

ECONOMIC RESEARCH REPORTS

RENEGOTIATION IN FINITELY REPEATED GAMES

BY

**Jean-Pierre Benoit
and
Vijay Krishna**

RR # 91-34

June, 1991

**C. V. STARR CENTER
FOR APPLIED ECONOMICS**



**NEW YORK UNIVERSITY
FACULTY OF ARTS AND SCIENCE
DEPARTMENT OF ECONOMICS
WASHINGTON SQUARE
NEW YORK, N.Y. 10003**

RENEGOTIATION IN FINITELY REPEATED GAMES

BY

Jean-Pierre Benoit

and

Vijay Krishna

**Jean Pierre Benoit
Department of Economics
New York University
New York, NY 10003**

**Vijay Krishna
Harvard Business School
Soldiers Field
Boston, MA 02163**

**July 1988
Revised November 1990**

We are grateful to Debraj Ray, Jorgen Weibull and two referees for helpful comments and to Scott Cantor for able research assistance. Research support was provided by the National Science Foundation (Grant No. SES 8611211) and the Division of Research at the Harvard Business School. Technical support from the C.V. Starr Center for Applied Economics is gratefully acknowledged.

Abstract

We study perfect equilibria of finitely repeated games which are immune to the possibility of "renegotiation" among players. The limiting properties of the set of payoffs from such equilibria are derived and the results contrasted with the "perfect folk theorem" for finitely repeated games.

1. Introduction

Interest in the study of repeated games stems largely from the intuition that equilibrium play in repeated settings may allow players to cooperate without the need for contracts enforced by outside parties. Over the years this intuition has been sharpened in many ways and results formalizing this idea are commonly called "folk theorems." Broadly, these results say that provided players value the future enough, any feasible and individually rational payoff in the one-shot game can be achieved as the average payoff in an equilibrium of the corresponding repeated game. Typically, formal proofs of these results are constructive and run something along the following lines. A "proposed" outcome path that achieves the required payoff is constructed. It is implemented by designing a set of strategies that deter any player from deviating by credibly threatening to "punish" any deviations. Of particular interest are the "cooperative" outcomes, those that yield payoffs on the Pareto frontier of the feasible set.

It was recognized very early that in these matters, there is a dichotomy between finite and infinite horizon games. There are games for which equilibrium payoffs of their finitely repeated versions differ in the limit from equilibrium payoffs of their infinite counterparts. In an earlier paper (Benoît and Krishna (1985)) we argued that this dichotomy arises only in games with a unique equilibrium.¹ We then showed that as long as the constituent game has multiple equilibrium payoffs for each player, the set of average payoffs from subgame perfect equilibria of the corresponding finitely repeated game converges to the set of feasible and individually rational payoffs as the

¹Strictly speaking the dichotomy arises when at least one player has a unique equilibrium payoff.

horizon increases.

This result, and indeed all "folk theorems," may be criticized on at least the following three related counts.

#1. The result is disappointing in that it demonstrates a lack of predictability in repeated game models since any feasible and individually rational payoff can arise in equilibrium for a long enough horizon.

#2. The theorem demonstrates that cooperative outcomes are possible, but does not favor them in any way. In fact, these outcomes are only a small subset of what can occur in equilibrium. Intuition and casual empiricism, however, suggest that repeated play leads to more cooperation than is indicated by such a neutral result.

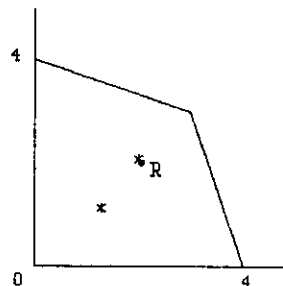
#3. A closer look at the proof reveals that for some games certain outcomes are enforced only by the threat of punishments which are Pareto dominated by other equilibrium paths. Since all players, "deviators" and "punishers" alike, suffer as a result, these equilibria are vulnerable to the possibility of renegotiation. Following a deviation, the deviating player can appeal to other players to let bygones be bygones and move to a Pareto dominating path. This appeal, if accepted, places in doubt the credibility of the original threat and hence that of the whole equilibrium.

To examine this last problem more closely, let us look at a simple example. Consider the game given below. There are two (pure strategy) equilibria marked by asterisks and the minmax payoff is 1 for each player. The game satisfies the hypotheses of the folk theorem for finitely repeated games and hence any feasible payoff vector that Pareto dominates (1,1) can be approximated as the average payoff in a finitely repeated game of long enough duration. For instance, the Pareto efficient payoff vector (3,3) may be approximated as follows. Consider any $T > 1$ and the outcome path that

results in the following sequence of payoffs: $(3,3)$ $(3,3)$... $[(T-1)$ times] ... $(3,3)$ $(2,2)$. Any deviations are met by reverting to repeated play of the one-shot equilibrium that results in $(1,1)$. This constitutes a

Example 0:

1,1*	0,0	4,0
0,0	2,2*	0,0
0,4	0,0	3,3



perfect equilibrium which for large T , approximates the payoff $(3,3)$ on average.

However, this equilibrium is vulnerable to the possibility of renegotiation and may be criticized for this reason. Suppose for the sake of argument that players can communicate with each other before the play in each period. Player 1 (say) could then deviate in period $(T-1)$ for a gain of 1 and then in period T suggest that his transgression be forgiven and instead of the sacrifice of playing the $(1,1)$ equilibrium, the $(2,2)$ equilibrium be played in the last period as originally planned. Since this proposal benefits player 2 as well as player 1, player 2 should accept it. Hence, both players should realize that, regardless of what may have been agreed upon beforehand, when period T comes up, they will actually play the $(2,2)$ equilibrium. Note that it is important that the last period play be an equilibrium and not just a Pareto efficient outcome, since it itself must be self-enforcing. Working backwards, in period $T-1$ the players must be playing a one-shot equilibrium since there are no believable threats available in period T . But again, a suggestion to play the $(1,1)$ equilibrium would

be rejected in favor of playing the Pareto dominating equilibrium that yields (2,2). Thus, (2,2) must result in the last two periods. Reasoning in this way, one can see that if renegotiation is allowed the only acceptable outcome is repeated play of (2,2). This is in marked contrast to what can be achieved as perfect equilibrium outcomes.

This paper explores the implications of allowing renegotiation in finitely repeated games. We are interested in subsets of perfect equilibria which are immune to criticism #3 of the folk theorem. The exact notion we employ is discussed below in section 3. Our main results concern the limiting behavior of the sets of payoffs resulting from renegotiation proof equilibria as the number of repetitions increases.

The problem of renegotiation in repeated games has been previously addressed in a number of papers² but by and large, all of this work concerns infinitely repeated games. For infinitely repeated games, there are substantial difficulties in formalizing these ideas and an acceptable notion of renegotiation proof equilibria remains elusive (see Farrell and Maskin (1989) for a discussion of the various proposals). By contrast, for finitely repeated games a satisfactory notion can be readily formulated and there seems to be little controversy about its adequacy. It is our understanding that the notion we adopt originated independently in an unpublished paper by J. Farrell and in Bernheim and Ray (1989). Bernheim and Whinston (1987) and earlier versions of Farrell and Maskin (1989) contain the same definition. To the best of our knowledge, no systematic study of renegotiation in finitely repeated games has previously been attempted.

²Farrell and Maskin (1987), Bernheim and Ray (1987), Pearce (1987), van Damme (1986).

2. Preliminaries

Let $G = (A_1, A_2; U_1, U_2)$ be a two-player game in normal form, where A_i is player i 's compact set of (pure) strategies, and $U_i : A_1 \times A_2 \rightarrow \mathbb{R}$ is player i 's payoff function, assumed to be continuous. Write $A = A_1 \times A_2$ as the set of outcomes of G with generic element a and $U : A \rightarrow \mathbb{R}^2$ as the function whose i th component is U_i . A payoff vector $u \in \mathbb{R}^2$ is said to be feasible if it is in the convex hull of $U(A)$; and is individually rational if it gives each player at least his/her (pure strategy) minmax payoff. Let F be the set of feasible and individually rational payoff vectors in G . Initially, we assume that G has at least one (pure strategy) Nash equilibrium.³

$G(T)$ is the game that results when G is successively played T times. The payoff to players from an outcome path (a^1, a^2, \dots, a^T) of $G(T)$ is the sum of the resulting per-period payoffs: $\sum_{t=1}^T U(a^t)$.

A (pure) strategy for player i in $G(T)$ is a function σ_i , which selects for any t and any history of play up to and including period $(t-1)$, an element of A_i . Given a strategy combination $\sigma = (\sigma_1, \sigma_2)$, let $a^t(\sigma)$ be the resulting outcome at time t and write the resulting payoff as $U(\sigma) = \sum_{t=1}^T U(a^t(\sigma))$. For $K < T$ let $h(K) = (a^1, a^2, \dots, a^K)$ denote a K -period history and let $\sigma|_{h(K)}$ denote the strategy combination induced by σ on the subgame $G(T-K)$ following $h(K)$. $U(\sigma|_{h(K)})$ will be called the continuation payoff prescribed by σ on $G(T-K)$ following $h(K)$.

The strategy combination $\sigma = (\sigma_1, \sigma_2)$ is a (subgame) perfect equilibrium of $G(T)$ if it is a Nash equilibrium of $G(T)$ and for any $K < T$ and

³Mixed strategies are considered in section 6.

any K -period history $h(K)$, the strategies $\sigma|_{h(K)}$ induced by σ on $G(T-K)$ also form a Nash equilibrium of $G(T-K)$.

Let $P(T)$ be the set of total payoffs resulting from perfect equilibria of $G(T)$, that is

$$P(T) = \{ U(\sigma) \mid \sigma \text{ is a perfect equilibrium of } G(T) \}.$$

It will be useful to recall the following theorem for two player games from Benoît and Krishna (1985). It characterizes the limiting behavior of the set of average payoffs from perfect equilibria of $G(T)$.⁴

Theorem 0: Suppose that each player has at least two Nash equilibrium payoffs in the game G . Then,

$$\lim_{T \rightarrow \infty} (1/T)P(T) = F.$$

⁴Whenever we speak of limits of sets, we use the Hausdorff metric on the non-empty closed subsets of F . See Hildenbrand (1974) for details.

3. Renegotiation Proof Equilibria

In this section we explore a model where players (a) can communicate in each period prior to choosing their actions; and (b) are unable to sign binding contracts. The process of communication and negotiation is not explicitly modelled but rather used to motivate the definition. First of all, (b) implies that players must respect incentive compatibility constraints and hence always play a (perfect) equilibrium. Next, (a) implies that from the set of "acceptable" paths, only those that are efficient will be played. These considerations lead to the following recursive definition. In a one-shot game (a) and (b) lead to the selection of equilibria that are not Pareto dominated by any other equilibria. Consistency requires that in $G(2)$, the second period play should involve an efficient one-shot equilibrium. Thus, in the first period, only outcomes which are sustained by threats of efficient one-shot equilibria are considered. Of these paths, only those that are efficient are played. Working backward in this manner, a complete definition may be obtained. A formal definition is given below.

We emphasize that the assumption of explicit communication is made for the purposes of motivation only, and does not indicate a significant departure from the traditional non-cooperative approach. In particular, when choosing actions, players must always respect incentive compatibility constraints. A proposal to select a Pareto dominating outcome which is not incentive compatible is not viable. Also a theory of play thus derived must be dynamically consistent in the sense that it requires that the theory be applied in every subgame at every stage. For this reason, a proposal to play a Pareto optimal perfect equilibrium is may not be viable since it may be

vulnerable to renegotiation off the equilibrium path.

y repeated games also and call such equilibria consistent.

For any set $S \subset \mathbb{R}^2$ define $\text{Eff } S$ to be the set of (weakly) efficient points in S : $\text{Eff } S = \{ x \in S \mid \text{there is no } y \in S \text{ such that } y \gg x \}$.⁵

Renegotiation proof equilibria of $G(T)$ form a subset of the set of perfect equilibria of $G(T)$, and are defined recursively as follows:

$$Q(1) = P(1)$$

$$R(1) = \text{Eff } Q(1)$$

and for $T > 1$,

$$Q(T) = \{ U(\sigma) \in P(T) \mid \text{all continuation payoffs prescribed by } \sigma \text{ on } G(T-1) \text{ lie in } R(T-1) \}$$

$$R(T) = \text{Eff } Q(T).$$

A perfect equilibrium σ of $G(T)$ is said to be renegotiation proof if $U(\sigma) \in R(T)$, and for any $K < T$ and any K -period history $h(K)$, $U(\sigma|_{h(K)}) \in R(T-K)$.

In what follows we will only be interested in the set of payoffs accruing from such equilibria and not the strategies themselves. However, as we will see, given the sets $R(T)$, a pair of renegotiation proof equilibrium strategies that sustain payoffs in $R(T)$ can be easily recovered.

We write $R(T)$ for $T > 1$ as follows:

$$W_i(T-1) = \inf \{ X_i \mid (X_1, X_2) \in R(T-1) \};$$

$$Q(T) = \{ U(a) + X \mid X \in R(T-1) \text{ and for } i = 1, 2 \\ [\max_{\alpha} U_i(\alpha, a_{-i})] + W_i(T-1) \leq U_i(a) + X_i \}$$

$$R(T) = \text{Eff } Q(T).$$

It is shown below that the sets $R(T)$ are compact and hence the infimum

⁵For $x, y \in \mathbb{R}^2$, $x \gg y$ denotes that for $i = 1, 2$, $x_i > y_i$.

in the definition of $W_i(T-1)$ is attained in $R(T-1)$. Any payoff in $R(T)$ can thus be sustained by threatening to meet any deviation by player i in period 1 by moving to the equilibrium that yields $W_i(T-1)$ in the game $G(T-1)$.

We claim that for all T , $Q(T)$ and $R(T)$ are non-empty and closed subsets of F . That $Q(T)$ and $R(T)$ are non-empty follows trivially from the definitions. To verify that they are closed we argue by induction. $Q(1)$, the set of Nash equilibrium payoffs of G , is closed and if S is a closed set then so is $\text{Eff } S$. Thus $R(1) = \text{Eff } Q(1)$ is also closed. Now suppose that $Q(T-1)$ and $R(T-1)$ are closed. If $Z^n \rightarrow Z$ and for all n , $Z^n \in Q(T)$, then we claim that $Z \in Q(T)$ also. Let $Z^n = U(a^n) + X^n$ where $X^n \in R(T-1)$. Then taking a convergent subsequence of (a^n, X^n) converging to (a, X) and using the fact that U is continuous, implies that $X \in R(T-1)$ and for $i = 1, 2$, $[\max_{\alpha} U_i(\alpha, a_{-i})] + W_i(T-1) \leq U_i(a) + X_i = Z_i$. Hence, $Z \in Q(T)$. Thus, $Q(T)$ is a closed set and then so is $R(T) = \text{Eff } Q(T)$.

Next, notice that if G has a single equilibrium that Pareto dominates all other equilibria, as in Example 0, then for all T , $(1/T)R(T) = R(1)$. Thus, to be of any interest, G must have at least two efficient equilibria.

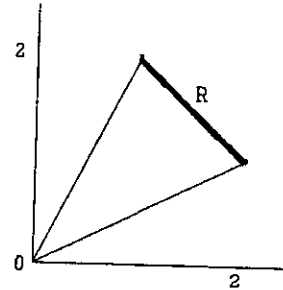
Just as Theorem 0 concerns the limiting behavior of $(1/T)P(T)$, in this paper we study the limiting behavior of $(1/T)R(T)$. We begin by analyzing the behavior of $(1/T)R(T)$ in a few representative examples. These will then serve to motivate the results of section 5.

4. Some Motivating Examples

In this section we present four examples chosen to exhibit a range of possible phenomena.

Example 1:

0,0	1,2*
2,1*	0,0

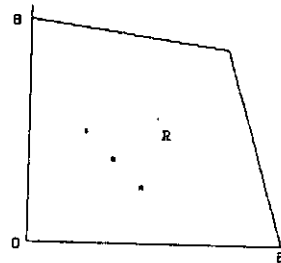


Our first example is the "battle of the sexes." There are two pure strategy equilibria both of which lie on the Pareto frontier of the feasible set (see the accompanying figure) and hence are in $R(1)$. It is easy to see that for all T , $R(T)$ can be derived by summing all the elements of $R(1)^T$ and that as a result,

$$\lim_{T \rightarrow \infty} (1/T)R(T) = R = \text{Co} \{(1,2), (2,1)\}.$$
⁶

Example 2:

0,0	2,4*	0,0	8,0
4,2*	0,0	0,0	0,0
0,0	0,0	3,3*	0,0
0,8	0,0	0,0	7,7



⁶For any set $S \subset \mathbb{R}^n$, $\text{Co } S$ denotes the convex hull of S .

Our next example, which is a variation of an example analyzed by Bernheim and Whinston (1987), displays more interesting behavior. In this case, $R(1)$ consists of three elements. To determine $R(2)$, first notice that $(10,10) \in Q(2)$. This is obtained by following the path that yields the payoffs $(7,7)$, $(3,3)$. The equilibrium path is sustained by the threat of playing the $(2,4)$ equilibrium in the second period if player 1 deviates and the $(4,2)$ equilibrium if player 2 does. All other elements of $Q(2)$ can be obtained from paths which consist of repeating or alternating between equilibria of G . Thus,

$$Q(2) = \{(4,8), (8,4), (5,7), (7,5), (6,6), (10,10)\}; \text{ and hence}$$

$$R(2) = \text{Eff } Q(2) = \{(10,10)\}.$$

But this implies that no player can be punished for deviating in the first period of $G(3)$. Thus, in this first period an element of $R(1)$ must be played, so that, $R(3) = R(1) + \{(10,10)\}$. The sequence of $R(T)$'s is given below:

$$\begin{aligned} R(1) &= \{(2,4), (3,3), (4,2)\} \\ R(2) &= \{(10,10)\} \\ R(3) &= \{(12,14), (13,13), (14,12)\} \\ &\quad \vdots \\ &\quad \vdots \\ R(2n) &= \{(10n,10n)\} \\ R(2n+1) &= \{(10n+2,10n+4), (10n+3,10n+3), (10n+4,10n+2)\} \end{aligned}$$

Hence, for all T , $R(T) = R(T-2) + \{(10,10)\}$; and

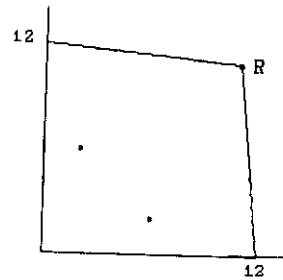
$$\lim_{T \rightarrow \infty} (1/T)R(T) = \{(5,5)\}.$$

This example is rather striking in that repetition yields a unique equilibrium payoff vector in the limit even though G itself has three

equilibria. Contrary to the lessons of the folk theorem, with renegotiation, repetition actually narrows the set of equilibrium payoffs in this example. Furthermore, the limiting set of payoffs is bounded away from the Pareto frontier of the feasible set. This highlights the difference between requiring equilibria to be renegotiation proof and simply selecting perfect equilibria which are Pareto efficient.

Example 3:

0,0	6,2*	12,0
2,6*	0,0	0,0
0,12	0,0	11,11



In our third example, $\lim (1/T)R(T)$ is again a singleton even though $R(T)$ never shrinks to a single point as in example 2. In addition, the limit lies on the Pareto frontier of the feasible set. The sequence of $R(T)$'s may be written as follows.

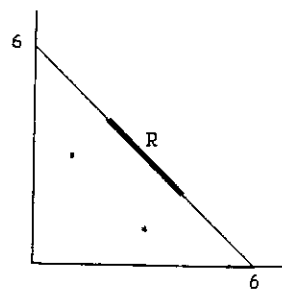
$$\begin{aligned}
 R(1) &= \{(2,6), (6,2)\} \\
 R(2) &= \{(6,14), (8,8), (14,6)\} \\
 R(3) &= \{(8,20), (19,19), (20,8)\} \\
 R(4) &= \{(19,31), (30,30), (31,19)\} \\
 &\quad \vdots \\
 &\quad \vdots \\
 R(T) &= \{(11T-25, 11T-14), (11T-14, 11T-14), (11T-13, 11T-25)\}
 \end{aligned}$$

For $T > 3$ we have the recursive relation, $R(T) = R(T-1) + \{(11,11)\}$.

Hence, $\lim_{T \rightarrow \infty} (1/T)R(T) = \{(11,11)\} \subset \text{Eff } F$.

Example 4:

0,0	1,3*	0,0
3,1*	0,0	6,0
0,0	0,6	0,0



Our last example is more complicated. Unlike our earlier examples, in this case the limit of $(1/T)R(T)$ is not a singleton. The sequence of $R(T)$'s is determined below. The following notation is useful. Let $X = (x^1, x^2, \dots, x^k)$ be an ordered array of k vectors from R^2 and let $y \in R^2$. Define $X+y$ to be the ordered array $(x^1+y, x^2+y, \dots, x^k+y)$. Then,

$$R(1) = \{(1,3), (3,1)\}$$

$$R(2) = \{(3,7), (4,4), (7,3)\}$$

$$R(3) = \{(4,10), (7,9), (9,7), (10,4)\}$$

$$R(4) = \{(7,15), (9,13), (10,10), (13,9), (15,7)\}$$

⋮

$$R(3n) = \{(6n-2, 12n-2), (6n+1, 12n-3), X(n), X(n)+(6,-6), X(n)+2(6,-6), X(n)+3(6,-6), \dots, X(n)+(n-1)(6,-6), (12n-3, 6n+1), (12n-2, 6n-2)\}$$

$$R(3n+1) = \{(6n+1, 12n+3), Y(n), Y(n)+(6,-6), Y(n)+2(6,-6), Y(n)+3(6,-6), \dots, Y(n)+(n-1)(6,-6), (12n+3, 6n+1)\}$$

$$R(3n+2) = \{Z(n), Z(n)+(6,-6), Z(n)+2(6,-6), \dots, Z(n)+n(6,-6)\}$$

where $X(n) = ((6n+3, 12n-5), (6n+4, 12n-8), (6n+7, 8n-9))$; $Y(n) = X(n)+(0,6)$; and $Z(n) = Y(n)+(0,6) = X(n)+(0,12)$.

It can be verified that

$$\lim_{T \rightarrow \infty} (1/T)R(T) = \text{Co} \{(2,4), (4,2)\} \subset \text{Eff } F.$$

5. The Limiting Behavior of $(1/T)R(T)$

Our main result, Theorem 1 below, shows that the (Hausdorff) limit of $(1/T)R(T)$, if it exists, must have at least one of two characteristics. Either $(1/T)R(T)$ shrinks to a single point (as in examples 0, 2 and 3) or it approaches a subset of the (weak) Pareto frontier of the feasible set (as in examples 1, 3 and 4).

Theorem 1: $\lim (1/T)R(T)$ is a singleton;
or $\lim (1/T)R(T) \subset \text{Eff } F$.

Notice that the theorem concerns the behavior of the limit of $(1/T)R(T)$.⁷ The limit was shown to exist in all of the examples in the previous section. While we have been unable to establish that it exists for arbitrary games, we strongly conjecture that it does.⁸

Although the proof of the theorem is somewhat involved, the basic underlying idea is rather simple and may be worth sketching informally. Let R denote the limit of $(1/T)R(T)$. First, notice that R is a closed set since the Hausdorff limit of a sequence of closed sets is by definition closed. It can be argued (Lemma 1) that R is a connected subset of F . Suppose it were not connected but consisted of many segments, as in Figure 1. Consider the points x and y . Although x and y may be close to each other, they

⁷For the class of finite symmetric games, a result stronger than Theorem 1--one that does not rely on the existence of the limit of $(1/T)R(T)$ --can be established. See the appendix for details.

⁸It can be established that the limit exists for a generic class of 2×2 games.

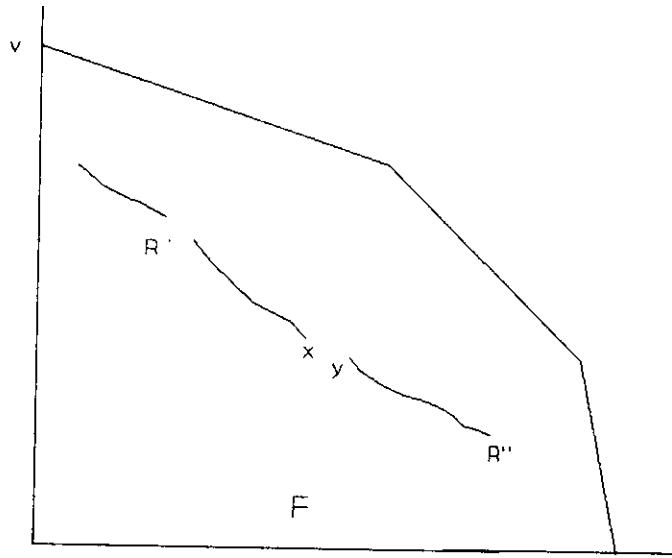


Figure 1

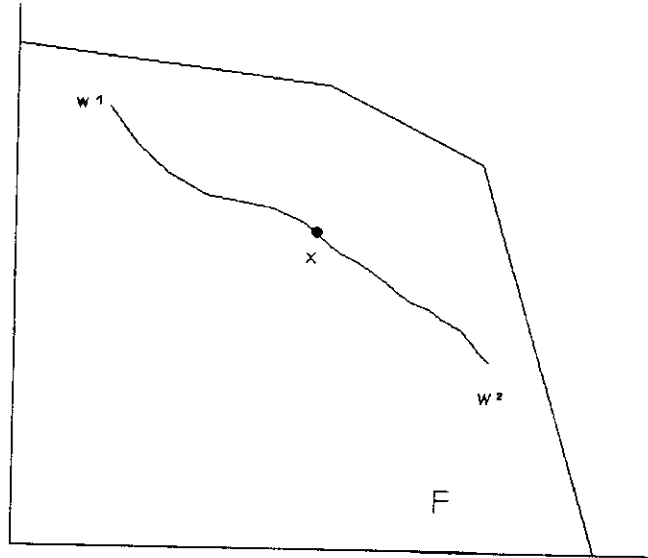


Figure 2

represent limits of average payoffs and the corresponding outcome paths may be made arbitrarily far in terms of total payoffs as T increases. To y (actually, to points close to y) we can repeatedly add the outcome v that yields player 2 his highest payoff in the game. This is because player 2 is at a best response at this outcome and player 1 can be threatened with outcome x (actually, points close to x). But this "moves" the outcome y upwards, contradicting the assumption that there are no limit points "between" x and y . Hence R is connected.

Next, suppose that there are points in R , such as x in Figure 2, that are inefficient. As in the argument above, x can be made arbitrarily far from w^1 and w^2 in terms of total payoffs. Thus for large T , there is no incentive constraint to prevent Pareto dominating payoffs from being added to elements of $R(T)$ whose averages are close to x . Hence, no point in the relative interior of R can be Pareto inefficient. The connectedness of R ensures that the same is also true of the end points. The formal proof proceeds along these lines but also takes care of cases where R may have horizontal or vertical segments.

We will need some additional notation in what follows. For two vectors $x, y \in \mathbb{R}^2$, define $x \setminus y$ to denote that $x_1 \leq y_1$ and $x_2 \geq y_2$. Let $R = \lim (1/T)R(T)$ and if $x, y \in R$, define $R(x, y) = \{z \in R \mid x \setminus z \setminus y\}$. For any closed set $S \subset \mathbb{R}^2$, $B(S; \epsilon) = \{y \mid \text{dist}(y, S) < \epsilon\}$ will denote the open ϵ -ball around S . Finally, if $X(T) \in R(T)$, we will write its lower case counterpart $x(T) = (1/T)X(T)$ as the corresponding average payoff.

Lemma 1: $\lim (1/T)R(T)$ is a connected set.

Proof: We argue by contradiction. Suppose R is not a connected set so that there exist points $x, y \in R$, $x \setminus y$, and $x \neq y$ such that $R(x, y) = \{x, y\}$. Define $R' = \{z \in R \mid z \setminus x\}$ and $R'' = \{z \in R \mid y \setminus z\}$. R' and R'' are closed sets such that $R' \cup R'' = R$ and $R' \cap R'' = \emptyset$. Furthermore, $z' \in R'$ and $z'' \in R''$ implies that $z' \setminus z''$.

There exists an $\varepsilon^* > 0$ such that for all $\varepsilon < \varepsilon^*$, there exists a $T(\varepsilon)$ such that for all $T > T(\varepsilon)$, $(1/T)R(T) \subset B(R; \varepsilon) = B(R'; \varepsilon) \cup B(R''; \varepsilon)$ and $B(R'; \varepsilon) \cap B(R''; \varepsilon) = \emptyset$. For each $T > T(\varepsilon)$, we can partition $R(T)$ as $R(T) = R'(T) \cup R''(T)$ satisfying $(1/T)R'(T) \subset B(R'; \varepsilon)$ and $(1/T)R''(T) \subset B(R''; \varepsilon)$. Let $x(T) \in (1/T)R'(T)$ satisfying for all $z \in (1/T)R'(T)$, $z \setminus x(T)$ and $y(T) \in (1/T)R''(T)$ satisfying for all $z \in (1/T)R''(T)$, $y(T) \setminus z$.

Let v_2 denote player 2's highest payoff in the game G and let $v = (v_1, v_2)$ be such that for all $z \in F$, $z_2 = v_2$ implies that $v_1 \leq z_1$.

There are two cases to consider.

Case (i): $v_2 > y_2$. Choose $\varepsilon < \min \{(y_1 - x_1)/3, (v_2 - y_2)/3, \varepsilon^*\}$. There exists a $T^* > T(\varepsilon)$ such that, for all $T > T^*$, $T[(y_1 - \varepsilon) - (x_1 + \varepsilon)]$ is greater than the gain from any one-shot deviation by player 1. For such a T , $v + Ty(T) \in Q(T+1)$. This is because player 2 has no incentive to deviate in the first period of $G(T+1)$ and player 1 can be made to conform by threatening him with $x(T)$ if he deviates. Let $Z(T+1) \in R(T+1)$ satisfying $Z(T+1) \geq v + Ty(T)$. Next, for $k = 2, 3, \dots$ inductively define $Z(T+k) \in R(T+k)$ to satisfy $Z(T+k) \geq v + Z(T+k-1)$ if $z(T+k-1) \in B(R''; \varepsilon)$. There exists a large enough T and a K such that for all $k < K$, $z(T+k) \in B(R''; \varepsilon)$ but $z_2(T+K) > y_2 + \varepsilon$ and $z_1(T+K) > x_1 + \varepsilon$. Thus, $z(T+K)$ lies neither in $B(R'; \varepsilon)$ nor in $B(R''; \varepsilon)$. This contradicts the fact that $(1/(T+K))R(T+K) \subset B(R; \varepsilon)$.

Case (ii): $v_2 = y_2$. Choose $\varepsilon < \min \{(y_1 - x_1)/3, \varepsilon^*\}$. As before there exists a $T^* > T(\varepsilon)$ such that, for all $T > T^*$, $T[(y_1 - \varepsilon) - (x_1 + \varepsilon)]$ is greater than the gain from any one-shot deviation by player 1. There also exists a $T^{**} > T^*$ such that for all $z'(T) \in (1/T)R'(T)$, for all $z''(T) \in (1/T)R''(T)$, and for all u in the range of the payoff function U , $[(u + Tz'(T))/(T+1)] \in B(R'; \varepsilon)$ and $[(u + Tz''(T))/(T+1)] \in B(R''; \varepsilon)$. First, note that for $T > T^{**}$, $v + Ty(T) \in Q(T+1)$. This is because player 2 once again has no incentive to deviate in the first period of $G(T+1)$ and player 1 can be made to conform by threatening him with $x(T)$ if he deviates. Furthermore, $v + Ty(T) \in R(T+1)$. This is because by the choice of $T > T^{**}$, no element of $R'(T+1)$ can Pareto dominate $v + Ty(T)$; and by construction, neither can any element of $R''(T+1)$. Now consider for $k = 2, 3, \dots$ the point $kv + Ty(T)$. There exists a K such that for all $k < K$, $kv + Ty(T) \in R(T+k)$, $Kv + Ty(T) \in Q(T+K)$ and either (a) $Kv + Ty(T) \in R(T+K)$ and $x_1 + \varepsilon < [Kv_1 + Ty_1(T)]/(T+K) < y_1 - \varepsilon$; or (b) $Kv + Ty(T) \notin R(T+K)$ because there is an element $[\sum_{k=1}^K u^k + Tz'(T)] \in R(T+K)$, where $Tz'(T) \in R'(T)$ and the u^k 's are in $U(A)$, which Pareto dominates $Kv + Ty(T)$. If (a) occurs we have a contradiction, since then $Kv + Ty(T)$ is not in $B(R; \varepsilon)$. If (b) occurs, we again have a contradiction since for some $L \leq K$, $x_1 + \varepsilon < [\sum_{k=1}^L u_1^k + Tz_1'(T)]/(T+L) < y_1 - \varepsilon$. □

Proof of Theorem 1: Let $w^1 = (w_1, b_2)$ and $w^2 = (b_1, w_2)$ be elements of $R = \lim (1/T)R(T)$ such that for all $z \in R$, $w^1 \setminus z \setminus w^2$. Suppose R contains more than one point so that $w^1 \neq w^2$.

Let $x \in R$ be such that $x \neq w^1$ and $x \neq w^2$. We will argue that $x \in \text{Eff } F$.

Suppose $x \notin \text{Eff } F$. Then there exists a set of K points u^1, u^2, \dots, u^K such that $u = (1/K)\sum_{k=1}^K u^k \gg x$, and each u^k is the payoff from some (pure)

action choices in G .

There are three separate cases to consider.

Case (i): $x_1 > w_1$ and $x_2 > w_2$. Let $x(T)$, $w^1(T)$ and $w^2(T)$ be sequences of elements in $(1/T)R(T)$ that converge to x , w^1 and w^2 respectively. Choose $0 < \varepsilon < \min \{\|x-w^1\|/3, \|x-w^2\|/3\}$. There exists a T^* such that for all $T > T^*$, $x(T)$, $w^1(T)$ and $w^2(T)$ are all within ε of x , w^1 and w^2 respectively. There also exists a $T^{**} > T^*$ such that for all $T > T^{**}$, the point $[(Tx(T)+kv)/(T+k)]$ is within ε of x for all v in the range of the payoff function U and for all $k \leq K$. Let D denote the largest gain that any player can get by a one shot deviation. There exists a $T^{***} > T^{**}$ such that for all $T > T^{***}$, $T\varepsilon > D$.

Choose a $T > T^{***}$ and consider the total payoff $Tx(T) + u^1$ resulting from a path where players' choices result in a payoff of u^1 in the first period followed by the equilibrium play resulting in a total payoff of $Tx(T)$ in the last T periods. The choice of T guarantees that $Tx(T) + u^1 \in Q(T+1)$ since we have that for $i=1,2$ $T[x_i(T)-w_i(T)] > D$. Thus, there exists a point $Z(T+1) \in R(T+1)$ satisfying $Z(T+1) \geq Tx(T) + u^1$. Now for $n = 2,3,\dots$ inductively define $Z(T+n) \in R(T+n)$ to satisfy $Z(T+n) \geq Z(T+n-1) + u^{n \pmod K}$. To see that such a $Z(T+n)$ exists, notice that for all n , $Z(T+n-1) + u^{n \pmod K} \in Q(T+n)$. This is because by our choice of T , for all n , for $i=1,2$ $[Z_i(T+n-1)-(T+n-1)w_i(T+n-1)] > D$. For large n , $z(T+n) \gg x(T+n)$ which is a contradiction.

Case (ii): $x_1 = w_1$ but $x_2 > w_2$. First, suppose that $b_1 = w_1$ so that R is a vertical line.

For any $\varepsilon > 0$ define $X = \bar{B}(R;\varepsilon) \cap \text{Co}\{\{\bar{u}\} \cup \bar{B}(x;\varepsilon)\}$, where \bar{u} is the payoff vector where player 1 attains his highest payoff in the game. $\bar{B}(R;\varepsilon)$ is the closure of $B(R;\varepsilon)$, etc.

Let $y \in X$ be such that $y_2 \leq z_2$ for all $z \in X$. By choosing a small enough ε , we can ensure that $y_2 > w_2$. There exists a T^* , such that for all $T > T^*$, $T[y_2 - w_2] > D$, the largest deviation by a player in the game.

For $T > T^*$ let $x(T) \in B(x; \varepsilon)$. Then, $\bar{u} + Tx(T) \in Q(T+1)$. This is because player 1 is at a best response in period 1 when \bar{u} is played and because 2 can be prevented from deviating by a threat of playing the worst equilibrium in $R(T)$, since $T > T^*$ and $x_2(T) > y_2$. Thus there exists a $Z(T+1) \in R(T+1)$ such that $Z(T+1) \geq \bar{u} + Tx(T)$. For $k = 2, 3, \dots$ inductively define $Z(T+k) \in R(T+k)$ to satisfy $Z(T+k) \geq \bar{u} + Z(T+k-1)$. Notice that by construction $z_1(T+k) \leq x_1 + \varepsilon$ implies that $z_2(T+k) > y_2$ so that player 2 can be prevented from deviating. For some $k = K$ we will have that $z_1(T+K) > x_1 + \varepsilon$ contradicting the fact that $(1/(T+K))R(T+K) \subset B(R; \varepsilon)$.

Now suppose that $b_1 > w_1$. Define the point $z \in R$ by $z_1 = x_1 = w_1$ and if $y \in R$ satisfies $y_1 = w_1$ then $z_2 \leq y_2$. By case (i) above, every point in the set $R(z, w^2) = \{y \in R \mid z \setminus y \setminus w^2\}$ save possibly z and w^2 is efficient. But since R is a connected set z and w^2 are also efficient. Finally since $x \geq z$, x is also efficient.

Case (iii): $x_1 = w_1$ and $x_2 = w_2$. Since R is connected, there exist points $y \in R$, $y \setminus x$ that are arbitrarily close to x . By reproducing the argument in case (ii), all such points $y \in \text{Eff } F$. Thus, $x \in \text{Eff } F$ also.

We have argued that if R does not consist of a single point then any $x \in R$ satisfying $x \neq w^1$ and $x \neq w^2$ is an element of $\text{Eff } F$. Since R is connected w^1 and w^2 are also efficient. □

It is possible for R to be a single point on the Pareto frontier of F . Trivial examples are "games of coordination", where a single Pareto

dominating equilibrium is efficient. A non-trivial instance of this occurs in Example 3.

Our next result says that the best payoff for a player in $R(T)$ must be, on average, at least as large as his best equilibrium payoff in the one-shot game.

Lemma 2: Let $B^1(T) = (B_1(T), W_2(T))$ and $B^2(T) = (W_1(T), B_2(T))$ be elements of $R(T)$ satisfying: for all $Z(T) \in R(T)$, $B^2(T) \setminus Z(T) \setminus B^1(T)$.

Then for $i = 1, 2$ and for all T ,

$$(1/T)B_i(T) \geq B_i(1). \quad (*)$$

Proof: The proof is by induction. By definition, (*) is true for $T = 1$.

Suppose that $B_i(T-1) \geq (T-1)B_i(1)$ and notice that for $i = 1, 2$

$[B^i(1) + B^i(T-1)] \in Q(T)$. Hence there exists a $Z(T) \in R(T)$ satisfying $Z(T) \geq B^i(1) + B^i(T-1)$. Since, $B_i(T) \geq Z_i(T)$, we have that

$$B_i(T) \geq B_i(1) + B_i(T-1) \geq TB_i(1). \quad \square$$

Lemma 2 immediately yields a simple sufficient condition on the game G that ensures that the limit of $(1/T)R(T)$ is a subset of the Pareto frontier of F .

$$\text{Let } b_i = B_i(1) = \max \{x_i \mid (x_1, x_2) \in R(1)\}.$$

Theorem 2: If $(b_1, b_2) \notin F$, then $\lim (1/T)R(T) \subset \text{Eff } F$.

Proof: Suppose that $(b_1, b_2) \notin F$. Lemma 2 then implies that R cannot consist of a single point. By Theorem 1, it must, therefore, be a subset of $\text{Eff } F$. □

6. Mixed Strategies

In the previous sections we restricted attention to pure strategies or equivalently to the case when players play possibly mixed strategies at every stage of $G(T)$, but their choice of the randomizing device is observable by other players prior to the next period. This is accomplished by interpreting the actions $a_i \in A_i$, as random variables over some underlying pure strategy sets. In this section we illustrate how our results extend, virtually unchanged, to the standard case when only the pure strategy outcomes chosen by the randomizing devices are observable and not the devices themselves.

For any set S , let $\Delta(S)$ denote the set of probability distributions over S . A behavior strategy for player i in $G(T)$ is a function $\tilde{\sigma}_i$, which selects for any t and any history of play $h(t-1) = (a^1, a^2, \dots, a^{t-1})$ an element of $\Delta(A_i)$. Given a behavior strategy combination $\tilde{\sigma} = (\tilde{\sigma}_1, \tilde{\sigma}_2)$, let $U(\tilde{\sigma})$ denote the sum of the expected payoffs in each period and for any $K < T$ and any K -period history $h(K)$, define $\tilde{\sigma}|_{h(K)}$ to be the behavior strategy combination induced by $\tilde{\sigma}$ on the subgame $G(T-K)$ following $h(K)$. Subgame perfect equilibria of $G(T)$ are defined as before and let $\tilde{P}(T)$ denote the set of expected payoffs from subgame perfect equilibria of $G(T)$ when players are allowed to randomize. Renegotiation proof equilibria may be defined analogously as follows:

$$\tilde{Q}(1) = \tilde{P}(1)$$

$$\tilde{R}(1) = \text{Eff } \tilde{Q}(1)$$

and for $T > 1$,

$$\tilde{Q}(T) = \{ U(\tilde{\sigma}) \in \tilde{P}(T) \mid \text{all continuation expected payoffs prescribed by } \tilde{\sigma} \text{ on } G(T-1) \text{ lie in } \tilde{R}(T-1) \}$$

$$\tilde{R}(T) = \text{Eff } \tilde{Q}(T).$$

A (mixed strategy) perfect equilibrium $\bar{\sigma}$ of $G(T)$ is renegotiation proof if $U(\bar{\sigma}) \in \tilde{R}(T)$ and for any $K < T$ and any K -period history $h(K)$, $U(\bar{\sigma}|_{h(K)}) \in \tilde{R}(T-K)$.

The following result is analogous to Theorem 1.

Theorem 1*: $\lim (1/T)\tilde{R}(T)$ is a singleton;
or $\lim (1/T)\tilde{R}(T) \subset \text{Eff } F$.

Proof: The proof follows that of Theorem 1 exactly. This is because both Lemma 1 and Theorem 1 are proved by constructing pure strategy renegotiation proof equilibrium paths to obtain a contradiction. Since the construction is still valid when players are allowed to randomize, the same contradictions can still be obtained. □

Notice that Theorems 1 and 1* are independent in the sense that neither directly implies the other. Also, for a given game the limit of $(1/T)R(T)$ may be very different from that of $(1/T)\tilde{R}(T)$.

Lemma 2 and Theorem 2 also apply to the case of unobservable mixed strategies with no change in their proofs.

7. Concluding Remarks

Renegotiation proof equilibria are obtained by imposing what may be termed "dynamic incentive efficiency", that is, players' choices are efficient given that they must respect dynamic incentive compatibility constraints. As such no outside enforcement of the agreed upon "contracts" is required since all renegotiation proof equilibria are perfect equilibria. One may argue that because of this, such equilibria remain within the bounds of traditional non-cooperative game theory. On the other hand, efficiency, albeit from the set of incentive compatible agreements, is assumed and not derived as the result of individual decisionmaking. From this perspective, renegotiation proof equilibria transcend the traditional boundaries of non-cooperative game theory.

Our results should be contrasted with the "folk theorem" for finitely repeated games. Recall the three intuitive grounds on which the folk theorem for repeated games may be criticised.

#1. It indicates a surfeit of equilibria and consequently, a lack of predictability.

#2. It is neutral, in the sense that efficient outcomes are no more likely to emerge than any other feasible and individually rational outcome.

#3. Perfect equilibria may be vulnerable to renegotiation.

Theorem 1 provides an intriguing link among the three criticisms. Our original goal was to address the third criticism, concerning the possibility of renegotiation. However, the theorem also addresses the first two criticisms in quite a strong fashion: if renegotiation is allowed, increasing the number of repetitions either leads to a unique equilibrium payoff or it

ensures that all equilibrium payoffs are Pareto efficient. Stated another way, the only way in which the limit of the set of payoffs can be inefficient is if it is a single point. Moreover, even though the idea of renegotiation is predisposed towards efficiency, it does not always result and is not even a generic property in any sense (perturbing the payoffs in Example 2 yields an open set of games for which the limit of payoffs continues to be a single inefficient point).

As mentioned earlier, The bulk of the work on renegotiation in repeated games has been concerned with infinitely repeated games. The absence of a last period introduces substantial difficulties in formalizing the basic idea of renegotiation in such games and many alternatives have been considered (Farrell and Maskin (1989) survey the various proposals). There does not appear to be any obvious sense in which any of the definitions are the "same" as that for the finite case. For this reason it is difficult to compare our results with those obtained in the infinite context. The reformulation of the renegotiation idea in a way that a unified definition emerges for both the finite and infinite cases is an important and active area of research.

8. References

- Benoit, J-P. and V. Krishna: "Finitely Repeated Games," Econometrica, 53 (1985), 890-904.
- Bernheim, B.D., B. Peleg and M.D. Whinston: "Coalition Proof Nash Equilibria, I: Concepts," Journal of Economic Theory, 42 (1987), 1-12.
- Bernheim, B.D., and M.D. Whinston: "Coalition Proof Nash Equilibria, II: Applications," Journal of Economic Theory, 42 (1987), 13-29.
- Bernheim, B.D., and D. Ray: "Collective Dynamic Consistency in Repeated Games," Games and Economic Behavior, 1 (1989), 295-326.
- Farrell, J. and E. Maskin: "Renegotiation in Repeated Games," Games and Economic Behavior, 1 (1989), 327-360.
- Hildenbrand, W.: Core and Equilibria of a Large Economy, Princeton University Press, Princeton, 1974.
- Pearce, D.: "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation," Mimeo, Yale University, June 1987.
- van Damme, E.: "Renegotiation-Proof Equilibria in Repeated Prisoners' Dilemma," Journal of Economic Theory, 47 (1989), 206-217.

Appendix

The results of section 5, especially Theorem 1, rely on the existence of the limit of $(1/T)R(T)$. We now provide a stronger result for the class of symmetric games--one that does not presume that this limit exists.

A two player game $G = (A_1, A_2; U_1, U_2)$ is said to be symmetric if $A_1 = A_2$ and for all $(a_1, a_2) \in A_1 \times A_2$, $U_1(a_1, a_2) = U_2(a_2, a_1)$. It is said to be finite if A_1 and A_2 are finite sets.

Theorem 3: Let G be a finite symmetric game. Then,

$\lim (1/T)R(T)$ exists and is a singleton;

or for all ϵ , there exists a $T(\epsilon)$ such that for all $T > T(\epsilon)$,

$$(1/T)R(T) \subset B(\text{Eff } F; \epsilon).$$

Although the proof of Theorem 3 involves many ideas familiar from section 5, we are unable to argue by contradiction as in the proof of Theorem 1. Because the arguments cannot be carried out directly at the limit, the proof of Theorem 3 involves a rather elaborate construction of equilibrium paths. The role of symmetry is to ensure that unless $(1/T)R(T)$ converges to a single point, there are always points in $R(T)$, close to the 45° line, which are "far" from both extreme points. To these points any payoff vector in the original game may be added and hence only efficient vectors will be added. The basic idea of the proof, which we sketch only informally, is to approximate any path by a suitably constructed path which is "far" from both extreme points and hence nearly efficient. As will become apparent, the

existence of paths close to the 45° line (because of symmetry) greatly facilitates this construction.

Sketch of Proof of Theorem 3: Since G is symmetric, so are the sets $R(T)$. As before, define $B_1(T)$ and $W_1(T)$ to be player 1's best and worst payoffs in $R(T)$ respectively. There are two cases to consider.

Case (i): Suppose that there exists an M such that for all T' there exists a $T > T'$ such that $B_1(T) - W_1(T) \leq M$.

By the finite nature of G , there exists a T and a smallest S such that $R(T+S) = R(T) + \{V\}$ for some vector $V = \sum_{k=1}^S v^{T+k}$ and $v^{T+k} \in U(A)$. In this case, the sets $R(T)$ cycle every S periods. Thus, $\lim (1/T)R(T)$ exists and is equal to $\{V/S\}$.

Case (ii): Suppose that for all M there exists a T' such that for all $T > T'$, $B_1(T) - W_1(T) > M$.

First, let T_0 be large enough so that for all $T > T_0$, $B_1(T) - W_1(T)$ is larger than twice the largest deviation for a player. For large $T > T_0$, consider any point $x(T) \in (1/T)R(T)$. Without loss of generality, suppose that $x(T)$ lies above the 45° line. We will construct two successive approximations, $y(T)$ and $z(T)$ to $x(T)$. In doing so, we will often refer to Figure A which captures all the essential features of the formal proof.

Step 1: Find the path of points $x(S) \in (1/S)R(S)$, for S between T_0 and T , each of which is an equilibrium continuation payoff of $x(T)$. Now construct an alternative path $y(S)$ that always lies above the 45° line as follows.

(a) If for all S between T_0 and T_1 , $x(S)$ always lies above the 45° line, let $y(S) = x(S)$.

(b) In general, as depicted by the solid dark line in Figure A, the path $x(S)$ could cross the 45° line many times. Let T_1 be the first time it does so and T_2 be the last time it does so. These are also marked in the figure.

If $x(T_0)$ lies below the 45° line (as drawn), $y(S)$ is the reflection of $x(S)$ about the 45° line for S between T_0 and T_1 . If $x(T_0)$ lies above the 45° line, then for S between T_0 and T_1 , $y(S) = x(S)$. After T_1 , $y(S)$ stays close to the 45° line. This can be achieved by alternating between adding the points v^* and v^{**} to $y(T_1)$. These are defined as the first point from $U(A)$ on the efficient frontier of F that lies above the 45° line and its reflection, respectively. (Of course, for some games it could be that $v^* = v^{**}$, that is the symmetric point on the efficient frontier is an element of $U(A)$.) After T_2 , the path $y(S)$ mimicks the $x(S)$ path. The path $y(S)$ is depicted by the dashed line in Figure A.

Step 2: Finally, construct a path $z(S)$ that lies to the right and below $y(S)$. The path $z(S)$ is constructed so that $Z_1(S) - Y_1(S)$ is large enough to prevent any deviation by player 1 in period $S+1$. The path $z(S)$ is depicted by the solid light line in Figure A. It can be shown that $Z(S)$ can be chosen so that it is not arbitrarily far from $Y(S)$. Because of this, for large S , its average, $z(S)$, is close to $y(S)$ and hence also to $x(S)$.

The key observation is this. As constructed, for all S between T_0 and T , $Z(S)$ is far from both extreme points of $R(S)$ in the sense that any element of $U(A)$ may be added to $Z(S)$ to obtain an element of $Q(S+1)$.

This means that between T_0 and T , only efficient combinations must have been used in constructing the path $z(S)$ and hence that for large T , $z(T)$ must be nearly efficient. Since by construction, $z(T)$ is close to $x(T)$, for large T , $x(T)$ must also be nearly efficient. □

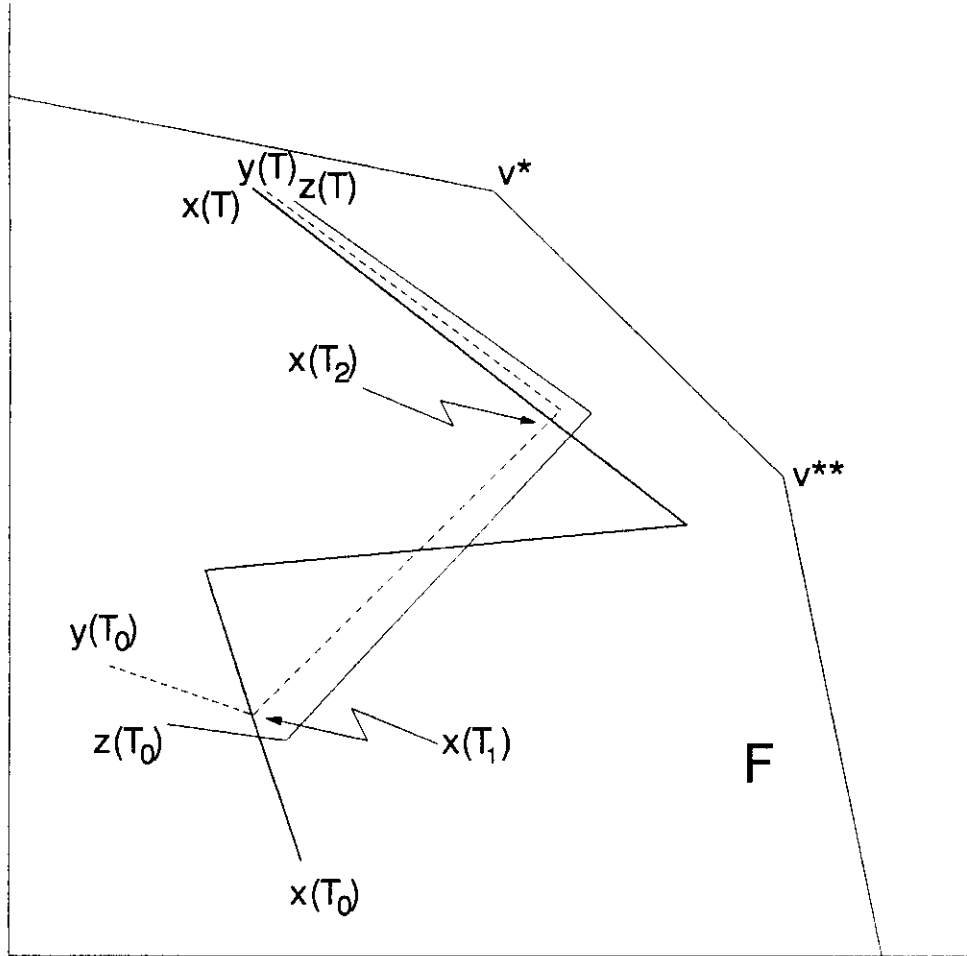


Figure A