg, "Linear programming and Markov decision chains," *Management Science* 25 (1979) 352-362.

[10] A. Hordijk, O.J. Vrieze and G.L. Wanrooy, "Semi-Markov strategies in stochastic games," *International Journal of Game Theory* 12 (1983) 81-89.

[11] C.E. Lemke and J. Howson, "Equilibrium points of bimatrix games," *Journal of the Society of Industrial and Applied Mathematics* 12 (1964) 413-423.

[12] O.L. Mangasarian and H. Stone, "Two-person nonzero-sum games and quadratic programming," *Journal of Mathematical Analysis and Applications* 9 (1964) 348-355.

[13] P.D. Rogers, "Non-zero-sum stochastic games," PhD Thesis, Report ORC 69-8 Operations Research Center, University of California (Berkeley, CA, 1969).

[14] J. Rosenmuller, "On a generalization of the Lemke-Howson algorithm to noncooperative n-person games," *SIAM Journal of Applied Mathematics* 21 (1971) 73-79.

[15] U.G. Rothblum, "Solving stopping stochastic games by maximizing a linear function subject to quadratic constraints," in: O. Moeschlin and D. Pallashke, eds., *Game Theory and Related Topics* (North-Holland, Amsterdam, 1979) pp. 103-105.

[16] T. Schultz, "Mathematical programming and stochastic games," PhD Thesis, Johns Hopkins University (Baltimore, MD, 1987).

[17] M. Sobel, "Noncooperative stochastic games," *The Annals of Mathematical Statistics* 42 (1971) 1930-1935.

[18] M. Takahashi, "Equilibrium points of stochastic, noncooperative n-person games," *Journal of Science of Hiroshima University Series A-I* 28 (1964) 95-99.

[19] O.J. Vrieze, *Stochastic Games with Finite State and Action Spaces, CWI-tract No. 33*, (Centre for Mathematics and Computer Science, Amsterdam, 1987).

[20] R. Wilson, "Computing Equilibria of n-person games," *SIAM Journal of Applied Mathematics* 21 (1971) 80-87.