GAMES AND ECONOMIC BEHAVIOR 6, 162–177 (1994)

Internally Renegotiation-Proof Equilibrium Sets: Limit Behavior with Low Discounting

DEBRAJ RAY

Boston University, Boston, Massachusetts 02215; and the Indian Statistical Institute, New Delhi 110016, India

Received October 24, 1989

Recent literature in the theory of games addresses the criticism that efficient outcomes in a dynamic game are often supported by punishment paths that do not have the same efficiency property. The outcome of this research is the notion of *renegotiation-proof* equilibria. In this paper, I analyze the notion of renegotiation-proof equilibrium sets that satisfy a natural criterion of internal consistency, one that the earlier notions do not satisfy. I analyze the limit points of such sets as discounting vanishes. The main result states that such limit sets must either be singletons or belong to the Pareto frontier of the convex hull of the feasible set of stage game payoffs. *Journal of Economic Literature* Classification Number: 026. © 1994 Academic Press, Inc.

1. INTRODUCTION

Recent literature in the theory of games addresses the criticism that "collusive" or "efficient" outcomes in a dynamic game are often supported by equilibrium punishments that do not have the same efficiency property.[1] This is a lack of consistency: if players negotiate to select an efficient equilibrium outcome, why should they not *renegotiate* to do so in *all* subgames?

This issue was first addressed by Farrell (1983) and Bernheim and Ray (1985) in independent work. The consideration of consistency leads to a refinement of subgame perfect equilibrium by imposing the same "selection norms" (Pareto efficiency in the class of all allowable equilibria) in

[1] See, for example, Bernheim and Ray (1989), Farrell and Maskin (1989), Pearce (1987), Asheim (1990), Benoit and Krishna (1990), and van Damme (1989).

0899-8256/94 $5.00
Copyright © 1994 by Academic Press, Inc.
All rights of reproduction in any form reserved.

every subgame. The literature has been concerned with repeated games. For *finitely* repeated games, an unambiguous definition can be arrived at through backward induction. However, in *infinitely* repeated games, there remain some serious definitional ambiguities, which we shall now quickly recall.[2]

The definitional issue in an infinitely repeated game has two parts to it, one dealing with "internal" consistency, and the other with "external" consistency. Consider an infinitely repeated game with discounting. Let us see how the issue of internal consistency is dealt with. Bernheim and Ray (1989) and Farrell and Maskin (1989) define a set W of equilibrium payoffs to be *weakly renegotiation-proof* (WRP)[3] if

(a) Each payoff in W is supportable as a subgame perfect equilibrium by using continuation payoffs drawn from W itself; and

(b) No payoff of W Pareto-dominates some other payoff in W.

Conditions (a) and (b) represent minimal notions of internal consistency. Of course, in general, many sets of payoffs satisfy these conditions. For example, the singleton set consisting of the discounted payoff obtained by repeating any equilibrium of the stage game is always WRP.

In this paper, I argue that a complete notion of internal consistency can take us further than this. Specifically, I claim that the notion of a WRP set does *not* adequately capture the requirement of internal consistency, and that a somewhat more stringent test must be passed. This will lead to what I will call an *internally renegotiation-proof* (IRP) set.[4]

Recall the definition of W as a WRP payoff set. Criterion (a) above implies that W is to be considered as a "stationary" theory of supportable payoffs, valid at every date. Criterion (b) imposes dynamic consistency at the collective level: players will not be able to "renegotiate away" from some payoff in W invoking some other better alternative *in W itself*.

Let us pause to reconsider the implications of (a): if W is thought of as the theory at time t, so it is at time $t + 1$. With this in mind, what are the supportable payoffs at time t? Clearly, they are the set of *all* payoffs that can be supported as equilibria with *all* continuation payoffs (at $t + 1$)

[2] See Bernheim and Ray (1989) and Farrell and Maskin (1989) for a discussion of the issues. I also note that the work of Pearce (1987) embodies an entirely different way of studying renegotiation-proofness, one which is not covered by this paper. On this, see Abreu and Pearce (1989) and Abreu *et al.* (1989).

[3] This is the terminology of Farrell and Maskin (1989); I adopt it here.

[4] After the first draft of this paper was written in 1989, I noted that van Damme in his book (van Damme, 1987) briefly considers an alternative notion of renegotiation-proofness which incorporates both the internal consistency notion that I will discuss here as well as a strong external consistency requirement. So IRP is one part of this definition. But the strong implications of the internal consistency requirement were not investigated by him.

restricted to lie in W. Call this set $\theta(W)$. Internal consistency demands, then, that the set of renegotiation-proof payoffs at time t must coincide with the efficient frontier of $\theta(W)$.[5] So this last set must be W itself! But WRP sets do *not*, in general, satisfy this property. I postpone further discussion to Section 2.

We call a set *internally renegotiation-proof* (IRP) if it satisfies the above criterion. Note that an IRP set is always WRP.

We have therefore identified a subset of the class of WRP sets, using the internal consistency requirement alone. However, it remains to be seen if this definitional issue has any practical "bite" in narrowing the set of outcomes. This is what I turn to now.

Consider, for the moment, all the WRP sets. Do they permit us to narrow down the possibilities, relative to the set of all subgame perfect payoffs? The paper by van Damme (1989) suggests that the WRP sets may not be very effective in doing so. He considers the infinitely repeated version of the prisoners' dilemma, and shows that every individually rational, feasible payoff of the repeated game is a member of *some* WRP set, provided that the discount factors of the players are close enough to unity. This implies, at least in the context of an interesting example, that the WRP notion does not narrow down the set of equilibrium outcomes as discounting vanishes.[6]

The behavior of IRP sets is, however, different, and this is the main result of the paper: the class of *limit payoff sets* of *all* sequences of IRP sets (as discounting vanishes) contains *at most* two types of sets, (i) singletons and (ii) subsets of the Pareto frontier of all feasible payoffs. This result stands in marked contrast to van Damme's example for WRP sets.

In Section 2, I motivate and introduce the concept of an IRP set, and study an explicit example. The example shows clearly why a WRP set may not have a natural internal consistency property. On the other hand, the example also reveals that in attempting to fill this gap, an IRP set may not always exist. However, existence is trivially guaranteed when the stage game has a unique Nash equilibrium with the property that it is undominated by any other Nash equilibrium. I also briefly comment on a nonstationary version of the IRP solution concept.

Section 3 contains the main result of my paper, described above, and Section 4 proves this result. Section 5 concludes by stating what I believe to be interesting open questions in this area.

I end the Introduction with two remarks on the main result of the paper.

First, the limit behavior of IRP sets bears a strong similarity to a theorem

[5] Or at least, it must be a subset of this frontier. I do not consider this alternative here.

[6] In general, though, *some* narrowing is achieved—see Farrell and Maskin (1989).

of Benoit and Krishna (1990) about renegotiation-proof equilibria in *finitely* repeated games. Indeed, my proof is guided by the ingenious arguments of these two authors. But there are some differences. Benoit and Krishna consider undiscounted games with the time horizon tending to infinity. There is no difficulty here in defining, unambiguously, *the* set of renegotiation-proof equilibria (Bernheim and Ray, 1989). So one considers here, the limit behavior of a *single* set, while the inherent ambiguities of an infinitely repeated game force us to look at a *class* of sets. In addition, it turns out that there are somewhat different analytical considerations involved in taking the discount factor to one, rather than the time horizon to infinity. However, the overall feature to be stressed is the similarity, not the difference. My result indicates that a natural extension of the finite horizon definition involves *stronger* restrictions than just WRP.

Second, this limit result suggests that the different definitions of external consistency may all yield the same results if there is sufficiently low discounting of the future. The problem is that, in general, there may be cyclical patterns of domination among different WRP (or IRP) sets, so that a maximal element may fail to exist. The various definitions of external consistency may be viewed as a way to get around this problem. However, given the result of this paper, it appears that the different definitions may all yield the same results in the limit, because all sequences of IRP sets that are not tending to singletons are tending to the (unrestricted) Pareto frontier. Cycles of domination will, therefore, disappear as discounting vanishes.

2. Internally Renegotiation-Proof Sets

2.1. *Basic Terminology and Assumptions*

Let $G = (A^1, A^2, \pi^1, \pi^2)$ be a two-person game in normal form.[7] The *action set*[8] of player i is A^i, and his *payoff function* is $\pi^i: A \equiv A^1 \times A^2 \to \mathbb{R}$. We assume that

(A.1) For each $i = 1, 2$, A^i is compact and π^i is continuous.

We normalize payoffs so that $\pi(\alpha) \equiv (\pi^1(a), \pi^2(a)) \geq 0$ for all $a \in A$. So,

[7] We consider only two-person repeated games. Additional conceptual issues, such as renegotiation at the coalitional level, arise with more persons.

[8] The interpretation of A^i that I feel most comfortable with is that of a set of *pure* strategies, though formally A^i may be regarded as a set of *observable* mixed strategies. I have not explored the consequences of extending the analysis to the case where mixtures are used, but only the realized outcomes are observed.

by (A.1), $F \equiv \pi(A)$ is a compact subset of $\mathbb{R}^2_+$. Let F^* be the convex hull of F. Then F^*, too, is compact.

Denote by (G, δ) the infinitely repeated version of G with a common discount factor $\delta \in (0, 1)$ for the two players. That is, if $\langle a_t \rangle_0^\infty \in A^\infty$ is a sequence of action vectors, then the normalized payoff to player i is $(1 - \delta)\ \Sigma_{t=0}^\infty \delta^t \pi^i(a_t)$. Denote by $F(\delta)$ the set of all normalized payoff vectors in the game (G, δ) as we range over all possible sequences in A^∞. It is easy to see that $F(\delta) \in F^*$ for all $\delta \in (0, 1)$.

We omit here the standard definitions of *strategy* and *(subgame) perfect equilibrium* for the game (G, δ).

An element $p \in F(\delta)$ is a *perfect equilibrium payoff* if there exists a perfect equilibrium of (G, δ) with equilibrium payoff p.

2.2. *Internal Consistency*

For any nonempty compact subset B of $\mathbb{R}^2$, define

$$f(B) = \{x \in B \mid \text{there is no } y \in B \text{ with } y \gg x\}. \tag{1}$$

The set of Pareto-efficient points of B (relative to B) is precisely $f(B)$.[9] For any $a \in A$, define

$$d_i(a) \equiv \max_{a_i' \in A^i} [\pi^i(a_i', a_j) - \pi^i(a)], \qquad i = 1, 2. \tag{2}$$

This represents the maximum value of the deviation for agent i from any given action vector $a \in A$.

Let $p \in \mathbb{R}^2$ and a nonempty compact $B \subseteq \mathbb{R}^2$ be given. Say that B *supports* p if there exist $a \in A$ and $\hat{p}, p^1, p^2 \in B$ such that for $i = 1, 2$,

$$p_i = (1 - \delta)\pi^i(a) + \delta \hat{p}_i \tag{3}$$

and

$$d_i(a) \leq \delta(1 - \delta)^{-1}(\hat{p}_i - p_i^i). \tag{4}$$

That is, p is constructing by "rewarding" the players with a continuation payoff of $\hat{p}$, drawn from B, if none deviate, and by reverting to a "punishment" of p^i, also drawn from B, should player i unilaterally deviate. Call the collection $(a, \hat{p}, p^1, p^2)$ the *supporter* of p. Whenever we write a

[9] The notation $y \gg x$ means that $y_i > x_i$ for $i = 1, 2$. So we are using the concept of weak Pareto efficiency here.

supporter, it will be taken for granted that the first entry is an action vector, the second a continuation payoff if there is no deviation, the third a continuation payoff when player 1 unilaterally deviates, and the fourth a continuation payoff if player 2 deviates.

Define

$$\theta(B) \equiv \{p \in \mathbb{R}^2 \mid B \text{ supports } p\} \tag{5}$$

A compact set $P \subseteq \mathbb{R}^2$ is internally renegotiation-proof (IRP)[10] if

$$P = f(\theta(P)). \tag{6}$$

I reiterate that this definition represents a more stringent requirement on the payoff sets than that implicit in a WRP set. As I have argued in the Introduction (and this will be made even clearer in Section 2.3), an IRP set captures the notion of internal consistency to a fuller extent than a WRP set. For comparison, we note that a WRP set P is characterized by

$$P = f(B) \qquad \text{for some } B \subseteq \theta(P)$$

It is easy to verify that WRP sets (and therefore all IRP sets) are indeed sets of perfect equilibrium payoffs, so that definition (6) is a refinement of the notion of subgame perfection.

Observe, moreover, that definition (6) is the natural extension of the unambiguous notion of renegotiation-proofness in finite horizon games (see Bernheim and Ray, 1985, 1989, and Benoit and Krishna, 1990). There, one simply starts with the Nash payoffs at the last date and works backwards recursively using the composed map $f(\theta(\cdot))$. So there is nothing particularly new or novel about the definition of IRP—it is *the* natural extension of the corresponding definition in the finite horizon case.[11]

[10] We are restricting ourselves to the consideration of compact sets of solutions. Given (A.1), this is hardly a serious limitation.

[11] However, one must be very careful in linking limits of IRP sets obtained in the finite horizon case by taking the time horizon to infinity, and the IRP sets that we have directly defined for the infinitely repeated game. For one thing, it is not surprising that there are IRP sets in the infinite game that cannot be achieved as limits of finite horizon IRP sets. But this failure of lower hemicontinuity as one moves from the finite to the infinite horizon is well known. What is more disturbing, however, is the possibility that the Hausdorff limits of finite horizon IRP sets may *not* be infinite-horizon IRP. This is related to the lack of continuity of the map $f(\theta(\cdot))$ in the Hausdorff metric and via this route to the nonexistence problem for IRP sets—see Section 2.3.

TABLE I

	A	B	C
a	10, 10	$-K$, $-K$	$-K$, $10 + \varepsilon$
b	$-K$, $-K$	$-K$, $-K$	1,2*
c	$10 + \varepsilon$, $-K$	2,1*	$-K$, $-K$

2.3. *An Example*

The following example is designed to make two points:

(1) WRP sets do not, in general, satisfy a natural "internal consistency" property.

(2) IRP sets, which get around the problem in (1), have a different drawback: such sets do not always exist.

Consider the bimatrix game depicted in Table I. There are two pure strategy Nash equilibria in the one-shot game: they are (b, C) and (c, B) and are indicated by (*) in Table I. Consider, for a given common discount factor δ, the set W of payoffs generated by *all* outcome paths which have as entries only (b, C) and (c, B). The set W is WRP. Now suppose both players hold W as a "theory" of the *set* of achievable present-value payoffs from *each* date onward. I claim that this theory lacks internal consistency, provided $\varepsilon < \delta^2$. For if W is believed to be achievable from date $t = 1$, the action vector (a, A) is supportable at date $t = 0$ using continuation payoffs in every subgame that come *only* from W (the condition $\varepsilon < \delta^2$ is sufficient for this). But then the Pareto frontier of the set of achievable payoffs at date $t = 0$ is *not* W, which is a contradiction.

Internal consistency is, therefore, only achieved when this recursive calculation yields, "today," precisely the same set of payoffs that are anticipated in the future.[12] This is the idea behind the definition of an IRP set.

Unfortunately, IRP sets may not always exist, as the very same game above demonstrates. I show this by considering only pure strategies and then indicate how the argument may be extended to mixed strategies.

Choose ε, δ, and K such that $\varepsilon > 0$, $K \geq 0$, $\varepsilon < \delta^2$, and $\delta < 1/(11 + K)$. I first claim that in any perfect equilibrium, cells apart from (a, A) and the two one-shot Nash equilibria will never be observed. The one-shot gain from deviation available to *some* player in *any* other cell is at

[12] This is only true for time-stationary solution concepts, of which the WRP notion is certainly one. However, there is no *a priori* reason why the solution concept must be time-stationary. See Section 2.4 for a brief discussion.

least 1. The *maximum* punishment (discounted to the current date) that can be inflicted on that player is clearly bounded above by $\delta(10 + K)/(1 - \delta)$. Because $\delta < 1/(11 + K)$, this cannot serve as an adequate deterrent.

Next I claim that if P is an IRP set, and if for *some* equilibrium supporting *some* payoff in P, the cell (a, A) is *ever* played, then $P = \{(10, 10)\}$. Suppose that (a, A) is played for some equilibrium supporting some payoff in P. Consider the continuation payoff $p \in P$ which involves the play of (a, A) in the very first period. Then $p \geq (10, 10)(1 - \delta)$. Now suppose that there is some other $q' \in P$ which involves a play of, say, (c, B). Then there is $q \in P$ which involves a play of (c, B) in the very first period. But then

$$q \leq (2, 1)(1 - \delta) + \delta(10, 10).$$

Now it is easy to check, using the above inequalities and the assumption $\delta < 1/(11 + K)$, that $q \ll p$. But this contradicts the definition of IRP. So $P = \{(10, 10)\}$. But then P cannot be IRP, for the payoff vector (10, 10) cannot be supported by $P = \{(10, 10)\}$.

So if P is an IRP set, the *only* outcome paths corresponding to *any* payoff vector $p \in P$ must involve combinations of the two Nash equilibria. It is then easy to see that P must contain *all* payoff vectors attainable in this way, i.e., $P = W$. But now (a, A) *is* supportable! (We use the same argument that we used to demonstrate the inconsistency of W.)

This proves that no IRP set exists for $\varepsilon^2 < \delta < 1/(11 + K)$.

With mixed strategies, one only needs to take K large. Then all play will be "almost" pure and restricted to the three cells (a, A), (c, B), and (b, C) with high probability. From this point on, a minor variant of the above argument applies.

Note that in this example, an IRP set *does* exist for δ large enough. Consider pure strategies. One can prove that if δ is close enough to 1, there exists a unique IRP set. This sequence (in δ) of IRP sets converges to the subset of the Pareto frontier of F^* formed by the individually rational payoffs (i.e., $p \geq (1, 1)$).

Whether this feature is general, i.e., *whether there exists, always, an IRP set for δ large enough* is an interesting open question.

Note that the existence of an IRP set is trivially guaranteed if the Pareto frontier of the NE payoffs of the stage game is a singleton.

2.4. *IRP Solutions*

I mention here, in passing, a weakening of the concept of an IRP set that may merit further investigation. The implicit assumption that underlies the notion of WRP and IRP sets is that the theory of equilibrium payoffs is

invariant with respect to time. There is no reason why this should be the case, even though the underlying repeated game is fully stationary.

Let $\langle P_t \rangle_{t=0}^{\infty}$ be a sequence of subsets of $\mathbb{R}^2$. Say that $\langle P_t \rangle_{t=0}^{\infty}$ is an *IRP solution* if for all t,

$$P_t = f(\theta(P_{t+1})).$$

Note that the definition is internally consistent and yet allows for a time-dependent theory. In particular, if players commonly adhere to the belief that P_{t+1} is the set of finite horizon payoffs available from time $t + 1$, then the twin requirements of supportability and no Pareto dominance yield *exactly* the set P_t at time t.

An IRP set, as defined before, is then any set that corresponds to a time-stationary IRP solution.

IRP solutions may be of interest in studying cyclical behavior, e.g., the periodic breakdown of "cooperation," especially in cases where IRP sets do not exist. But a detailed examination of such solutions will take us far afield of my main objective, to which I now turn.

3. A Limit Theorem for IRP Sets

For the game (G, δ), denote by $\mathcal{P}(\delta)$ the collection of all IRP sets. We are interested in the members of the following collection:

$$\mathcal{P} \equiv \{\text{all limit points of sequences in } \mathcal{P}(\delta), \quad \text{as } \delta \to 1\}.$$

(A sequence P^k of compact subsets of $\mathbb{R}^2$ will be said to *converge* to some set P if the Hausdorff distance (relative to Euclidean distance) between P^k and P tends to zero as $k \to \infty$.)

So $\mathcal{P}$ is the collection of all possible sets to which sequences of IRP sets can converge as discounting vanishes. It turns out that $\mathcal{P}$ has the following property:

THEOREM 1. *Each element of $\mathcal{P}$ is either a singleton or a subset of $f(F^*)$, the efficient frontier of F^*.*

This is the analogue of the "folk theorem" for internally renegotiation-proof equilibrium sets. We reiterate two points already made in the Introduction. First, this result stands in sharp contrast to the wide range of limiting behavior that WRP sets appear to predict, at least in examples (van Damme, 1989). Second, our result complements the Benoit–Krishna theorem for undiscounted finitely repeated games (Benoit and Krishna, 1990). Of course, the setting is different from that of the latter paper. For

instance, if G has a unique Pareto-undominated stage equilibrium, the unique renegotiation-proof equilibrium set in finite repetitions of the game is obtained by simply repeating the stage equilibrium. Nevertheless, in the infinitely repeated game, there may still be more than one IRP set.

In this context, I should note that *both* possibilities allowed for in Theorem 1 may actually occur in the same game. An example is the prisoners' dilemma. There, it is possible to show that there are *two* limit IRP sets. One is formed by the payoff vector yielded by the one-shot Nash equilibrium. This is not on the Pareto frontier, and it is a singleton. The other is precisely the *entire* individually rational Pareto frontier of F^*.

4. Proof of the Theorem

For each $i = 1, 2$, define $\bar{p}^i \in F^*$ by first maximizing p_i over $p \in F^*$ and then minimizing p_j over the set of maximizers. Observe that $\bar{p}^i$ is unique, and in fact that $\bar{p}^i \in F$.

Also, define $D \equiv \max_{i=1,2}\{\max_a d_i(a)\}$. Clearly, $0 \leq D < \infty$.

Pick any $P \in \mathcal{P}$. Suppose for some pair p', $p'' \in P$ and for some $i \in \{1, 2\}$, we have $p_i' < p_i''$. Define the following subset of $\mathbb{R}^2$:

$$L(i, p', p'') \equiv \{p \mid \exists \lambda \in [0, 1] \text{ s.t. } p = (1-\lambda)\bar{p}^j + \lambda p'', j \neq i \text{ and } p_i > p_i'\} \tag{7}$$

Lemma 1. *For each $p \in L(i, p', p'')$, there exists $q \in F^*$ such that $q \geq p$ and $q \in P$.*

Proof. Without loss of generality take $i = 1$. Pick any $p \in L(1, p', p'')$. Then there is $\lambda \in [0, 1]$ such that

$$p = (1-\lambda)\bar{p}^2 + \lambda p'' \tag{8}$$

$$p_1 > p_1' + 2\beta \qquad \text{for some } \beta > 0. \tag{9}$$

There exists a sequence $\delta^k \to 1$ and $P(\delta^k) \in \mathcal{P}(\delta^k)$ such that $P(\delta^k) \to P$. For convenience, drop the "k": $P(\delta) \to P$ as $\delta \to 1$. By a property of convergence in the Hausdorff metric, there exist $p'(\delta)$, $p''(\delta) \in P(\delta)$ with $p'(\delta) \to p'$ and $p''(\delta) \to p''(\delta) \to p''$ as $\delta \to 1$.

Recall that $\bar{p}^2$ (defined above) is an element of F. So there is $\bar{a} \in A$ such that $\pi(\bar{a}) = \bar{p}^2$. For each δ in the sequence, pick some nonnegative integer $T(\delta)$, and define $p(\delta)$ by

$$p(\delta) = (1-\delta)\sum_{t=0}^{T(\delta)} \delta^t \pi(\bar{a}) + \delta^{T(\delta)+1} p''(\delta). \tag{10}$$

We can (and will) choose the sequence $T(\delta)$ so that $p(\delta) \to p$ as $\delta \to 1$. To see that we can do this, note that

$$p(\delta) = (1 - \delta^{T(\delta)+1})\overline{p}^2 + \delta^{T(\delta)+1}p''(\delta) \tag{11}$$

and also that $p''(\delta) \to p''$ as $\delta \to 1$. So it suffices to choose $T(\delta)$ for each δ such that $\delta^{T(\delta)+1} \to \lambda$ as $\delta \to 1$. This is easy to do.

Because $p_1 > p_1'$ and $p_1'' > p_1'$, it is possible to choose $\underline{\delta} \in (0, 1)$ such that for all $\delta \geq \underline{\delta}$,

$$\min\{p_1(\delta) - p_1'(\delta); p_1''(\delta) - p_1'(\delta)\} \geq \beta > 0. \tag{12}$$

Define, for each $t = 1, \ldots, T(\delta) + 1$,

$$\hat{p}(\delta, t) \equiv (1 - \delta^t)\overline{p}^2 + \delta^t p''(\delta). \tag{13}$$

Note that $\hat{p}(\delta, T(\delta) + 1) = p(\delta)$. I claim that if $\delta \geq \underline{\delta}$, then

$$\hat{p}_1(\delta, t) - p_1'(\delta) \geq \beta \qquad \text{for all } t = 1, \ldots, T(\delta) + 1. \tag{14}$$

To establish this claim, note that by combining (13) and (11) we have, for $t = 1, \ldots, T(\delta) + 1$,

$$\hat{p}(\delta, t) = \frac{1 - \delta^t}{1 - \delta^{T(\delta)+1}} p(\delta) + \left[1 - \frac{1 - \delta^t}{1 - \delta^{T(\delta)+1}}\right] p''(\delta). \tag{15}$$

Observing that $1 - \delta^t \leq 1 - \delta^{T(\delta)+1}$, we see from (15) that

$$\hat{p}_1(\delta, t) - p_1'(\delta) \geq \min\{p_1(\delta) - p_1'(\delta); p_1''(\delta) - p_1'(\delta)\}$$

and using (12), (14) is established.

Finally, pick $\overline{\delta} \in (\underline{\delta}, 1)$ such that

$$D < \beta(1 - \delta)^{-1} \qquad \text{for } \delta \geq \overline{\delta}. \tag{16}$$

We now complete the proof by showing that for each $\delta \geq \overline{\delta}$, there exists $q(\delta) \gg p(\delta)$ such that $q(\delta) \in P(\delta)$. If this is true, then we are done. For pick any limit point q of $q(\delta)$. Then $q \in P$, because $P(\delta) \to P$. Moreover, $q \geq p$, as desired.

Pick any $\delta \geq \overline{\delta}$. Observe that $P(\delta)$ supports $\hat{p}(\delta, 1)$, using the supporter $(\overline{a}, p''(\delta), p'(\delta), p''(\delta))$. One can easily verify that this is indeed a supporter,

using (12), (16), and the fact that $\bar{a}$ maximizes 2's one-period payoff. So there is $\hat{q}(\delta, 1) \geq \hat{p}(\delta, 1)$ such that $\hat{q}(\delta, 1) \in P(\delta)$. Recursively, for $t = 1, \ldots, T(\delta) + 1$, suppose that for some t we have defined $\hat{q}(\delta, t) \geq \hat{p}(\delta, t)$ with $\hat{q}(\delta, t) \in P(\delta)$. For $t + 1$, define

$$q'(\delta, t + 1) \equiv (1 - \delta)\pi(\bar{a}) + \delta\hat{q}(\delta, t). \tag{17}$$

It is easy to check that $P(\delta)$ supports $q'(\delta, t + 1)$ via the supporter $(\bar{a}, \hat{q}(\delta, t), p'(\delta), \hat{q}(\delta, t))$, by using (14), (16), the definition of $\bar{a}$, and the fact that $\hat{q}(\delta, t) \geq \hat{p}(\delta, t)$. So there exists $\hat{q}(\delta, t + 1) \in P(\delta)$ with $\hat{q}(\delta, t + 1) \geq q'(\delta, t + 1) \geq \hat{p}(\delta, t + 1)$ (to check the last inequality simply use (13), (17), and $\hat{q}(\delta, t) \geq \hat{p}(\delta, t)$). This completes the recursive definition. Finally, defining $q(\delta) \equiv \hat{q}(\delta, T(\delta) + 1)$, the proof is complete. ■

To proceed further, define for $i = 1, 2$, $p^i \in P$ by minimizing p_i over $p \in P$, and then maximizing p_j, $j \neq i$ over the set of minimizers. Because P is the Hausdorff limit of compact sets in an ambient compact space (F^*), P is compact too and so p^i is well (and uniquely) defined for each i.

LEMMA 2. *Suppose that P is not a singleton and that P is not a subset of $f(F^*)$. Then there exist $p \in P$, $q \in F^*$ such that $q \gg p$ and for $i = 1, 2$,*

$$p_i > p_i^i. \tag{18}$$

Proof. Suppose the conditions of the lemma hold; then for some $\tilde{p} \in P$ and $q \in F^*$, we have $q \gg \tilde{p}$. Of course, $\tilde{p}_i \geq p_i^i$ for $i = 1, 2$. First, we claim that

$$\bar{p}_i^i > p_i^i \tag{19}$$

for $i = 1, 2$. Suppose not. Then, for some i, equality holds. But then for all $p \in P$, we have $p_i = \bar{p}_i^i$. But this would mean that $P \subseteq f(F^*)$, a contradiction.

Now, if $p \equiv \tilde{p}$ satisfies (18), we are done. Otherwise, $\tilde{p}_i = p_i^i$ for some i. In this case we claim that in fact, $\tilde{p} = p^i$. Suppose not; then by the definition of p^i, we have $\tilde{p}_j < p_j^i$ for $j \neq i$. Now consider any $p \in L(j, \tilde{p}, p^i)$. Using (19) and the definition of L, $p \gg \tilde{p}$. By Lemma 1, there is $q \geq p$ with $q \in P$. But then $q \gg \tilde{p}$, contradicting our supposition that $\tilde{p} \in P$.

So, if (18) fails then $\tilde{p} = p^i$ for some i. Say $\tilde{p} = p^2$. Now consider some $p \in L(1, p^1, p^2)$. Because P is not a singleton and because (19) holds for $i = 2$, we have (18) holding for each such p. Moreover, p can be chosen arbitrarily close to p^2, so that $q \gg p$. This p satisfies all the requirements of the lemma, and the proof is complete. ■

Now we prove the theorem. Suppose, on the contrary, that P is not a singleton but that P is not a subset of $f(F^*)$. Then, by Lemma 2, there exist $q' \in F^*$ and $p \in P$ such that $q' \gg p$ and

$$\beta \equiv \frac{\min\{p_1 - p_1^1; p_2 - p_2^2\}}{2} > 0. \tag{20}$$

Because $q' \in$ convex hull (F), there exists $q \in F^*$ ("close to" q) so that $q \gg p$ and there are a finite number of action vectors $a_0, \ldots, a_M$ with

$$\frac{1}{M+1}\sum_{t=0}^{M} \pi(a_t) = q. \tag{21}$$

Because $q \gg p$, there is a scalar $\alpha > 1$ such that

$$q \gg \alpha p. \tag{22}$$

By the definition of P, there is a sequence $\delta \to 1$ and corresponding $P(\delta) \in \mathcal{P}(\delta)$ such that $P(\delta) \to P$. So there are sequences $p(\delta), p^1(\delta), p^2(\delta) \in P(\delta)$ converging, as $\delta \to 1$, to p, p^1, and p^2, respectively.

We claim that there exists $\delta_1 \in (0, 1)$ such that for all $\delta \geq \delta_1$,

$$q^0(\delta) \equiv (1-\delta)\sum_{t=0}^{M} \delta^t \pi(a_t) + \delta^{M+1} p(\delta) \gg p(\delta). \tag{23}$$

To see this, note that $p(\delta) \to p$ as $\delta \to 1$, so, using (22), there exists $\underline{\delta} \in (0, 1)$ and $\alpha' > 1$ such that for all $\delta \geq \underline{\delta}$,

$$\frac{1}{M+1}\sum_{t=0}^{M} \pi(a_t) \gg \alpha' p(\delta). \tag{24}$$

For all such δ, we have, remembering that $\pi(a) \geq 0$ by normalization,

$$\begin{aligned} q^0(\delta) &\geq (1+M)(1-\delta)\delta^M \left[\frac{\sum_{t=0}^{M} \pi(a_t)}{M+1}\right] + \delta^{M+1} p(\delta) \\ &\gg [(1+M)(1-\delta)\delta^M \alpha' + \delta^{M+1}]\, p(\delta). \end{aligned} \tag{25}$$

Observe that there exists $\delta_1 \in (\underline{\delta}, 1)$ such that for all $\delta \geq \delta_1$, $g(\delta) \equiv (1 + M)(1 - \delta)\delta^M \alpha' + \delta^{M+1} > 1$. To see this, note that $g(\delta)$ is differentiable

and that $g(1) = 1$. Moreover the derivative of $g(\cdot)$ evaluated at 1 is $g'(1) = (M + 1) - \alpha'(M + 1) > 0$. So we have established (23).

Next, pick $\delta_2 \in (\delta_1, 1)$ such that for all $\delta \geq \delta_2$, the collection

$$q^s(\delta) \equiv (1 - \delta) \sum_{t=s}^{M} \pi(a_t) + \delta^{M+1-s} p(\delta) \tag{26}$$

for $s = 0, \ldots, M$ has the property that

$$\min_s \{q_1^s(\delta) - p_1^1(\delta); q_2^s(\delta) - p_2^2(\delta)\} > \beta \tag{27}$$

and

$$\min \{p_1(\delta) - p_1^1(\delta); p_2(\delta) - p_2^2(\delta)\} > \beta. \tag{28}$$

This is possible because $q^s(\delta) \to p$ as $\delta \to 1$ (for each s), $p^1(\delta) \to p^1$, $p^2(\delta) \to p^2$, and (20) holds.

Finally, choose $\underline{\delta} \in (\delta_2, 1)$ such that for all $\geq \underline{\delta}$,

$$D < \beta(1 - \delta)^{-1}. \tag{29}$$

Consider any δ in $[\underline{\delta}, 1)$. We are going to show that there exists $r \geq q^0(\delta)$ such that $r \in P(\delta)$. This will yield a contradiction, because by (23), $q^0(\delta) \gg p(\delta)$ and $p(\delta) \in P(\delta)$.

We proceed recursively. Consider, first, $q^M(\delta)$. It is easy to check, using (28) and (29), that $P(\delta)$ supports $q^M(\delta)$ via the supporter $(a_M, p(\delta), p^1(\delta), p^2(\delta))$. So there is $r^M(\delta) \in F^*$ such that $r^M(\delta) \geq q^M(\delta)$, and $r^M(\delta) \in P(\delta)$. Recursively, suppose that for some $s + 1$ $(s = 0, \ldots, M - 1)$, we have defined $r^{s+1}(\delta) \in P(\delta)$ with $r^{s+1}(\delta) \geq q^{s+1}(\delta)$. Define

$$\tilde{q}^s(\delta) \equiv (1 - \delta)\pi(a_s) + \delta r^{s+1}(\delta). \tag{30}$$

Then, using (27), (29), and $r^{s+1}(\delta) \geq q^{s+1}(\delta)$, we see that $\tilde{q}^s(\delta) \in \theta(P(\delta))$—use the supporter $(a_s, r^{s+1}(\delta), p^1(\delta), p^2(\delta))$.

Therefore, there exists $r^s(\delta) \geq \tilde{q}^s(\delta)$ with $r^s(\delta) \in P(\delta)$. Also, recalling that $r^{s+1}(\delta) \geq q^{s+1}(\delta)$, we have, combining (26) and (30), that $\tilde{q}^s(\delta) \geq q^s(\delta)$. So $r^s(\delta) \geq q^s(\delta)$, and the recursion is complete.

Finally, observe that $r^0(\delta) \in P(\delta)$ and $r^0(\delta) \gg q^0(\delta) \gg p(\delta)$, so that we have a contradiction as desired.

5. Further Directions

This paper leaves a number of issues unresolved.

(1) First, there is the question of the existence of an IRP set. Of course, existence is a trivial issue if the stage game has a Nash equilibrium that is the unique one which is not Pareto-dominated by any other Nash equilibrium. Otherwise, it is problematic. I conjecture that IRP sets always exist if the discount factor is close enough to unity.

(2) Next, there is the question of characterizing the class of games that admit a limit IRP set on the Pareto frontier of F^*. Evans and Maskin (1989) proved that for generic finite action games, there is always some payoff on the frontier that is supportable as a WRP equilibrium. Is a corresponding result true for IRP sets?

(3) Farrell and Maskin (1989, Theorem 3) describe the subset of points of the Pareto frontier of F^* (again, for finite action games) that can be supported as WRP equilibria. It is easy to see that a limit IRP set, whenever it exists, must be a subset of this set. But does it coincide with this set? Ongoing research suggests that the answer is, in general, no, but more work is needed to conclusively settle the question.

Acknowledgments

I am most grateful to Vijay Krishna for encouraging me to work on this problem, and to an Associate Editor and a referee of this journal for suggestions concerning a revision. My debt to Benoit and Krishna [1990] is obvious. I also record my gratitude to Doug Bernheim for many helpful discussions on this subject.

References

Abreu, D., and Pearce, D. (1989). "A Perspective on Renegotiation in Repeated Games," Hoover Institution Working Papers in Economics No. E-89-31.

Abreu, D., Pearce, D., and Stacchetti, E. (1989). "Renegotiation and Symmetry in Repeated Games," mimeograph.

Asheim, G. (1990). "Extending Renegotiation-Proofness to Infinite Horizon Games," *Games Econ. Behav.*, forthcoming.

Benoit, J., and Krishna, V. (1990). "Renegotiation in Finitely Repeated Games," mimeograph.

Bernheim, D., and Ray, D. (1985). "Pareto-Perfect Nash Equilibria," mimeograph. Stanford Universty.

Bernheim, D., and Ray, D. (1989). "Collective Dynamic Consistency in Repeated Games," *Games Econ. Behav.* **1**, 295–326.

Evans, R., and Maskin, E. (1989). "Efficient Renegotiation-Proof Equilibria in Repeated Games," *Games Econ. Behav.* **1**, 361–369.

FARRELL, J. (1983). "Renegotiation-Proof Equilibrium in Repeated Games," mimeograph.

FARRELL, J., AND MASKIN, E. (1989). "Renegotiation in Repeated Games," *Games Econ. Behav.* **1,** 361–369.

PEARCE, D. (1987). "Renegotiation-proof Equilibria: Collective Rationality and Intertemporal Cooperation," mimeograph.

VAN DAMME, E. (1987). *Stability and Perfection of Nash Equilibria*. Springer-Verlag, Berlin, New York.

VAN DAMME, E. (1989). "Renegotiation-Proof Equilibria in the Prisoners' Dilemma." *J. Econ. Theory* **47,** 206–217.