*Int. Journal of Game Theory, Vol. 9, Issue 1, page 1–12. ©Physica-Verlag, Vienna.*

# Strong Perfect Equilibrium in Supergames[1])

By *A. Rubinstein*, Jerusalem[2])

*Abstract*: The set of payoffs for the strong equilibria is characterized for supergames when the evaluation relations are according to the limit of the means and where no coalition can correlate its strategies in a randomized way.

It is proven that this set is identical to the set of payoffs of the strong perfect equilibria. On the other hand an example is given to demonstrate that perfection is a significant notion in supergames where the evaluation relations are according to the overtaking criterion.

## 1. The Model[3])

The single game $G$ is a game in strategic form

$$G = \langle \{S_i\}_{i=1}^n, \{\pi_i\}_{i=1}^n \rangle.$$

The set of players is $N = \{1, \ldots, n\}$. For each $i \in N$, the set of strategies of $i$ is $S_i$; $S_i$ is assumed non-empty and compact. $S = \prod_{i=1}^n S_i$ is the set of outcomes. An element in $S$ is called an outcome of $G$.

Each player $i$ has a payoff function $\pi_i: S \to \mathbf{R}$ ($\mathbf{R}$ = the reals), which is continuous in the product topology.

Given $\sigma \in S$, the payoff vector of $\sigma$ is the $n$-tuple $\pi(\sigma) = \langle \pi_1(\sigma), \ldots, \pi_n(\sigma) \rangle$.

The supergame $G^\infty$ is $\langle G, \preccurlyeq_1, \ldots, \preccurlyeq_n \rangle$ where $G$ is a single game and the $\preccurlyeq_i$'s are evaluation relations on real number sequences; more exactly, $\preccurlyeq_i$ is a binary relation on $\mathbf{R}^\mathbf{N}$ which is transitive, anti-symmetric, but not necessarily a total order. A strategy for $i$ in $G^\infty$ is a set $\{f_i(t)\}_{t=1}^\infty$, where $f_i(1) \in S_i$, and for $t \geqslant 2, f_i(t): S^{\{1,\ldots,t-1\}} \to S_i$. Thus a supergame strategy is a choice of strategies at every stage, where each choice is possibly dependent on the outcomes of the preceding games, and where all players know all the choices made by all the players in the past.

The set of supergame strategies of player $i$ is denoted by $F_i$. $F$ is the set of $n$-tuples of strategies; $F = \prod_{i=1}^n F_i$.

[1]) I wish to thank Professor B. Peleg for his advice and guidance.

[2]) Dr. *Ariel Rubinstein*, Department of Economics, The Hebrew University, Jerusalem, Israel.

[3]) See *Roth* [1975].

Given $f \in F$, the outcome at time $t$ is denoted by $\sigma(f)(t)$, and is defined inductively by

$$\sigma(f)(1) = (f_1(1), \ldots, f_n(1))$$
$$\sigma(f)(t) = (\ldots, f_i(t)(\sigma(f)(1), \ldots, \sigma(f)(t-1)), \ldots).$$

Define a relation $\bar{\preccurlyeq}_i$ on $F$, induced by $\preccurlyeq_i$, as follows:

For all $f, g \in F$, $f \bar{\preccurlyeq}_i g$ if

$$\{\pi_i(\sigma(f)(t))\}_{t=1}^{\infty} \preccurlyeq_i \{\pi_i(\sigma(g)(t))\}_{t=1}^{\infty}.$$

Before passing to the main definitions, some generally usefull notation is introduced:

1. Let $A$ be a set, $a \in A^n$. The $(n-1)$-tuple $(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ is denoted by $a^{-i}$, and the pair $(a^{-i}, a_i)$ is identified with $a$.
2. If $B \subseteq N$, $\{a_i\}_{i \in B}$ is denoted by $a^B$, and $(a^B, a^{N-B})$ is identified with $a$.
3. Given sets $\{A_i\}_{i \in B}$, $\prod_{i \in B} A_i$ is denoted by $A^B$.
4. Given $f \in F$ and $r(1), \ldots, r(T) \in S$, the $n$-tuple of strategies determined by $f$ after a "history" $r(1), \ldots, r(T)$ is denoted by $f_{|r(1),\ldots,r(T)}$; thus $(f_{|r(1),\ldots,r(T)})_i(t)(s(1), \ldots, s(t-1)) =$ $f_i(T+t)(r(1), \ldots, r(T), s(1), \ldots, s(t-1))$.
5. Given $x, y \in \mathbf{R}^\mathbf{B}$, $x << y$ denotes that $x_i < y_i$ for all $i \in B$.

*Definition.* $f \in F$ is a (Nash) equilibrium in the supergame $G^\infty$ if for all $i$, and for all $h_i \in F_i$, $f \not\bar{\preccurlyeq}_i (f^{-i}, h_i)$.

This definition of equilibria in the supergame is known to be too general [see for example *Aumann*, 1976; and *Rubinstein*, 1977]. One reasonable restriction is by the notion of perfect equilibrium.

*Definition.* $f \in F$ is a perfect equilibrium in the supergame $G^\infty$ if for all $r(1), \ldots, r(T) \in S$ $(0 \leqslant T)$, $f_{|r(1),\ldots,r(T)}$ is an equilibrium in the supergame.

This paper investigates the strong equilibria of supergames (s.g.) [see *Aumann*, 1959, 1960], that is those $n$-tuples of strategies in s.g. where no coalition of players can alter their strategies to bring profit to all members of the coalition. Formally:

*Definition.* $f \in F$ is a strong equilibrium in $G^\infty$ if there does not exist $\emptyset \ddagger B \subseteq N$, and $g^B \in F^B$ such that for all $i \in B$, $(f^{N-B}, g^B) \bar{\succcurlyeq}_i f$.

The notion of perfection is applied to strong equilibrium as follows:

*Definition.* $f \in F$ is a strong perfect equilibrium in $G^\infty$ if for all $r(1), \ldots, r(T) \in S$ $f_{|r(1),\ldots,r(T)}$ is a strong equilibrium in $G^\infty$.

The main evaluation relation considered is the Limit of the Means Evaluation Relation (L.M.E.R.), defined by

$$x \lessdot y \text{ if } 0 < \underline{\lim} \frac{\sum_{t=1}^{n} y_t - x_t}{n} \qquad (x, y \in \mathbf{R}^{\mathbf{N}}).$$

The other evaluation relation considered is the Overtaking Criterion Evaluation Relation (O.T.E.R.) defined by

$$x \lessdot y \text{ if } 0 < \underline{\lim} \sum_{t=1}^{n} y_t - x_t \qquad (x, y \in \mathbf{R}^{\mathbf{N}}).$$

(For an axiomatic characterization of the O.T.E.R. see *Brock* [1977]).

An $n$-tuple of strategies $f$ is summable if $\lim \frac{\sum_{t=1}^{T} \pi_i(\sigma(f)(t))}{T}$ exists for all $i$. The vector of the limits is called the payoff of $f$ in $G^\infty$, and is denoted by $\pi(f)$.

$f \in F$ is stationary if there exists $\sigma \in S$ such that for all $t$, $\sigma(f)(t) = \sigma$. Denote this $\sigma$ by $\hat{\sigma}(f)$.

It is proved in *Aumann/Shapley* [1976] and *Rubinstein* [1977] that if the $\lessdot_i$'s are L.M.E.R.'s, then the payoff set of the summable/stationary perfect equilibria is equal to the payoff set of the summable/stationary equilibria. If the $\lessdot_i$'s are O.T.E.R.'s, the difference between the two sets is "marginal" [proved in *Rubinstein,* 1979].

Together these results indicate that when considering Nash equilibrium in supergames, the concept of perfection does not enable the isolation of a smaller solution set.

In section 2 the set of payoffs of the summable strong equilibria is characterized in a supergame where all the $\lessdot_i$'s are L.M.E.R.'s. It is proved that this set remains unchanged when the requirement of perfection is added. But when considering strong equilibria in supergames where the $\lessdot_i$'s are O.T.E.R.'s, perfection becomes significant. The matrix game described in section 4, the "Amnesty Dilemma" has a stationary equilibrium but no strong perfect equilibrium. Furthermore, this property is insensitive to "small" changes in the payoffs.

## 2. Strong and Strong Perfect Equilibria in Supergames with L.M.E.R.'s

Let $A$ be a set. Define $C(A) =$
$\{c \mid c : A \to [0,1], c(a) > 0$ for a finite number of $a \in A$, and $\sum_{a \in A} c(a) = 1\}$.

Thus $C(A)$ is the set of the densities on $A$ which take positive values on a finite set only. For every $c^B \in C(S^B)$ and every $\gamma^{N\text{-}B} \in S^{N\text{-}B}$, write

$$\pi_i(c^B, \gamma^{N\text{-}B}) \text{ for } \sum_{s^B \in S^B} c^B(s^B) \cdot \pi_i(s^B, \gamma^{N\text{-}B}).$$

*Definition.* $a \in \mathbf{R}^n$ is a pareto-optimal payoff (in $G$) if $a \in \text{conv } \pi(S)$ and for no $b \in \text{Conv } \pi(S)$ is $a \ll b$.

*Definition.* $\acute{a} \in \mathbf{R}^n$ is a desired payoff if it is pareto-optimal, and if for all $\emptyset \neq B \subset N$, there exists $\gamma^{N\text{-}B} \in S^{N\text{-}B}$ such that for no $c^B \in C(S^B)$ is $\pi_i(c^B, \gamma^{N\text{-}B}) > a_i$ for all $i \in B$. A desired payoff has the property that for all coalitions $\emptyset \neq B \subset N$, there is a corresponding "punishment" $\gamma^{N\text{-}B}$, inflicted by $N - B$, such that even if the players in $B$ could randomly coordinate their strategies in a single game, they could not guarantee more than the desired payoff offers for them all.

The following lemma, taken from *Aumann* [1960, Lemma 5.2] will be used in the main propositions of this section.

*Lemma 2.1.* Let $Z$ be a finite set, and let $y \in C(Z)$. For all maps $\psi : \mathbf{N} \to Z$, for all natural numbers $k$, and for all $z \in Z$, define $\rho_\psi(k, z)$ by

$$\rho_\psi(k, z) = |\{j \mid \psi(j) = z, j \leqslant k\}|$$

($\rho_\psi(k, z)$ is the number of times $\psi$ takes the value $z$ up to time $k$).

Then there exists $\psi : \mathbf{N} \to Z$ such that for all $z \in Z$, $\lim\limits_{k \to \infty} \dfrac{\rho_\psi(k, z)}{k} = y(z)$.

The following proposition gives a necessary condition for summable $f \in F$ to be a strong equilibrium in a supergame with L.M.E.R.'s.

*Proposition 2.2.* Let $G^\infty = \langle G, \leqslant_1, \ldots, \leqslant_n \rangle$ be a supergame where all the $\leqslant_i$'s contain[4] the L.M.E.R. If the summable strategy $f \in F$ is a strong equilibrium in $G^\infty$, then $\pi(f) = a$ is a desired payoff.

*Proof*: conv $\pi(S)$ is closed, hence $a \in$ conv $\pi(S)$. Suppose $a$ is not desired. Then there is a $B \neq \emptyset$ such that for all $s^{N\text{-}B} \in S^{N\text{-}B}$ there exists $c^B \in C(S^B)$ where $\pi_i(c^B, s^{N\text{-}B}) > a_i$ for all $i \in B$. Now, $\sup\limits_{c^B \in C(S^B)} \min\limits_{i \in B} \{\pi_i(c^B, s^{N\text{-}B}) - a_i\}$ is a l.s.c. function of $s^{N\text{-}B}$.

Therefore there is an $\epsilon > 0$

$$0 < \epsilon < \min_{s^{N\text{-}B} \in S^{N\text{-}B}} \ \sup_{c^B \in C(S^B)} \ \min_{i \in B} \{\pi_i(c^B, s^{N\text{-}B}) - a^i\}.$$

Since $\pi_i$, $i = 1, \ldots, n$, are uniformly continuous, there is a finite open cover $U_1, \ldots, U_k$ of $S$ such that for all $s, t \in U_j$, and for all $i$, $|\pi_i(s) - \pi_i(t)| \leqslant \epsilon/2$. Since $S^{N\text{-}B}$ and $S^B$ are compact, there is an open cover of $S^{N\text{-}B}$ $\{0_1, \ldots, 0_L\}$. such that for all $0_j$ and for all $s^B \in S^B$ there exists $m$ such that $U_m \supseteq 0_j \times \{s^B\}$. Take $r^{N\text{-}B}(j) \in 0_j$, and $c^B(j)$ satisfying

$$\pi_i(c^B(j), r^{N\text{-}B}(j)) - a_i > \epsilon \quad \text{for all } i.$$

Let $\psi_j : \mathbf{N} \to S^B$ be maps satisfying, for all $s^B \in S^B$,

[4] Let $R$, $S$ be a binary relation on a set $A$. $S$ contains $R$ if for all $a, b \in A$ $a \, R \, b \Rightarrow a \, S \, b$.

$$\lim_{k\to\infty} \frac{\rho_{\psi_j}(k, s^B)}{k} = c^B(j)\,(s^B).$$

Their existence is guaranteed by lemma 2.1.

Define $g_i \in F_i$ for all $i \in B$ inductively, together with $m_j(t)$, the number of times $B$ use $c^B(j)$ up to time $t$, as follows:

Let $j_1$ satisfy $f^{N\text{-}B}(1) \in 0_{j_1}$.

Then for all $i \in B$, $1 \leqslant j \leqslant L$

$$g_i(1) = (\psi_{j_1}(1))_i$$

$$m_j(1) = \begin{cases} 1 & j = j_1 \\ 0 & \text{otherwise.} \end{cases}$$

Proceeding by induction, let $j_{t+1}$ satisfy

$$f^{N\text{-}B}(t+1)\,(\{\sigma(f^{N\text{-}B}, g^B)\,(h)\}_{h=1}^{t}) \in 0_{j_{t+1}}.$$

Then $g_i\,(t+1)\,(\{\sigma(f^{N\text{-}B}, g^B)\,(h)\}_{h=1}^{t}) = [\psi_{j_{t+1}}(m_{j_{t+1}}(t)+1)]_i$ and $g_i(t+1)$ is defined arbitrarily at other "histories" in its domain.

$$m_j(t+1) = \begin{cases} m_j(t) + 1 & j = j_{t+1} \\ m_j(t) & \text{otherwise.} \end{cases}$$

Write

$T_j(T)$ for $|\{t \mid j_t = j, 1 \leqslant t \leqslant T\}|$; then $\underline{\lim}\ \dfrac{\sum_{t=1}^{T} \pi_i(\sigma(f^{N\text{-}B}, g^B)\,(t))}{T} \geqslant$

$$\geqslant \underline{\lim}\ \frac{\sum_{t=1}^{T} \pi_i(\sigma^B(f^{N\text{-}B}, g^B)\,(t), r^{N\text{-}B}(j_t))}{T} - \frac{\epsilon}{2} =$$

$$= \underline{\lim}\ \frac{\sum_{j=1}^{L} \sum_{s^B \in S^B} \rho_{\psi_j}(T_j(T), s^B) \cdot \pi_i(s^B, r^{N\text{-}B}(j_t))}{T} - \frac{\epsilon}{2} =$$

$$= \underline{\lim}\ \frac{\sum_{j=1}^{L} T_j(T) \sum_{s^B \in S^B} ((\rho_{\psi_j}(T_j(T), s^B))/(T_j(T))) \cdot \pi_i(s^B, r^{N\text{-}B}(j_t))}{T} - \frac{\epsilon}{2}.$$

For $j$ satisfying $T_j\,(T)\,\overrightarrow{T}^{\infty}$,

$$\lim \sum_{s^B \in S^B} \frac{\rho_{\psi_j}(T_j(T), s^B)}{T_j(T)} \cdot \pi_i(s^B, r^{N\text{-}B}(j)) = \pi_i(c^B(j), r^{N\text{-}B}(j)) > a_i + \epsilon .$$

Thus

$$\underline{\lim} \frac{\sum_{t=1}^{T} \pi_i(\sigma(f^{N\text{-}B}, g^B)(t))}{T} \geqslant a_i + \frac{\epsilon}{2} .$$

Therefore for all $i \in B$, $(g^B, f^{N\text{-}B}) \overline{\geqslant}_i f$.

The following proposition will give a sufficient condition for a payoff in $\pi(S)$ to be the payoff of a *strong perfect* equilibrium of a supergame with L.M.E.R.'s.

*Proposition 2.3.* Let $G^\infty = \langle G, \leqslant_1, \ldots, \leqslant_n \rangle$ be a supergame where for all $1 \leqslant i \leqslant n$, $\leqslant_i$ is an L.M.E.R.

If $a \in \mathbf{R}^n$ is a desired payoff, then there exists $f \in F$, a strong summable perfect equilibrium such that $\pi(f) = a$.

*Proof.* For every $\emptyset \ddagger B \subset N$ let $\gamma^{N\text{-}B}$ be some strategy in $S^{N\text{-}B}$ which guarantees that for all $c^B \in C(S^B)$ there is $i \in B$ such that $\pi_i(c^B, \gamma^{N\text{-}B}) \leqslant a_i$.

The existence of such a $\gamma^{N\text{-}B}$ follows from the definition of desired payoff.

Now, $a \in \operatorname{conv} \pi(S)$, thus there exists $c \in C(S)$ such that $\pi(c) = a$. Let $\psi$ be a map from the natural numbers into $S$, satisfying

$$\lim_{k \to \infty} \frac{\rho_\psi(k, s)}{k} = c(s).$$

By Lemma 2.1 such a map exists.

For every $s(1), \ldots, s(T) \in S$, define $B(s(1), \ldots, s(T))$ the set of profit-making deviants after $s(1), \ldots, s(T)$, in parallel with the definition of $f \in F$; all this by induction on $T$.

$$B(\emptyset) = \emptyset$$

$$f_i(1) = \psi(1)_i .$$

$$B(s(1), \ldots, s(T)) = \left\{ \begin{array}{ll} A & \text{if } A = B(s(1), \ldots, s(T-1)) \cup \\ & \cup \{ i \mid s_i(T) \ddagger f_i(T)(s(1), \ldots, s(T-1)) \ddagger \emptyset \} \\ & \text{and if for all } i \in A \\ & \sum_{t=1}^{T} \frac{\pi_i(s(t))}{T} \geqslant a_i + \frac{1}{\sqrt{T}} \\ \emptyset & \text{otherwise} \end{array} \right.$$

$$f_i\,(T+1)\,(s(1),\dots,s(T)) = \begin{cases} \gamma_i^{N\text{-}B(s(1),\dots,s(T))} \text{ if } B(s(1),\dots,s(T)) \neq \emptyset \\ \qquad \text{and } i \notin B(s(1),\dots,s(T)) \\ \text{arbitrary if } B(s(1),\dots,s(T)) \neq \emptyset \text{ and} \\ \qquad i \in B(s(1),\dots,s(T)) \\ \psi_i\,(T-k) \text{ if } B(s(1),\dots,s(T)) = \emptyset \text{ and} \\ \qquad k = \max\,\{t \mid B(s(1),\dots,s(t)) \neq \emptyset \\ \qquad\qquad \text{or } t = 0\}. \end{cases}$$

Let $r(1),\dots,r(T) \in S (T \geqslant 0)$. Write $\bar{f}$ for $f_{|r(1),\dots,r(T)}$. To prove that $f$ is a strong perfect equilibrium it suffices to prove that $\bar{f}$ is a strong perfect equilibrium.

*Lemma*. Let $h^B \in F^B$; define $D(t)$ by

$$D(t) = B(r(1),\dots,r(T),\sigma(h^B,\bar{f}^{N\text{-}B})\,(1),\dots,\sigma(h^B,\bar{f}^{N\text{-}B})\,(t)).$$

Then, for all $t_0$ there exists $t_1 \geqslant t_0$ such that $D(t_1) = \emptyset$.

*Proof*. Suppose not. From the definition of the set of profit-making deviants, for all $t_0 \leqslant t \leqslant s$

$$\emptyset \neq D(t_0) \subseteq D(t) \subseteq D(s) \subseteq N.$$

Thus there exists $t_0 \leqslant t'$ such that for all $t' \leqslant t$, $\emptyset \neq D(t) = D(t')$. Hence for all $t' \leqslant t$,

$$\sigma^{N\text{-}D(t')}\,(h^B,\bar{f}^{N\text{-}B})\,(t) = \gamma^{N\text{-}D(t')}.$$

Now for all $t' \leqslant t_1$,

$$\frac{1}{T+t_1}\,[\sum_{t=1}^{T} \pi_i(r(t)) + \sum_{i=1}^{t_1} \pi_i(\sigma(\bar{f}^{N\text{-}B},h^B)\,(t))] =$$

$$\frac{1}{T+t_1}\,[\sum_{t=1}^{T} \pi_i(r(t)) + \sum_{t=1}^{t'} \pi_i(\sigma(\bar{f}^{N\text{-}B},h^B)\,(t)) + \sum_{t=t'+1}^{t_1} \pi_i(\sigma^B(\bar{f}^{N\text{-}B},h^B)\,(t),\gamma^{N\text{-}B})]$$

and for some $i \in B$ (dependent on $t_1$) $\leqslant \frac{1}{T+t_1}\,[C_i + (t_1 - t')\,a_i]$ ($C_i$ is the sum of the first two terms). The last term is less than $a_i + \frac{1}{\sqrt{T+t_1}}$ for sufficiently large $t_1$, contradicting $D(t_1) \neq \emptyset$. From the lemma, applied to the case $B = \emptyset$, and from the definition of $f$ it follows that for all $i$,

$$\frac{1}{p}\sum_{t=1}^{p} \pi_i(\sigma(\bar{f})\,(t)) \to a_i.$$

Let $h^B \in F^B$. From the lemma, there exists $T_1$ such that for all $T_1 \leqslant t$, $D(t) \subseteq B$. Two separate cases arise:

1. There exists $T_1 \leqslant \bar{T}$ such that for all $\bar{T} \leqslant t$, $D(t) = \emptyset$. Then there exists $K$ such that for all $K \leqslant t$,

$$\frac{1}{p} \sum_{t=1}^{p} \pi_i(\sigma(\bar{f})(t)) = \sigma(h^B, \bar{f}^{N\text{-}B})(t) = \psi(t-K).$$

Hence

$$\frac{1}{p} \sum_{t=1}^{K} \pi_i(\sigma(\bar{f})(t)) + \frac{p-K}{p} \cdot \frac{1}{p-K} \sum_{t=K+1}^{p} \pi_i(\sigma(\bar{f})(t)) \rightarrow a_i .$$

Thus for all $i \in B$, $\bar{f} \not\preceq_i (\bar{f}^{N\text{-}B}, h^B)$.

2. There are an infinite number of $t$ such that $D(t) \ddagger \emptyset$.
Let $t_0 \geqslant T$ be such that $D(t_0) \ddagger \emptyset$ $(D(t_0) \subseteq B)$. From the lemma it follows that there exist $\hat{t} \geqslant t_0$ and $i \in B$ such that

$$\frac{1}{T+\hat{t}} \left[ \sum_{t=1}^{T} \pi_i(r(t)) + \sum_{t=1}^{\hat{t}} \pi_i(\sigma(\bar{f}^{N\text{-}B}, h^B)(T)) \right] < a_i + \frac{1}{\sqrt{T+\hat{t}}} .$$

For sufficiently large $t_0$, $\hat{t} \geqslant t^0$ implies that

$$\frac{1}{\hat{t}} \sum_{t=1}^{\hat{t}} \pi_i(\sigma(\bar{f}^{N\text{-}B}, h^B)(t) < a_i + t^{-1/4} .$$

Hence there exists $i \in B$ such that at an infinite number of times $\hat{t}$ his mean payoff from $(\bar{f}^{N\text{-}B}, h^B)$ is less than $a_i + t^{-1/4}$, and thus $\bar{f} \not\preceq_i (\bar{f}^{N\text{-}B}, h^B)$.
The following theorem is a consequence of 2.2. and 2.3.

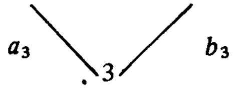
*Theorem*. In a supergame with L.M.E.R.'s, the set of payoffs of the summable strong equilibria = the set of payoffs of the summable strong perfect equilibria = the set of desirable payoffs.

## Example and Comment

In matrix games, the desired payoffs are contained in the $\beta$-core, but the converse is not necessarily true. Let us look at the following 3-person game, where each player has two pure strategies, $a_i$ and $b_i$. The payoff matrix is represented by two matrices, where player 1 is the row player, player 2 is the column player, and player 3 chooses the matrix.

|  | $a_2$ | $b_2$ |
|---|---|---|
| $a_1$ | (2, 2, 2) | (1, 1, 0) |
| $b_1$ | (1, 1, 3) | (0, 0, 4) |

|  | $a_2$ | $b_2$ |
|---|---|---|
| $a_1$ | (0, 0, 4) | (1, 1, 3) |
| $b_1$ | (1, 1, 0) | (2, 2, 2) |



For every $1 \leqslant i \leqslant 3$, $S_i$ is the set of mixed strategies naturally identified with the interval $[0, 1]$, where the choice $p \in [0, 1]$ corresponds to the strategy $p \cdot a_i + (1-p) \cdot$ $\cdot\, b_i$ [5]). The payoff functions are the expected utilities. (2, 2, 2) is a payoff in the $\beta$-core, since the only coalition with a possible profitable strategy deviating from $(a_1, a_2, a_3)$ is $\{3\}$; but $\{1, 2\}$ have a punishing strategy $(1/2) \cdot (a_1, b_2) + (1/2) \cdot (a_2, b_1)$ which reduces 3's expected payoff, whatever his strategy, to $1.5 < 2$.

However, (2, 2, 2) is not a desired payoff since for $p \in S_1$, $q \in S_2$, player 3 may assure himself

$$\max\,\{2pq + 3(1-p)\,q + 4(1-p)\,(1-q), 4pq + 3p(1-q) + 2\,(1-p)\,(1-q)\}$$

which is strictly greater than 2.

## 3. Strong Equilibrium with the O.T.E.R.

The O.T.E.R. contains the L.M.E.R., and thus from 2.2 the following proposition is obtained.

*Proposition 3.1*: Let $f \in F$ be a strong summable equilibrium in a supergame with O.T.E.R.'s. Then $\pi(f)$ is a desired payoff.

With the O.T.E.R., the mean is not a good characterization of payoff sequences. (Thus, for example, for all *a, b, c*, where $a < b$, the sequence $(b, c, c, \ldots)$ is preferred to $(a, c, c, \ldots)$, despite the fact that the means are the same.) Hence it is pertinent to examine the strong stationary equilibria.

In continuation of *Rubinstein* [1979], where the strong Nash equilibria in a supergame with O.T.E.R.'s were characterized, it might be though that if $s \in S$ has the property that for all $\emptyset \neq B \subseteq N$ either

1. No $c^B \in C(S^B)$ exists such that for all $i \in B$

$$\pi_i(c^B, s^{N-B}) > \pi_i(s).$$

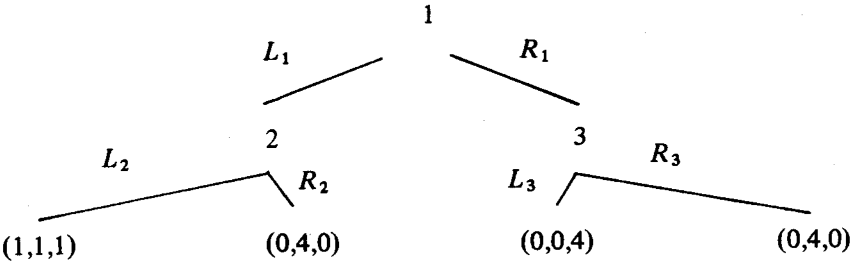
or

2. There exists $\gamma^{N-B}$ such that for all $c^B \in C(S^B)$ a player $i \in B$ can be found for whom $\pi_i(c^B, \gamma^{N-B}) < \pi_i(s)$.

Then, there must exist a strong stationary equilibrium $f$ such that $\hat{\sigma}(f) = \sigma$.

[5]) $p \cdot a_i + (1-p) \cdot b_i$ denotes the strategy "play $a_i$ with probability $p$, and $b_i$ with probability $(1-p)$."

But consider the following 3-person game presented in extended form. For every player, $S_i = \{R_i, L_i\}$. The evaluation relations are according to the overtaking criterion.



(1,1,1) is not a stationary payoff of a strong equilibrium in the supergame, despite the fact that for every $S \subseteq \{1,2,3\}$, either no deviation profitable to every player in the supergame exists, or $\{1,2,3\} - S$ can "retaliate", punishing at least one of the players in $S$.

This example, demonstrates the possibility of deviation by stages. An equilibrium strategy must punish player 2 for his deviation $R_2$, as follows: every time 2 gains 4, 2 and 3 will play $(R_1, L_3)$ at least three times; however, already after the first punishing game, both 2 and 3 will have averaged more than 1. Thus the coalition {2,3} can plan the following trick: 2 deviates; after being punished once, 3 plays alternatively $L_3$, $R_3$, while 2 plays $L_2$. This strategy is preferable for 2 and 3 (according to the overtaking criterion) to a strategy with a constant flow of 1.

On the other hand, the following proposition may be proven in a similar way to 2.2.

*Proposition 3.2.* Let $a$ be a pareto optimal payoff where for all $N \supset B \supset \emptyset$ there is a $\gamma^{N\text{-}B} \in S^{N\text{-}B}$, such that for all $c^B \in C(S^B)$, a player $i \in B$ can be found for whom $\pi_i(c^B, \gamma^{N\text{-}B}) < a_i$. If $\preccurlyeq_i$'s are O.T.E.R.'s then there is a strong stationary equilibrium such that $\hat{\sigma}(f) = s$.

## 4. The Amnesty Dilemma

Consider the following two-person game: $S_i = \{a_i, b_i\}$ and $\pi_i$ are represented by the matrix

| | $a_2$ | $b_2$ |
|---|---|---|
| $a_1$ | (2,2) | (1,1) |
| $b_1$ | (4,0) | (1,1) |

The following interpretation of the game explains its name. The column player is society, the row player is a citizen. When free, the citizen can behave well ($a_1$) or commit a crime ($b_1$). Society can jail him ($b_2$) or let him free ($a_2$). Commission of a crime benefits the individual but damages society; punishing the criminal is damaging to both.

From 3.3 it follows that (2,2) is the payoff of a strong (even stationary) equilibrium in the supergame with O.T.E.R.'s. But:

*Proposition 4.1*. This supergame has no perfect equilibrium.

The following will first be proved.

*Proposition 4.2*. Let $f$ be a strong perfect equilibrium in a supergame with O.T.E.R.'s. Let $r(1), \dots, r(T) \in S$. Then there exists no $\tau \in S$ such that for all $i \in N$

$$\pi_i(f(T+1)(r(1), \dots, r(T))) < \pi_i(\tau).$$

*Proof*. Suppose the proposition is false. Writing $\sigma$ for $f(T+1)(r(1), \dots, r(T))$, let $\tau \in S$ be a such that for all $i \in N$, $\pi_i(\sigma) < \pi_i(\tau)$.

Write $\bar{f} = f_{|r(1),\dots,r(T)}$.

Define the following $n$-tuple of strategies: $g_i(1) = \tau_i$.

For all $t \geqslant 2, g_i(t)(s(1), \dots, s(t-1)) = \bar{f}_i(t)(\sigma, s(2), \dots, s(t-1))$. Clearly $\pi(\sigma(g)(1)) = \pi(\tau) \gg \pi(\sigma) = \pi(\sigma(\bar{f})(t))$ and for all $t \geqslant 2$, $\pi(\sigma(g)(t)) = \pi(\sigma(\bar{f})(t))$.

Thus for all $i$, $\bar{f} \leqslant_i g$, contradicting the perfection of $f$.

*Proof of 4.1*. Suppose that $f \in F$ is a strong perfect equilibrium. From 4.2 there exist no $r(1), \dots, r(T)$ such that $f_2(T+1)(r(1), \dots, r(T)) = b_2$. It follows that $f_2(t) \equiv a_2$ for all $t$.

Clearly $f_1(t) = b_1$ for all $t$, otherwise perfection is contradicted.

Now the deviation of player 2, given by $f_2(t) = b_2$ guarantees him a utility flow of 1, contradicting $f$ being an equilibrium.

Intuitively, the situation is as follows: society and the citizen can agree to play $(a_1, a_2)$, with society threatening to punish the citizen by playing $b_2$ should he play $b_1$.

But after the citizen indeed deviates and plays $b_1$, society agrees to "forgive" him, since it is in its own interest to do so. The original threat is not a deterrent.

This example, in my opinion, exposes the limitations of the concepts of solution given in this paper. There is lacking a notion of "precedent"; if we are not to ignore the player's expectations, we have, for example, to introduce somehow, into player 2's considerations the possible consequence of not punishing 1. I hope to treat this subject in a forthcoming paper.

*Example. The Prisoner's Dilemma*

In contrast to the Amnesty Dilemma, the Prisoner's Dilemma has a strong perfect equilibrium.

This game is described like the previous one, but with the following matrix:

|  | $a_2$ | $b_2$ |
|---|---|---|
| $a_1$ | (2,2) | (0,3) |
| $b_1$ | (3,0) | (1,1) |

Define the $i$-strategy at time $t$, $f_i(t)$, and the "length of time $i$ has to be punished because of his record up to time $t-1$", $m_i(t)$, by induction on $t$:

$m_i(1) = 0$
$f_i(1) = a_i$

$$m_i(t)(s(1), \ldots, s(t-1)) = \begin{cases} 1 \text{ if } i \text{ is the only player for whom } s_i(t-1) \neq a_i \text{ and} \\ \quad m_j(t-1)(s(1), \ldots, s(t-2)) = 0 \quad \text{for all } j. \\ m_i(t-1)(s(1), \ldots, s(t-2)) + 1 \text{ if } s_i(t) \neq a_i \text{ and} \\ \qquad\qquad m_i(t-1)(s(1), \ldots, s(t-2)) > \\ \max\{0, m_i(t-1)(s(1), \ldots, s(t-2)) - 1\} \text{ otherwi} \end{cases}$$

$$f_i(t)(s(1), \ldots, s(t-1)) = \begin{cases} b_i \text{ if for the other player } j,\ m_j(t)(s(1), \ldots, s(t-1) \\ \\ a_i \text{ otherwise.} \end{cases}$$

The players are planning to play $(a_1, a_2)$ unless one of them deviates. If 2 deviates, 1 punishes him, by forcing $(b_1, a_2)$. If 2 does not co-operate in his own punishment, player 1 will increase the period of punishment. Clearly, in contrast to the previous example, the punishing player profits from the punishing arrangement and he has no motivation to forgive the deviant. It is easily verified that $f$ is indeed a strong perfect equilibrium.

## References

*Aumann, R.J.*: Acceptable Points in General Cooperative $n$-person Game. Contributions to the Theory of Games, IV. Ed. by A.W. Tucker and R.C. Luce. Annals of Math. Studies, No. **40**. Princeton, N.J., 1959, 287–324.

–: Acceptable Points in Games of Perfect Information. Pac. J. Math. **10**, 1960, 381–417.

–: Lectures On Game Theory. Stanford University, 1976, unpublished manuscript.

*Aumann, R.J.*, and *L. Shapley*: Long Term Competition – A Game Theoretic Analysis, 1976, unpublished manuscript.

*Brock, W.A.*: An Axiomatic Basis for the Ramsey Weizsacker Overtaking Criterion. Econometrica **38**, 1970, 927–929.

*Roth, A.E.*: Self Supporting Equilibrium in the Supergame, 1975, unpublished manuscript.

*Rubinstein, A.*: Equilibrium in Supergames. Center for Research in Math. Economics and Game Theory, The Hebrew University, Jerusalem, R.M. No. 25, May 1977.

–: Equilibrium in Supergames with the Overtaking Criterium. Journal of Economic Theory **21**, 1979, 1–9.

Received December, 1978.