Review of Economic Studies (2002) **69**, 339–356 0034-6527/02/00130339$02.00
© 2002 The Review of Economic Studies Limited

Gradualism and Irreversibility

BEN LOCKWOOD
University of Warwick

and

JONATHAN P. THOMAS
University of St Andrews

First version received May 1999; *final version accepted June* 2001 (*Eds.*)

This paper considers a class of two-player dynamic games in which each player controls a one-dimensional action variable, interpreted as a level of cooperation. The dynamics are due to an irreversibility constraint: neither player can ever reduce his cooperation level. Payoffs are decreasing in one's own action, increasing in one's opponent's action. We characterize efficient symmetric equilibrium action paths; actions rise gradually over time and converge, when payoffs are smooth, to a level strictly below the one-shot efficient level, no matter how little discounting takes place. The analysis is extended to incorporate sequential moves and asymmetric equilibria.

1. INTRODUCTION

We consider a model in which, in every period, there is a Prisoner's Dilemma structure: agents have some mutual interest in cooperating, despite the fact that it is not in any agent's individual interest to cooperate. We suppose that this situation is repeated over time, and, crucially, subject to irreversibility, in the sense that an agent can never reduce her level of cooperation, only increase it or leave it unchanged. In this setting, irreversibility has two opposing effects. First, it aids cooperation, through making deviations in the form of reduced cooperation impossible. Second, it limits the ability of agents to punish a deviator. We consider the complex interplay of these two forces.

The key role of irreversibility in affecting cooperation can be explained more precisely as follows. In the above model, suppose that every player has a scalar action variable, which we interpret as a level of cooperation. We say that *partial cooperation* occurs in some time period if some player chooses a level of this action variable higher than the stage-game Nash equilibrium level. Due to the Prisoner's Dilemma structure, the latter is the smallest feasible value of the action variable. *Full cooperation* occurs when both players choose a level of this action variable that maximizes the joint payoff of the players.[1] In general, partial or full cooperation in any time-period can only be achieved if deviation by any agent can be punished by the other agents in some way.

Without reversibility, the above model is just a version of a repeated Prisoner's Dilemma, and in that case, it is well-known that the most effective (credible) punishments take the form of "sticks", *i.e. reductions* in cooperation back to the stage-game Nash equilibrium, temporarily or permanently. With irreversibility, such punishments are no longer feasible. Instead, deviators can only be credibly punished by withdrawal of "carrots", that is, withdrawal of promised *increases* in cooperation in future. It follows immediately from this fact that irreversibility causes *gradualism*:

1. Our model is symmetric, *i.e.* players have identical per-period payoffs given a permutation of their actions, so we assume that both players choose the same level of the action variable in full cooperation (see Section 2 for details).

any (subgame-perfect) equilibrium sequence of actions cannot involve an immediate move to full cooperation, since there can be no carrot to enforce such a move.

Our first contribution is to refine and extend this basic insight. First, we show that any (subgame-perfect) symmetric[2] equilibrium sequence of actions involving partial cooperation must have the level of cooperation always rising, that is, not attaining its limit value, which in turn cannot exceed full cooperation. We focus on the symmetric equilibrium sequence that is efficient within the set of all symmetric equilibria, and refer to it as the *efficient symmetric equilibrium path*. A key question then is: to what value does this efficient symmetric equilibrium path converge? It turns out that if payoffs are smooth (differentiable) functions of actions, convergence will be to a level *strictly below* the full cooperation level, *no matter how patient* agents are.[3] For the case where payoffs are linear up to some joint cooperation level, and constant or decreasing thereafter (the linear kinked case), the results are different—above some critical discount factor, equilibrium cooperation can converge asymptotically to the fully efficient level. Below this critical discount factor, no cooperation *at all* is possible. So, even for discount factors close to one, the efficient path in our model is quite different from that in the same model without irreversibility: in the latter case, above some critical discount factor, the fully efficient cooperation level can be attained exactly and immediately as an equilibrium outcome.

The reason for the asymptotic inefficiency in the smooth payoff case is that close to full cooperation, returns from additional mutual cooperation are *second-order*, whereas the benefits to deviation (not increasing cooperation when the equilibrium path calls for it) remain *first-order*. The future gains from sticking to an increasing mutually cooperative path will be insufficient to offset the temptation to deviate. It follows that it will be impossible to sustain equilibrium paths that become too close to full cooperation.

In many economic applications, irreversibility arises more naturally when the level of "cooperation" is a stock variable which may benefit both players, and it is (costly) incremental investment in cooperation that is non-negative, implying the stock variable is irreversible. Therefore, in Section 4, we present an "adjustment cost" model with these features, and show that it can be reformulated so that it is a special case of our base model. We then apply the adjustment cost model to study sequential public good contribution games (Admati and Perry (1991), Marx and Matthews (1998)).

In Section 5 we relax the restriction to symmetric equilibrium paths. Moreover, the base model also assumes that the (two) players move simultaneously. In Section 6 we show that if players are constrained to move sequentially, the equilibrium payoffs in this game are a subset of those in the simultaneous move game, but that as discounting goes to zero, the efficient symmetric payoff in the symmetric move game can be arbitrarily closely approximated by equilibrium payoffs in the sequential game, so that asymptotically, the order of moves has little effect on achievable payoffs. Further extensions are analysed in Section 7, where we briefly discuss partial reversibility (considered in detail in Lockwood and Thomas (1999)).

Turning to related literature, there is a small literature on games with the features we consider here. The basic insight that irreversibility implies gradualism is not entirely new; perhaps Schelling (1960, p. 45) was the first to make the point. Admati and Perry (1991) and Marx and Matthews (1998) present equilibria of dynamic voluntary contribution games which exhibit gradualism, although in the first paper, gradualism is partly due to the strict convexity of the player's cost functions. The "level" of cooperation in these models is the sum of an

2. Due to the symmetry of the model, we focus initially on *symmetric* equilibrium sequences, where in equilibrium both players choose the same level of the cooperation variable at all dates.

3. Despite this result, inefficiency disappears in the limit as players become patient in the sense that the limit value of the efficient symmetric equilibrium path, and player payoffs from this path, both converge to fully efficient levels as discounting goes to zero.

individual's past contributions, and this is irreversible. Gale (2000) has considered a quite general class of sequential move games with irreversible actions which he calls "monotone games". For games with "positive spillovers", which include the class of games considered here, he characterizes long-run efficient outcomes when there is no discounting. In particular, his results imply that in a sequential-move version of our model without discounting, first-best outcomes are eventually attainable.[4] Finally, Salant and Woroch (1992) study a game between a regulator and a firm, where the former can set a price ceiling, and latter can make irreversible investments to a (depreciating) capital stock. In their model, the efficient level of capital is never achieved, but is approached asymptotically at a rate that depends on the discount factor. However, there are also convex costs of adjustment of capital in their model. So, to the best of our knowledge, our paper provides the first *general* treatment of a class of games *with discounting* in which gradualism in cooperation due to irreversibility arises naturally (see however Compte and Jehiel (1995) for a related idea in a bargaining context).

Of the papers just mentioned, possibly the closest is Marx and Matthews (1998). The relationship between the two papers is as follows. The two papers consider quite different models, although there is some overlap. Marx and Matthews consider a wide class of voluntary contribution games, where a number of players simultaneously make contributions to a public project over T periods, and where T may be finite or infinite. Each player gets a payoff that is linear in the sum of cumulative contributions, plus possibly a "bonus" when the project is completed. One case of their model (T infinite, two players, no bonus) can be reformulated as an "adjustment cost" variant of our model with linear kinked payoffs, as argued in detail in Section 4. In this version of their model, Marx and Matthews construct a subgame-perfect equilibrium which is approximately efficient when discounting is negligible,[5] whereas we are able to characterise efficient subgame-perfect equilibria for *any* fixed value of the discount factor. Specifically, our results show that in their model, the equilibrium with completion which they construct is in fact efficient for *any* discount factor above a critical value, and conversely when the discount factor is below the critical value, there are *no* contributions made in the efficient equilibrium.

We see our model as being applicable to a variety of situations. For example, in an earlier version of this paper (Lockwood and Thomas (1999)) we applied our "adjustment cost" model to capacity reduction in a declining industry, under the assumption usually made in this literature that capacity reductions are irreversible (Ghemawat and Nalebuff (1990)). While it may be desirable to move immediately to reduce capacity in an industry to some level, this is not an equilibrium because either firm would prefer to have the other reduce capacity while retaining its own capacity. In other situations, increases in co-operation may not be irreversible, but very costly to reverse, in which case our basic results about gradualism continue to hold (see Section 7). Disarmament between two warring parties is one example—here cooperation would be measured by the extent of disarmament, which may be very difficult to reverse, as complex weapons, once destroyed, may be difficult to rebuild. A further application is to environmental problems. Environmental cooperation may take the form of installation of costly abatement technology. Once installed, this technology may be very expensive to replace with a "dirtier" technology, *e.g.* conversion of automobiles to unleaded fuel would be expensive to reverse.

4. In fact, for the two-person case, Gale (2000) shows that any individually-rational point on the Pareto-frontier can be eventually attained. Gale's set-up is considerably more general than ours, and in particular, allows for the possibility that a player's payoff may be increasing in his or her own cooperation level (on completion of the project in the public good model). The lack of this feature here allows us to obtain sharp results without the need to impose no discounting.

5. Corollary 3(ii), Marx and Matthews (1998). Note that their results are stated for $n > 2$ players also, and hold even when only the sum of past contributions is observable.

Consequently it will again be difficult to punish deviants by reversing the investment. Similarly, destruction of capital (*e.g.* fishing boats) which leads to over-exploitation of a common property resource will also fit into the general framework of the paper if it is difficult to reverse.

There are, of course, alternative explanations for gradualism. For example, GATT negotiations on tariff reductions are notable for their gradualism, and a small theoretical literature now rationalizes this in terms of self-enforceability (Staiger (1995), Devereux (1997), Furusawa and Lai (1999)). The general idea[6] is that initially, full liberalization cannot be self-enforcing, as the benefits of deviating from free trade are too great to be dominated by any credible punishment. But if there is partial liberalization, some structural economic change induced by the initial liberalization reduces the benefits of deviation from further trade liberalization, and/or raises the costs of punishment to the deviator.[7] So, in this case, irreversibility is neither plausible (tariff cuts can always be reversed) nor required to explain gradualism.

2. THE MODEL AND PRELIMINARY RESULTS

There are two players[8] $i = 1, 2$. In each period, $t = 1, 2, \ldots$ both players $i = 1, 2$ simultaneously choose an action variable $c_i \in \Re_+$, measuring $i's$ level of cooperation.[9] The per-period payoff to player 1 is $\pi(c_1, c_2)$ with that of player 2 being $\pi(c_2, c_1)$. So, payoffs of the two players are identical following a permutation of the pair of actions. Also, we assume that π is continuous, strictly decreasing in c_1 and strictly increasing in c_2. The last two conditions ensure that the one-shot game has a Prisoner's Dilemma structure. Payoffs over the infinite horizon are discounted by common discount factor δ, $0 < \delta < 1$. Finally, our crucial assumption is that the choice of action is irreversible in every period, *i.e.*

$$c_{i,t} \geq c_{i,t-1}, \qquad i = 1, 2, \qquad t = 1, 2, \ldots, \tag{2.1}$$

where $c_{i,t}$ is i's action in period t. Without loss of generality, we set $c_{1,0} = c_{2,0} = 0$. These irreversibility constraints imply that the game is dynamic, rather than repeated.

We now make the following further weak assumption on π. First, define $w(c) := \pi(c, c)$, and let c^* be the smallest value of c that maximises w, if it exists.[10] In what follows, we refer to c^* as the *first-best efficient level of cooperation*. Our interest in c^* follows from the fact that we focus on symmetric equilibrium paths, as defined below.

A1. *There exists a maximiser of $w(c)$, c^*, such that $w(c)$ is strictly increasing in c for all $0 \leq c \leq c^*$.*

A *game history* at time t is defined in the usual way as sequence of action pairs $\{c_{1,\tau}, c_{2,\tau}\}_{\tau=1}^{t-1}$, and is observable to both players at t. A *pure strategy* for player $i = 1, 2$

6. The individual papers differ in their description of the structural change induced by partial liberalization. Staiger (1995) endows workers in the import competing sector with specific skills, making them more productive there than elsewhere in the economy. When they move out of this sector, they lose their skills with some probability. In Devereux (1997), there is dynamic learning-by-doing in the export sector. In Furusawa and Lai (1999), there are linear adjustment costs incurred when labor moves between sectors.

7. A formal treatment of a related idea in the negotiation context is in Compte and Jehiel (1995) who consider the impact of outside options in a negotiation model where concessions by one party increase the payoff of the other in a dispute resolution phase.

8. Our main results generalise straightforwardly to more than two players.

9. The action spaces can also be bounded, *i.e.* $c_i \in [0, \bar{c}]$; see the end of Section 3.

10. In general, it is possible that the sum of players' payoffs could be made higher than $2w(c^*)$ by some asymmetric pair (c_1^*, c_2^*) with $c_1^* \neq c_2^*$, in which case *both* players would be better off with a $50 : 50$ randomization over (c_1^*, c_2^*) and (c_2^*, c_1^*) than with (c^*, c^*). One assumption sufficient to rule this out is that $\phi(c_1, c_2) := \pi(c_1, c_2) + \pi(c_2, c_1)$ has a unique global maximum on $\Re_+^2$. From this assumption, and the fact that $\phi(c_1, c_2)$ is symmetric, it follows that ϕ is maximised when $c_1 = c_2 = c^*$.

is defined as a sequence of mappings from game histories in periods $t = 1, 2, \ldots$ to values of $c_{i,t}$ in $\Re_+$, and where every pair $(c_{i,t-1}, c_{i,t})$ satisfies (2.1). An *outcome path* of the game is a sequence of actions $\{c_{1,t}, c_{2,t}\}_{t=1}^{\infty}$ that is generated by a pair of pure strategies. We are interested in characterizing pure-strategy subgame-perfect Nash equilibrium[11] outcome paths. For the moment, we restrict our attention to *symmetric* equilibrium outcome paths where $c_{1,t} = c_{2,t} = c_t$, $t = 1, 2, \ldots$ and we denote[12] such paths by $\{c_t\}_{t=1}^{\infty}$. In view of the fact that the underlying model is symmetric, this is a reasonable restriction. It is relaxed in Section 5.

We now derive necessary and sufficient conditions for some fixed symmetric outcome path $\{c_t\}_{t=1}^{\infty}$ to be an equilibrium. Consider some deviation c_t' by player i at t. It is clear from the fact that π is decreasing in its first argument that the following is a subgame-perfect equilibrium path in the continuation game following the deviation: both players immediately and permanently stop increasing their levels of cooperation, *i.e.* $c_{i,\tau} = c_t'$, $c_{j,\tau} = c_t$ all $\tau > t$. It is also clear that this path imposes the worst punishment on i that j can inflict, given the irreversibility constraint (2.1). The continuation payoff to i from this punishment equilibrium is $\pi(c_t', c_t)/(1-\delta)$. As π is decreasing in its first argument, it is clear that if i anticipates this punishment equilibrium, the optimal deviation for i at any date t is to set c_t' as low as possible, *i.e.* $c_t' = c_{t-1}$.

Consequently, for a non-decreasing sequence $\{c_t\}_{t=1}^{\infty}$ to be a (symmetric) equilibrium outcome path it is necessary and sufficient that the optimal deviation is never profitable at any $t \geq 1$, *i.e.* $\{c_t\}_{t=1}^{\infty}$ satisfies:

$$\frac{\pi(c_{t-1}, c_t)}{1-\delta} \leq \pi(c_t, c_t) + \delta\pi(c_{t+1}, c_{t+1}) + \ldots, \tag{2.2}$$

all $t \geq 1$, where the L.H.S. is the punishment payoff, and the R.H.S. is the payoff from the non-decreasing equilibrium path. Let C_{SE} be the set of non-decreasing paths $\{c_t\}_{t=1}^{\infty}$ that satisfy (2.2). We now note[13] two basic properties of sequences in C_{SE}.

Lemma 2.1. *If $\{c_t\}_{t=1}^{\infty}$ is an equilibrium path, then (i) $c_t < c^*$, for all $t \geq 1$, and (ii) if $c_t > c_{t-1}$ for some $t > 0$, then for all $\tau \geq 0$, there exists a $\tau' > \tau$ such that $c_{\tau'} > c_\tau$ (i.e. the sequence never attains its limit).*

Next, say that the path $\{\widehat{c}_t\}_{t=1}^{\infty} \in C_{SE}$ is an *efficient* symmetric equilibrium path if there does not exist another sequence $\{c_t'\}_{t=1}^{\infty} \in C_{SE}$ such that $\sum_{t=1}^{\infty} \delta^{t-1}\pi(c_t', c_t') > \sum_{t=1}^{\infty} \delta^{t-1}\pi(\widehat{c}_t, \widehat{c}_t)$. We refer to such a path simply as an *efficient path* in what follows.[14] Define $\widehat{c}_\infty := \lim_{t\to\infty} \widehat{c}_t$ (which exists by Lemma 2.1). We now have:

Lemma 2.2. *An efficient path $\{\widehat{c}_t\}_{t=1}^{\infty}$ exists, and any efficient path satisfies inequalities (2.2) with equality, i.e. for all $t \geq 1$,*

$$\frac{\pi(\widehat{c}_{t-1}, \widehat{c}_t)}{1-\delta} = \pi(\widehat{c}_t, \widehat{c}_t) + \delta\pi(\widehat{c}_{t+1}, \widehat{c}_{t+1}) + \ldots. \tag{2.3}$$

Lemma 2.2 does not rule out the possibility of multiple efficient paths. The next lemma shows that any efficient path is the upper envelope of all equilibrium paths, and hence that the efficient path is unique, as there can be only one upper envelope.

11. In the sequel, it is understood that "equilibrium" refers to subgame-perfect Nash equilibrium.
12. This is a slight abuse of notation in the interests of brevity, as in fact a symmetric outcome path is $\{c_t, c_t\}_{t=1}^{\infty}$.
13. All Lemmas in this section are proved in the Appendix.
14. We use the term "first-best" to refer to unconstrained efficient paths (not constrained to lie in C_{SE}).

Lemma 2.3. *The efficient path* $\{\widehat{c}_t\}_{t=1}^{\infty}$ *is unique and is the upper envelope of all equilibrium paths, i.e. there does not exist a* $\{c'_t\}_{t=1}^{\infty} \in C_{SE}$ *with* $c'_t > \widehat{c}_t$, *for some* t.

We are now able to show, using Lemmas 2.1 and 2.2, that the efficient path must satisfy a simple second-order difference equation. Say that a difference equation in c_t has a *bounded solution* if (given the initial conditions), $|c_t| < b$, all t, for some $b \in \Re_+$.

Lemma 2.4. *Any path* $\{c_t\}_{t=1}^{\infty}$ *with* $c_1 \geq 0$ *is non-decreasing and solves* (2.3) *if and only if it is a bounded solution to the difference equation*

$$\pi(c_t, c_{t+1}) = \frac{1}{\delta}[\pi(c_{t-1}, c_t) - \pi(c_t, c_t)] + \pi(c_t, c_t), \qquad t > 1, \tag{2.4}$$

with initial conditions $\bar{c}_0 = 0$, $\bar{c}_1 = c_1$.

Now, since the efficient path is non-decreasing and solves (2.3), it must, by the above lemma, solve the difference equation (2.4) with initial conditions $c_0 = 0$ and c_1 yet to be determined. Let the sequence $\{c_t(c_1; \delta)\}_{t=1}^{\infty}$ be a solution to the difference equation (2.4) with some fixed initial condition c_1, and consider the set of initial conditions c_1 such that $\{c_t(c_1; \delta)\}_{t=1}^{\infty}$ converges to a finite limit, *i.e.* $C_1(\delta) := \{c_1 | \lim_{t\to\infty} c_t(c_1; \delta) < +\infty\}$. Then we have our final result of this section:

Lemma 2.5. *If, for any* $c_1 \geq 0$, $\{c_t(c_1; \delta)\}_{t=1}^{\infty}$ *is a convergent sequence, then it is also an equilibrium path. Moreover, the efficient path is the sequence* $\{c_t(\widehat{c}_1; \delta)\}_{t=1}^{\infty}$, *where* $\widehat{c}_1 = \max\{c_1 | c_1 \in C_1(\delta)\}$.

3. MAIN RESULTS

The first main result consolidates and extends the preliminary results to get a characterisation of the efficient path:

Proposition 3.1. *A unique efficient path* $\{\hat{c}_t\}_{t=1}^{\infty}$ *exists, and on this path, either there is no cooperation at all* ($\widehat{c}_t = 0$, $t = 0, 1, \ldots$), *or the level of cooperation must strictly increase in every period* ($\widehat{c}_{t+1} > \widehat{c}_t$, *all* $t > 0$). *In either case, the efficient path solves the difference equation* (2.4) *with initial conditions* $\bar{c}_0 = 0$, $\bar{c}_1 = \widehat{c}_1 = \max\{c_1 | c_1 \in C_1(\delta)\}$.

Proof. The first part of the Proposition is from Lemma 2.3, and the third is from Lemma 2.5. To prove the second part, note that if there is ever any cooperation, there is a date τ at which $c_\tau > c_{\tau-1} = 0$. Then, by an induction argument as in the sufficiency part of the proof of Lemma 2.4, $c_t > c_{t-1}$, all $t \geq \tau$. Now suppose that $\tau > 1$: then the path could not be efficient, as clearly the path $\{c'_t\}_{t=1}^{\infty}$ with $c'_t = c_{t+1}$, all t, is an equilibrium path, and gives each player a higher present-value payoff, as it brings each payoff forward one period. ‖

This characterization of the efficient path allows the efficient path to be approximately computed in particular examples. For example, if $\pi(c_i, c_j) = -c_i^2/2 + c_j$, and $\delta = 0{\cdot}8$, then Figure 1 shows the solution to the difference equation (2.4) for different start values c_1. The highest value consistent with convergence is $c_1 = 0{\cdot}4$, in which case $c_t \to 0{\cdot}8$. (This is formally confirmed in Corollary 3.4 below, which implies for this case that the limit of the efficient sequence $\widehat{c}_\infty = \delta = 0{\cdot}8$; for this case $c^* = 1$.)

However, it is obviously of interest to have a *general* characterization of the limit of the efficient sequence, $\widehat{c}_\infty$, and we now turn to this issue. Whenever π is differentiable at (c, c),

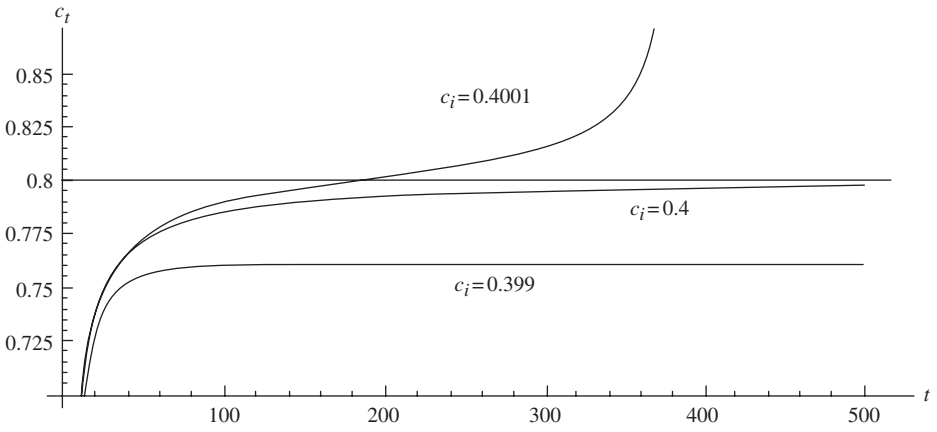

FIGURE 1
Simulation of difference equation

$c \in (0, c^*)$, define the function

$$\gamma(c) := \frac{-\pi_1(c,c)}{\pi_2(c,c)} > 0,$$

where π_i denotes the derivative of π with respect to its ith argument. Note that $\gamma(c)$ is the ratio of the cost $-\pi_1$, to the benefit π_2, of a small increase in cooperation by *both* players, starting at c. Moreover, define $\gamma(0) := \lim_{c \downarrow 0} \gamma(c)$ and $\gamma(c^*) := \lim_{c \uparrow c^*} \gamma(c)$ whenever these limits exist. In many special cases, the cost-benefit ratio may be increasing in c. More generally, if π is twice continuously differentiable, then from A1, we must have $w''(c^*) \leq 0$; if this inequality is strict then[15] $\gamma'(c^*) > 0$ and so γ is increasing on an interval $[c', c^*]$ for some $c' < c^*$. We cannot assert, however, that γ is everywhere increasing on $[0, c^*]$ on the basis of assumptions made so far.

Our main result characterising $\widehat{c}_\infty$ can now be stated.

Proposition 3.2. *Assume that A*1 *is satisfied. (i) If π is continuously differentiable*[16] *at the limit of the efficient symmetric path, $\widehat{c}_\infty$, then $\widehat{c}_\infty$ satisfies $\gamma(\widehat{c}_\infty) = \delta$. (ii) Moreover, if π is continuously differentiable on some interval $(0, \varepsilon)$ with $\gamma(0) < \delta$, then $\widehat{c}_\infty > 0$, and if π is continuously differentiable on some interval $(c^* - \varepsilon, c^*)$ with $\gamma(c^*) > \delta$, then $\widehat{c}_\infty < c^*$.*

We now have an immediate corollary, which gives quite weak sufficient conditions for cooperation on the efficient path to be uniformly bounded below the first-best level.

Corollary 3.3. *If π is continuously differentiable on some interval $(c^* - \varepsilon, c^* + \varepsilon)$, $\varepsilon > 0$, then $\widehat{c}_\infty < c^*$. So, the efficient path is uniformly bounded below the first-best efficient level of cooperation; i.e. $\widehat{c}_t < \widehat{c}_\infty < c^*$ for all t.*

15. Note that $w'' = \pi_{11} + 2\pi_{12} + \pi_{22}$, and $\gamma'(c) = \frac{-1}{\pi_2}[\pi_{11} + \pi_{12} + \gamma(\pi_{22} + \pi_{12})]$. So, the result follows from $w''(c^*) < 0$ and $\gamma(c^*) = 1$.

16. By π being differentiable at c is meant that $\pi(c_1, c_2)$ is differentiable at $c_1 = c_2 = c$ and by π being continuously differentiable on some interval (c', c'') is meant that $\pi(c_1, c_2)$ is continuously differentiable for $c_1, c_2 \in (c', c'')$.

This follows from part (ii) of the Proposition, noting that $\gamma(c^*) = 1 > \delta$ in this case.

Proof of Proposition 3.2. (i) By the assumption that π is continuously differentiable at $\hat{c}_\infty, \hat{c}_\infty > 0$. (a) Assume, first, that $\gamma(\hat{c}_\infty) > \delta$; a contradiction will be established. Suppose that $c_{t-2} < c_{t-1} < c_t$, and that π is continuously differentiable on some open interval enclosing $[c_{t-2}, c_t]$. By the Mean Value Theorem, $\pi(c_{t-1}, c_t) - \pi(c_{t-1}, c_{t-1}) = \pi_2(c_{t-1}, \theta_t)\Delta c_t$, for some $\theta_t \in (c_{t-1}, c_t)$, and $\pi(c_{t-2}, c_{t-1}) - \pi(c_{t-1}, c_{t-1}) = -\pi_1(\theta_{t-1}, c_{t-1})\Delta c_{t-1}$, for some $\theta_{t-1} \in (c_{t-2}, c_{t-1})$, where $\Delta c_t := c_t - c_{t-1}$. So, substituting in (2.4) and rearranging, we get

$$\Delta c_t = -\frac{\pi_1(\theta_{t-1}, c_{t-1})}{\delta \pi_2(c_{t-1}, \theta_t)} \Delta c_{t-1} \equiv a(c_{t-2}, c_{t-1}, c_t)\Delta c_{t-1}. \tag{3.1}$$

The limit of $a(c_{t-2}, c_{t-1}, c_t)$ as $c_{t-2}, c_{t-1}, c_t \to \hat{c}_\infty$, with $c_{t-2} < c_{t-1} < c_t$ exists, by $\pi(\cdot,\cdot)$ being continuously differentiable, and equals $-\pi_1(\hat{c}_\infty, \hat{c}_\infty)/\delta\pi_2(\hat{c}_\infty, \hat{c}_\infty) = \gamma(\hat{c}_\infty)/\delta > 1$. Consequently, there must exist a T such that

$$a(\hat{c}_{t-2}, \hat{c}_{t-1}, \hat{c}_t) > 1, \qquad t > T. \tag{3.2}$$

Also, as the equilibrium path is strictly increasing, $\Delta\hat{c}_T > 0$. But then from (3.1) and (3.2), the increments Δc_t are increasing when $t > T$ and so $\hat{c}_t$ cannot converge, contrary to hypothesis.
(b) Next assume that $\gamma(\hat{c}_\infty) < \delta$. We shall again establish a contradiction. By the continuous differentiability of π, find a neighbourhood around $\hat{c}_\infty$, $(\hat{c}_\infty - \varepsilon, \hat{c}_\infty + \varepsilon)$, and a $k < 1$, such that

$$-\pi_1(c, c')/(\delta\pi_2(c'', c''')) < k \qquad \text{for } c, c', c'', c''' \in (\hat{c}_\infty - \varepsilon, \hat{c}_\infty + \varepsilon). \tag{3.3}$$

Define $\psi := (1-k)\varepsilon/2$, and consider T such that $\hat{c}_{T-2} > \hat{c}_\infty - \psi$ (this must exist by virtue of $\hat{c}_\infty$ being the limit of $\{\hat{c}_t\}$). Now, since $\hat{c}_{T-2} < \hat{c}_{T-1} < \hat{c}_T < \hat{c}_\infty$, then by $c_t(c_1; \delta)$ being continuous in c_1, we can find $\tilde{c}_1 > \hat{c}_1$ such that, defining $\tilde{c}_t := c_t(\tilde{c}_1; \delta)$ all t, $\tilde{c}_{T-2}$, $\tilde{c}_{T-1}$ and $\tilde{c}_T \in (\hat{c}_\infty - \psi, \hat{c}_\infty)$. Consequently $0 < \Delta\tilde{c}_T \equiv \tilde{c}_T - \tilde{c}_{T-1} < \psi$.

We will show that this new sequence still converges. We first claim that, for $t > T$, if $\tilde{c}_{t-1} < \hat{c}_\infty + \varepsilon/2$ and $\Delta\tilde{c}_{t-1} < \psi$, then $\Delta\tilde{c}_t \le k\Delta\tilde{c}_{t-1}$. Rearranging (2.4):

$$\pi(\tilde{c}_{t-1}, \tilde{c}_t) - \pi(\tilde{c}_{t-1}, \tilde{c}_{t-1}) = [\pi(\tilde{c}_{t-2}, \tilde{c}_{t-1}) - \pi(\tilde{c}_{t-1}, \tilde{c}_{t-1})]/\delta. \tag{3.4}$$

We have $\pi(\tilde{c}_{t-1}, \tilde{c}_t) - \pi(\tilde{c}_{t-1}, \tilde{c}_{t-1}) \ge \underline{\pi}_2\Delta\tilde{c}_t$ provided $\tilde{c}_t < \hat{c}_\infty + \varepsilon$, where $\underline{\pi}_2 := \inf_{c,c' \in (\hat{c}_\infty - \varepsilon, \hat{c}_\infty + \varepsilon)} \pi_2(c, c')$, and $0 < [\pi(\tilde{c}_{t-2}, \tilde{c}_{t-1}) - \pi(\tilde{c}_{t-1}, \tilde{c}_{t-1})]/\delta \le -\underline{\pi}_1\Delta\tilde{c}_{t-1}/\delta$, where $\underline{\pi}_1 := \inf_{c,c' \in (\hat{c}_\infty - \varepsilon, \hat{c}_\infty + \varepsilon)} \pi_1(c, c')$ (recall that $\pi_1 < 0$). Also $k\Delta\tilde{c}_{t-1} < k\psi = k(1-k)\varepsilon/2 < \varepsilon/2$ by $0 < k < 1$, and so $\tilde{c}_{t-1} + k\Delta\tilde{c}_{t-1} < \hat{c}_\infty + \varepsilon$. Thus as $\tilde{c}_t$ varies between $\tilde{c}_{t-1}$ and $\tilde{c}_{t-1} + k\Delta\tilde{c}_{t-1}$, the L.H.S. of (3.4) varies between 0 and at least $\underline{\pi}_2 k\Delta\tilde{c}_{t-1}$ while by (3.3) $\underline{\pi}_2 \ge -\underline{\pi}_1/(k\delta)$, so $\underline{\pi}_2 k\Delta\tilde{c}_{t-1} \ge -\underline{\pi}_1\Delta\tilde{c}_{t-1}/\delta$, and thus $\underline{\pi}_2 k\Delta\tilde{c}_{t-1}$ is an upper bound on the R.H.S. of (3.4). So, given c_{t-1}, c_{t-2}, there must be a solution to (3.4) for $\tilde{c}_t \in (\tilde{c}_{t-1}, \tilde{c}_{t-1} + k\Delta\tilde{c}_{t-1})$, which implies $\Delta\tilde{c}_t \le k\Delta\tilde{c}_{t-1}$. Since the solution to (3.4) is clearly unique, this establishes the claim.

Next we show that $\{\tilde{c}_t\}$ converges to a limit no greater than $\hat{c}_\infty + \varepsilon/2$. Suppose to the contrary there exists $\tau > T$ such that $\tilde{c}_\tau > \hat{c}_\infty + \varepsilon/2$, with $\tilde{c}_{\tau-1} \le \hat{c}_\infty + \varepsilon/2$. By the fact that $\tilde{c}_T < \hat{c}_\infty$, $\Delta\tilde{c}_T < \psi$, and that $\Delta\tilde{c}_t \le k\Delta\tilde{c}_{t-1}$ for $T < t \le \tau$, we have $\tilde{c}_\tau \le \tilde{c}_T + k(1 - k^{\tau-T})\psi/(1-k) = \tilde{c}_T + k(1-k^{\tau-T})\varepsilon/2 < \hat{c}_\infty + \varepsilon/2$, which is a contradiction.

So, since $\{\tilde{c}_t\}_{t=1}^\infty$ is a convergent path it is also an equilibrium path (Lemma 2.5(i)). Also, by construction, $\tilde{c}_1 > \hat{c}_1$, which contradicts the envelope property of the efficient equilibrium (Lemma 2.3). So, $\gamma(\hat{c}_\infty) < \delta$ is also impossible.

(c) From parts (a) and (b), it follows immediately that $\gamma(\hat{c}_\infty) = \delta$, as was to be proved.
(ii) If $\gamma(0) < \delta$, then suppose to the contrary that $\hat{c}_\infty = 0$. By the assumption that $\gamma(0)$ exists and is less than δ, $\gamma(c) < k\delta$, where $k < 1$, on some interval $(0, \varepsilon)$. Next, choose, by continuity,

$c_1' > 0$ such that c_1', $c_2(c_1';\delta)$ and $c_3(c_1';\delta) \in (0,(1-k)\varepsilon/2)$. Then repeating the argument of part (i)(b) above, since $\Delta c_3 < (1-k)\varepsilon/2$, $c_\infty(c_1';\delta) < (1-k)\varepsilon/2 + \varepsilon/2 < \varepsilon$, and we have constructed a higher equilibrium path, which is again a contradiction. Finally if $\gamma(c^*) > \delta$, then the argument of (i)(a) above applies *mutatis mutandis* to show that $\widehat{c}_\infty = c^*$ is impossible. ‖

Our main result was deliberately stated making minimal assumptions on π, and therefore γ. We now consider two special cases for which we can get a sharper characterization of $\widehat{c}_\infty$. In one of these cases, we can also solve explicitly for the efficient path.

The differentiable monotonic case. π is everywhere continuously differentiable and $\gamma(c)$ is strictly increasing on $(0, c^*)$.

In this case, we can define $\widehat{c}(\delta)$ to be the unique solution to the equation $\gamma(\widehat{c}) = \delta$, unless $\gamma(0) > \delta$, in which case we define $\widehat{c}(\delta) = 0$. Clearly $\widehat{c}(\delta) < c^*$ with $\lim_{\delta\to 1}\widehat{c}(\delta) = c^*$, and $\widehat{c}(\delta)$ can easily be computed in special cases. It follows now from Proposition 3.2 that:

Corollary 3.4. *In the differentiable monotonic case,* $\widehat{c}_\infty = \widehat{c}(\delta)$.

Proof. In this case, $\gamma(c^*) = 1$, so by Proposition 3.2(ii), $\widehat{c}_\infty < c^*$. If $\gamma(0) < \delta$, then by Proposition 3.2(ii), $\widehat{c}_\infty > 0$, and thus the result follows immediately from the definition of $\widehat{c}(\delta)$ and Proposition 3.2(i). If $\gamma(0) \geq \delta$, then as γ is increasing, $\gamma(c) > \delta$ on $(0, c^*)$, and moreover, γ is continuously differentiable at all $c \in (0, c^*)$. So, $\widehat{c}_\infty \notin (0, c^*)$: otherwise, by Proposition 3.2(i), $\gamma(\widehat{c}_\infty) = \delta$, contradicting the assumed properties of γ. Consequently, $\widehat{c}_\infty = 0$. ‖

Note that the differentiable monotonic case also satisfies the assumptions of Corollary 3.3, so for all $\delta < 1$, the efficient path is uniformly bounded below the first-best efficient level of cooperation; *i.e.* $\widehat{c}_t < \widehat{c}(\delta) < c^*$ for all t. The key feature of the differentiable monotonic case is that we have an operational formula for $\widehat{c}_\infty$. For example, if $\pi(c_i, c_j) = c_j - 0{\cdot}5(c_i)^2$, then $c^* = 1$, $\widehat{c}_\infty = \hat{c}(\delta) = \delta$.

The linear kinked case.

$$\pi = \begin{cases} \pi_1 c_1 + \pi_2 c_2 & \text{if } c_1 + c_2 \leq 2c^*, \\ \pi_1 c_1 + \pi_2(2c^* - c_1) & \text{if } c_1 + c_2 > 2c^*, \end{cases}$$

where $\pi_1 < 0$, $\pi_2 > 0$ are constants[17] with $\pi_1 + \pi_2 > 0$.

Note that in the linear kinked case, Assumption A1 above on the shape of $w(c)$ is automatically satisfied: $w(c)$ is linear and increasing in c until c reaches the efficient level c^*, and after that, higher cooperation yields negative benefit. In this case, we have the following striking result.

Corollary 3.5. *Assume the linear kinked case. If there is sufficiently little discounting* ($\delta > -\pi_1/\pi_2$), *then* $\widehat{c}_\infty = c^*$, *i.e. first-best efficient cooperation can be asymptotically obtained. In this case, the efficient path can be solved for explicitly as* $\widehat{c}_t = (1 - a^t)c^*$, $a := -\pi_1/\delta\pi_2$. *Otherwise (i.e. if* $\delta \leq -\pi_1/\pi_2$), *then* $\widehat{c}_\infty = 0$, *so no cooperation can ever be obtained* ($\widehat{c}_t = 0$, *all t).*

17. An interpretation is that payoffs depend positively on $(c_1 + c_2)$ up to $2c^*$ with a coefficient of π_2, but there is a marginal utility cost of $(\pi_2 - \pi_1)$ to increasing one's own c_i. For $c_1 + c_2 > 2c^*$, there is no more benefit from joint contributions, only the cost remains, so that joint payoffs are declining in $(c_1 + c_2)$. For $c_1 + c_2 > 2c^*$, all that is needed for the results is that joint payoffs are nonincreasing in $(c_1 + c_2)$ and also own payoffs are declining in own c_i.

Proof. Assume first $-\pi_1/\pi_2 < \delta$. Here, by definition, $\gamma(c) = -\pi_1/\pi_2$, $c \leq c^*$. So, if $\widehat{c}_\infty < c^*$, then π is differentiable at $\widehat{c}_\infty$ but $\gamma(\widehat{c}_\infty) < \delta$, contradicting Proposition 3.2(i). Since $\widehat{c}_\infty \leq c^*$ by Lemma 2.1, it follows that $\widehat{c}_\infty = c^*$. In this case, we can solve explicitly for the initial condition that gives the efficient path. Rearranging (2.4) for the kinked linear case, we get: $\Delta c_t = a\Delta c_{t-1}$, where $\Delta c_t = c_t - c_{t-1}$. So $c_t = \sum_{\tau=1}^{t} a^{\tau-1}c_1$ which converges to $c_\infty = \frac{1}{1-a}c_1 = (1 + \frac{\pi_1}{\delta\pi_2})^{-1}c_1$ if and only if $a < 1$. So, $\widehat{c}_1 = (1 + \frac{\pi_1}{\delta\pi_2})c^*$, and consequently, $\widehat{c}_t = (1 - a^t)c^*$.

For the case $-\pi_1/\pi_2 \geqslant \delta$, a symmetric argument implies that if $c_1 > 0$, $c_t \to \infty$ as $t \to \infty$, contradicting the assumption that $c_t < c^*$, all t. So, we must have $c_1 = 0$, implying $\widehat{c}_\infty = 0$. ‖

Note that in both the differentiable monotonic and kinked linear cases, we have shown that as $\delta \to 1$, the limiting level of cooperation on the efficient equilibrium path, $\widehat{c}_\infty$, tends to the first-best efficient level, c^*. It turns out that this fact implies that payoffs also converge to their efficient levels as $\delta \to 1$; *i.e.* there is no limiting inefficiency in this model.

Corollary 3.6. *In either the differentiable or linear kinked cases, as* $\delta \to 1$, *the normalized discounted payoff from the efficient path,* $\hat{\Pi} = (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\pi(\widehat{c}_t, \widehat{c}_t)$, *converges to the first-best payoff* $\pi(c^*, c^*)$.

Proof. Rewrite the equilibrium condition (2.2) as

$$\pi(c_{t-1}, c_t) \leq (1-\delta)\sum_{\tau=t}^{\infty}\delta^{\tau-t}\pi(c_\tau, c_\tau), \qquad t \geq 1. \tag{3.5}$$

Now, if $\{c_t\}_{t=1}^{\infty}$ is an equilibrium sequence at δ, then $\{c_t\}_{t=1}^{\infty}$ is also an equilibrium at any $\delta' > \delta$ since, as $\pi(c_t, c_t)$ is a non-decreasing sequence, the R.H.S. of (3.5) is non-decreasing in δ, and the L.H.S. is constant.

Take $\widehat{c}(\delta)$ as already defined for the differentiable case, and in the linear kinked case, define $\widehat{c}(\delta)$ as:

$$\widehat{c}(\delta) = \begin{cases} c^* & \text{if } \delta > -\pi_1/\pi_2, \\ 0 & \text{otherwise.} \end{cases}$$

So, for any $\varepsilon > 0$, find a $\overline{\delta}$ such that $\pi(\widehat{c}(\overline{\delta}), \widehat{c}(\overline{\delta})) > \pi(c^*, c^*) - \varepsilon$ (where in the differentiable case, we use the continuity of $\pi(\cdot,\cdot)$, and, as already remarked, $\lim_{\delta\to 1}\widehat{c}(\delta) = c^*$). From Corollaries 3.4 and 3.5, at $\overline{\delta}$, $\widehat{c}_t \to \widehat{c}(\delta)$, so holding $\{\widehat{c}_t\}_{t=1}^{\infty}$ fixed, $\lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\pi(\widehat{c}_t, \widehat{c}_t) \to \pi(\widehat{c}(\delta), \widehat{c}(\delta))$, and hence there exists a $\delta' > \overline{\delta}$ such that for δ satisfying $\delta' < \delta < 1$, $(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\pi(\widehat{c}_t, \widehat{c}_t) > \pi(c^*, c^*) - \varepsilon$. Since $\{\widehat{c}_t\}_{t=1}^{\infty}$ is an equilibrium sequence for such δ, the efficient path at such δ must also give a payoff greater than $\pi(c^*, c^*) - \varepsilon$. As ε is arbitrary, this completes the proof. ‖

An alternative way of viewing this result is to note that if we shrink the period length, holding payoffs per unit of time constant, then inefficiency disappears as period length goes to zero.[18] Note also that the proof establishes an interesting comparative statics result: when δ increases, every component of the path $\{\widehat{c}_t\}_{t=1}^{\infty}$ weakly increases for the simple reason that the original path remains an equilibrium path for higher δ; the upper envelope property of the efficient sequence (Lemma 2.3) then implies the result.

18. If π is discontinuous but otherwise satisfies our assumptions then asymptotic efficiency can fail. Consider an example in which player i benefits only from j's c_j, with an upwards jump in payoff at completion ($c_j = c^*$), and suffers continuous (increasing) costs from c_i. To be specific, suppose that $\pi(c_1, c_2) = -0{\cdot}5c_1 + \phi(c_2)$, where $\phi(c_2) = c_2$ for $c_2 < 1$, $\phi(c_2) = 2$ for $c_2 \geq 1$. Lemma 2.1 still applies, so $c_{i,t} < 1 \equiv c^*$, all t, and the payoff jump is never realised no matter how patient the players. As $\delta \to 1$, average payoffs converge to 0·5, whereas first-best payoffs are 1·5.

Finally, note that if the action space is bounded, so that $c_i \in [0, \bar{c}]$, then all the analysis of this section is unchanged as long as $\bar{c} > c^*$. If in fact $\bar{c} \leq c^*$, so that $\bar{c}$ now effectively replaces c^*, then in view of Lemma 2.1 (i), which holds *mutatis mutandis*, Proposition 3.2 can still be applied. For example, provided that π is differentiable with bounded first derivatives, and w has slope bounded above zero on $(0, \bar{c})$, $\gamma(c)$ is bounded below 1 and there will thus be, as in the linear kinked case, a critical discount factor above which $\bar{c}$ will be the limit of the efficient symmetric equilibrium path.

4. A MODEL WITH ADJUSTMENT COSTS

The model studied above is very stylized. In many economic applications, irreversibility arises more naturally when there is a stock variable which benefits both players, and a flow or incremental variable which is costly to increase, and is nonnegative. This non-negativity constraint implies that the value of the stock variable can never fall, *i.e.* the stock variable is irreversible. Here, we present a model with these features, and show that it can be reformulated so that it is a special case of our base model.

Player $i's$ payoff at time t is

$$u(c_{i,t}, c_{j,t}) - \alpha(c_{i,t} - c_{i,t-1}), \tag{4.1}$$

with u increasing in both arguments, and with $\alpha > 0$ being the cost of adjustment. Here, $c_{i,t}$ is to be interpreted as $i's$ *cumulative investment* in, or the stock level of, the cooperative activity. We assume that the investment flow is nonnegative, which implies that the stock level of cooperation is irreversible, *i.e.* $c_{i,t} \geq c_{i,t-1}$, $i = 1, 2$.

We now proceed as follows. The present value payoff for i in this model is

$$\begin{aligned}\Pi_i &= u(c_{i,1}, c_{j,1}) - \alpha(c_{i,1} - c_{i,0}) + \delta[u(c_{i,2}, c_{j,2}) - \alpha(c_{i,2} - c_{i,1})] + \dots \\ &= \sum_{t=1}^{\infty} \delta^{t-1}[u(c_{i,t}, c_{j,t}) - \alpha(1-\delta)c_{i,t}] + \alpha c_{i,0}.\end{aligned}$$

As initial levels of cooperation $c_{1,0}, c_{2,0}$ are fixed, we can think of this model as a special case of the model of the previous section (*i.e.* without adjustment costs) where per-period payoffs are

$$\pi(c, c') = u(c, c') - \alpha(1-\delta)c. \tag{4.2}$$

Of course, we require that π defined in (4.2) satisfies the conditions imposed in Section 2, and also satisfies the relevant conditions of either the differentiable or linear kinked case. If this is the case, then Corollaries 3.4 and 3.5 apply directly.

We now study an important economic application using this extension of our basic model, dynamic voluntary contribution games. This is not the only topic that can be studied in this way, it is chosen because it has already been studied quite intensively (Admati and Perry (1991), Fershtman and Nitzan (1991), Marx and Matthews (1998)), but nevertheless we are able to extend existing results: this, we believe, illustrates the power and flexibility of our approach.

A dynamic voluntary contribution game is one where players can simultaneously or sequentially make contributions towards the cost of a public project over a number of time periods. Marx and Matthews (1998) is the paper in this literature that is closest to our work. In their paper contributions are made simultaneously and benefits from the project are proportional to the amount contributed (up to a maximum, at which point the project is completed). We will show that a special case of their model can be written as an adjustment cost game as above, and that Corollary 3.5 above can be applied to extend some of their results. They consider a model in which N individuals simultaneously make nonnegative private contributions, in each of a finite or infinite number of periods, to a public project. We assume that $N = 2$, and let $c_{i,t}$ be the

cumulative contribution of a numeraire private good by i towards the public project. Individuals obtain at t a flow of utility $u = (1-\delta)v(c_{1,t} + c_{2,t})$ from the aggregate cumulative contribution $c_{1,t} + c_{2,t}$, where $v(\cdot)$ is piecewise linear:

$$v(c_1, c_2) = \begin{cases} \lambda(c_1 + c_2) & \text{if } c_1 + c_2 < 2c^* = C^*, \\ \lambda C^* + b & \text{if } c_1 + c_2 \geq C^*, \end{cases}$$

where we follow as closely as possible the notation of Marx and Matthews. Thus agents get benefit λ from each unit of cumulative contribution, and an additional benefit $b \geq 0$ when the project is "completed", *i.e.* when the sum of cumulative contributions reaches C^*. Also, the cost to i of an increment $c_{i,t} - c_{i,t-1}$ in his own cumulative contribution is simply $c_{i,t} - c_{i,t-1}$. It is assumed that $0{\cdot}5 < \lambda < 1$, so that it is socially efficient to complete the project (immediately, in fact), but not privately efficient to contribute anything. We consider the case where $b = 0$ and the time horizon is infinite (the $b = 0$ case unravels otherwise, in the sense that in the final period it is optimal to contribute nothing, which implies the same is true of the penultimate period, and so on).

Then, from (4.2), per period payoffs in the equivalent dynamic game are

$$\begin{aligned} \pi(c_1, c_2) &= (1-\delta)v(c_1, c_2) - (1-\delta)c_1 \\ &= \begin{cases} (1-\delta)[(\lambda - 1)c_1 + \lambda c_2] & \text{if } c_1 + c_2 < 2c^* = C^*, \\ (1-\delta)\lambda C^* - (1-\delta)c_1 & \text{if } c_1 + c_2 \geq C^*. \end{cases} \end{aligned}$$

This payoff function is clearly of the kinked linear type, where $\pi_1 = (1-\delta)(\lambda - 1) < 0$, $\pi_2 = (1-\delta)\lambda > 0$. So Corollary 3.5 applies directly to this version of the Marx–Matthews model. In particular, the critical value of δ in Corollary 3.5 is $\hat{\delta} = -\pi_1/\pi_2 = (1-\lambda)/\lambda$. Two results then follow directly from our Corollary 3.5 and its proof:

1. If $\delta > \hat{\delta}$, there is a class of equilibria, indexed by the initial condition c_1, where each player's cumulative contribution c_t converges to some value less than or equal to c^*, with the limit value increasing in c_1. Along the equilibrium path, incremental contributions fall at rate $\frac{(1-\lambda)}{\delta\lambda}$. The *efficient* symmetric equilibrium has initial contribution $c_1 = c^*(1 - \frac{(1-\lambda)}{\delta\lambda})$, and each player's cumulative contribution c_t converges to c^*.
2. If $\delta \leq \hat{\delta}$, then no contributions are made in any equilibrium.

Result 1 sharpens Proposition 3 and Corollary 3(ii) of Marx and Matthews, who show that for $\delta > \hat{\delta}$, there is an equilibrium with $c_t \to c^*$, and that for $\delta \simeq 1$, this equilibrium is approximately efficient. In the special case of $n = 2$ and $b = 0$, we not only confirm their results, but also show that the equilibrium they construct *is* the efficient equilibrium for *any* $\delta > \hat{\delta}$. Also, Result 2 is a complete converse result to their Proposition 3.

5. ASYMMETRIC COOPERATION

In the simultaneous move game, we only considered symmetric paths, *i.e.* where $c_{1,t} = c_{2,t} = c_t$. One question is whether the agents could both achieve higher (expected) equilibrium payoffs by playing asymmetrically. More generally, we are interested in the shape of the equilibrium payoff possibility frontier. Let $\{c_{1,t}, c_{2,t}\}_{t=1}^{\infty}$ be an arbitrary (possibly asymmetric) path. Then, by a similar argument to that given in Section 2, such a path is an equilibrium path if and only if

$$\frac{\pi(c_{1,t-1}, c_{2,t})}{(1-\delta)} \leq \sum\nolimits_{\tau=t}^{\infty} \delta^{\tau-t}\pi(c_{1,\tau}, c_{2,\tau}), \qquad t = 1, 2, \ldots, \tag{5.1}$$

$$\frac{\pi(c_{2,t-1}, c_{1,t})}{(1-\delta)} \leq \sum\nolimits_{\tau=t}^{\infty} \delta^{\tau-t}\pi(c_{2,\tau}, c_{1,\tau}), \qquad t = 1, 2, \ldots. \tag{5.2}$$

So, now we need a *pair* of sequences of incentive constraints to hold. Now let C_E be the set of equilibrium paths satisfying (5.1), (5.2), and Π_E in $\Re^2$ be the corresponding set of normalized[19] present discounted payoff pairs generated by paths in C_E. Let an equilibrium path in C_E that maximises the sum of present-value payoffs

$$\sum_{t=1}^{\infty} \delta^{t-1}[\pi(c_{1,t}, c_{2,t}) + \pi(c_{2,t}, c_{1,t})],$$

be denoted $\{\hat{c}_{1,t}, \hat{c}_{2,t}\}_{t=1}^{\infty}$: at least one such path exists by the arguments of the proof of Lemma 2.2. We refer to this as an *efficient equilibrium path*. In view of our previous restriction to symmetric equilibrium paths, a major question of interest is whether (one of) the efficient equilibrium path(s) is symmetric.

Proposition 5.1. *In the linear kinked case, Π_E is convex and symmetric about the* 45° *line. Moreover, one of the efficient equilibrium paths is symmetric, i.e. $\hat{c}_{1,t} = \hat{c}_{2,t}$, all t.*

Proof. Adapting Lemma 2.1, any sequence in C_E must have $c_{1,t} + c_{2,t} < 2c^*$, all t. Given this, the constraints (5.1), (5.2) are linear. Consequently, if $\{c'_{1,t}, c'_{2,t}\}_{t=1}^{\infty}$ and $\{c''_{1,t}, c''_{2,t}\}_{t=1}^{\infty}$ satisfy them, a convex combination of the two must also satisfy them and so C_E is a convex set. Consequently, Π_E is also convex, by linearity of payoffs. The symmetry claims follow straight forwardly. ‖

In fact, in the kinked linear case, we can say more[20] about the shape of Π_E as δ varies. As far as symmetric equilibria are concerned, we know from Corollary 3.5 if $\delta \leq \hat{\delta} = -\pi_1/\pi_2$, no cooperation is possible, so $\Pi_E(\delta) = \{0, 0\}$. The non-trivial case is where $\delta > \hat{\delta}$, in which case equilibria with positive levels of cooperation exist. Moreover, about the 45° line the efficient frontier of $\Pi_E(\delta)$ turns out to be linear (with slope -1) as in the segment AB in Figure 2. (The linear part of the frontier consists of payoffs from sequences which satisfy the incentive constraints with equality.) As $\delta \to 1$, the linear section extends to, but never attains, to the axes (with origin corresponding to the no cooperation payoffs $\pi(0, 0)$), and the entire frontier converges to the first-best efficient frontier.

6. SEQUENTIAL MOVES

So far, we have assumed that players can move simultaneously. However, it may be that players can only move sequentially, *e.g.* Admati and Perry (1991), Gale (2000). In certain public good contribution games, the assumption made can affect the conclusions substantially. In the Admati–Perry model, where players move sequentially, a no-contribution result holds when no player individually would want to complete the project, even though it might be jointly optimal to do so, but this result may disappear if the players can move simultaneously (see Marx and Matthews (1998) for a full discussion of this issue). By contrast, we shall find that in our model, equilibria in the two cases are closely related; indeed, the efficient symmetric equilibrium of the symmetric move game can "approximately" be implemented in the sequential move game.

Suppose w.l.o.g. that player 1 can move at even periods and player 2 at odd periods. Let the set of all non-decreasing paths that satisfy this restriction be C^{seq}. To be an equilibrium in the sequential game, any path $\{c_{1,t}, c_{2,t}\}_{t=1}^{\infty}$ must satisfy the following incentive constraints. When player 1 moves at $t = 2, 4, \ldots$ he prefers to raise his level of cooperation from $c_{1,t-2}$ to $c_{1,t}$

19. That is, multiplied by $1 - \delta$.
20. For proof of these claims, see Lockwood and Thomas (1999).

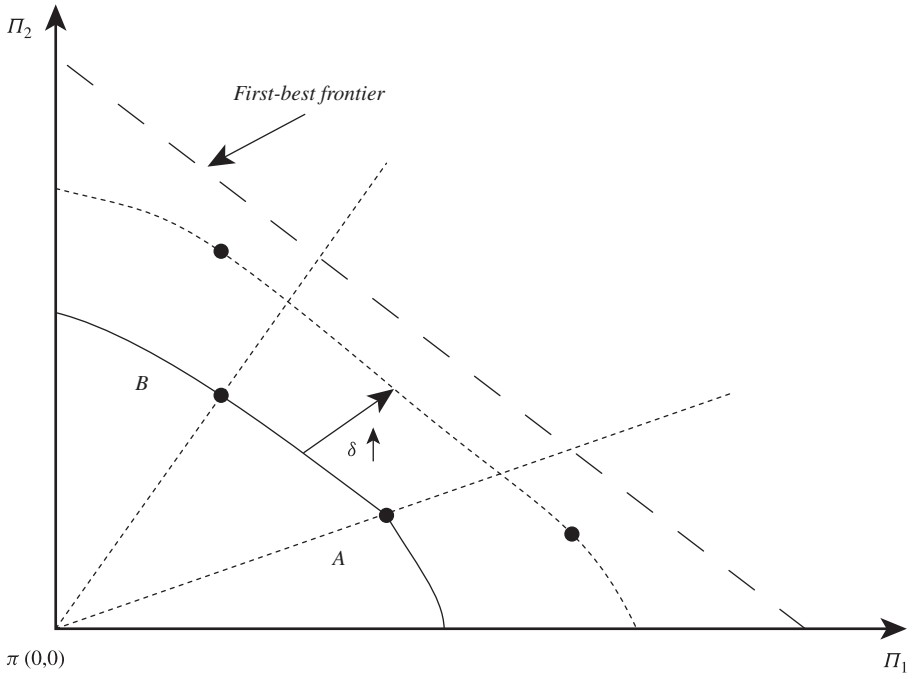

FIGURE 2
Asymmetric equilibria

only if

$$\frac{\pi(c_{1,t-2}, c_{2,t-1})}{1-\delta} \leq \pi(c_{1,t}, c_{2,t-1}) + \delta\pi(c_{1,t}, c_{2,t+1}) + \ldots, \qquad t = 2, 4, 6, \ldots. \tag{6.1}$$

Similarly, when player 2 moves at $t = 3, 5 \ldots$, he prefers to raise his level of cooperation from $c_{2,t-2}$ to $c_{2,t}$ only if

$$\frac{\pi(c_{2,t-2}, c_{1,t-1})}{1-\delta} \leq \pi(c_{2,t}, c_{1,t-1}) + \delta\pi(c_{2,t}, c_{1,t+1}) + \ldots, \qquad t = 3, 5, 7, \ldots. \tag{6.2}$$

When player 2 moves at period 1, (6.2) is modified by the fact that 2 can revert to $c_0 = 0$, rather than c_{-1}, but otherwise the incentive constraint is the same, *i.e.*

$$\frac{\pi(0,0)}{1-\delta} \leq \pi(c_{2,1}, 0) + \delta\pi(c_{2,1}, c_{1,2}) + \ldots. \tag{6.3}$$

Let the set of paths in C^{seq} that satisfy (6.1), (6.2) and (6.3) be $C_E^{\text{seq}} \subset C^{\text{seq}}$.

However, note that a path in C^{seq} is also in C_E^{seq} *if and only if* it is an (asymmetric, in general) equilibrium path in the simultaneous move game studied earlier. This is because in the simultaneous move game, the incentive constraints in the periods where agents do not have to move are *automatically* satisfied, as no agent likes to choose a higher $c_{i,t}$ than necessary (from π decreasing in its first argument). So, recalling the definition of C_E from the previous section, we have shown that $C_E^{\text{seq}} = C_E \cap C^{\text{seq}}$. Also, define Π_E^{seq} similarly to Π_E as the set of pairs of equilibrium normalised present-value payoffs in the sequential game. As $C_E^{\text{seq}} \subseteq C_E$, then $\Pi_E^{\text{seq}} \subseteq \Pi_E$; that is, players can *always* do at least as well with a simultaneous move structure as with a sequential one.

To say more than this, we shall go to the linear kinked case. Let $\hat{\pi}$ be the normalised present value payoff from the efficient symmetric path in the simultaneous move game.[21] By Proposition 5.1, we know that $(\hat{\pi}, \hat{\pi})$ is the equal utility point on the Pareto-frontier for that game. Finally, note that all equilibrium payoff sets depend on δ. Then:

Proposition 6.1. *Assume the linear kinked case. Then,* Π_E^{seq} *is convex. For any fixed* $\varepsilon > 0$*, there is a* $\delta(\varepsilon) < 1$*, and a point* $(\hat{\pi}_1^{\text{seq}}, \hat{\pi}_2^{\text{seq}}) \in \Pi_E^{\text{seq}}$ *such that* $\hat{\pi}_i^{\text{seq}} > \hat{\pi} - \varepsilon$*,* $i = 1, 2$ *for* $\delta \geq \delta(\varepsilon)$.

Proof. The proof that Π_E^{seq} is convex follows the proof of Proposition 5.1 exactly. Next, recall that $\{\widehat{c}_t\}_{t=1}^{\infty}$ is the (unique) symmetric efficient path in the simultaneous move game. Define the *asymmetric* path $\{\widetilde{c}_{1,t}, \widetilde{c}_{2,t}\}_{t=1}^{\infty}$ in C^{seq} as follows:

$$\begin{aligned} \widetilde{c}_{1,t} &= \widetilde{c}_{1,t+1} = \widehat{c}_t, \qquad t = 0, 2, 4, 6, \ldots; \\ \widetilde{c}_{2,t} &= \widetilde{c}_{2,t+1} = \widehat{c}_t, \qquad t = 1, 3, 5, \ldots. \end{aligned}$$

This is simply the path where an agent whose turn it is to move at t chooses $\widehat{c}_t$. We show that $\{\widetilde{c}_{1,t}, \widetilde{c}_{2,t}\}_{t=1}^{\infty} \in C_E^{\text{seq}}$. Define as before $\Delta\widehat{c}_t := \widehat{c}_t - \widehat{c}_{t-1}$, and recall $\Delta\widehat{c}_t = a\Delta\widehat{c}_{t-1}$ on the efficient path. For the player who moves at $t \geq 2$, and writing Δ for $\Delta\widehat{c}_{t-1}$, the constraints (6.1) and (6.2), evaluated on the path $\{\widetilde{c}_{1,t}, \widetilde{c}_{2,t}\}_{t=1}^{\infty}$, can be written as:

$$\begin{aligned} \frac{\pi_1\widehat{c}_{t-2} + \pi_2(\widehat{c}_{t-2} + \Delta)}{1-\delta} &\leq \pi_1(\widehat{c}_{t-2} + \Delta + a\Delta) + \pi_2(\widehat{c}_{t-2} + \Delta) \\ &\quad + \delta(\pi_1(\widehat{c}_{t-2} + \Delta + a\Delta) + \pi_2(\widehat{c}_{t-2} + \Delta + a\Delta + a^2\Delta)) \\ &\quad + \delta^2(\pi_1(\widehat{c}_{t-2} + \Delta + \cdots + a^3\Delta) \\ &\quad + \pi_2(\widehat{c}_{t-2} + \Delta + a\Delta + a^2\Delta)) + \ldots, \end{aligned} \tag{6.4}$$

or rearranging,

$$\frac{\pi_2\Delta}{1-\delta} \leq \frac{(1+a)\pi_1\Delta + (1-\delta^2a^2 + \delta a + \delta a^2)\pi_2\Delta}{(1-\delta)(1-\delta^2a^2)},$$

which holds with equality as $a = -\pi_1/(\delta\pi_2)$. Thus $\{\widetilde{c}_{1,t}, \widetilde{c}_{2,t}\}_{t=1}^{\infty}$ satisfies the equilibrium conditions (with equality) in the sequential game from $t = 2$ onwards. At $t = 1$ the equilibrium condition would hold with equality if player 2's inherited c was $-\widehat{c}_1/a$ as opposed to zero since it is higher, the condition will be slack (as $\pi_1 < 0$). So, we have established that $\{\widetilde{c}_{1,t}, \widetilde{c}_{2,t}\}_{t=1}^{\infty} \in C_E^{\text{seq}}$. Payoffs from the path $\{\widetilde{c}_{1,t}, \widetilde{c}_{2,t}\}$ are $\hat{\pi}_i^{\text{seq}} = (1-\delta)\{[\pi_j\widehat{c}_1] + \delta[\pi_i\widehat{c}_2 + \pi_j\widehat{c}_1] + \delta^2[\pi_i\widehat{c}_2 + \pi_j\widehat{c}_3] + \cdots\}$, $i, j = 1, 2$, $i \neq j$. Also, the payoffs from the efficient symmetric path in the simultaneous move game are $\hat{\pi} = (1-\delta)\{[\pi_1\widehat{c}_1 + \pi_2\widehat{c}_1] + \delta[\pi_1\widehat{c}_2 + \pi_2\widehat{c}_2] + \delta^2[\pi_1\widehat{c}_3 + \pi_2\widehat{c}_3] + \cdots\}$. Consequently, we get

$$\begin{aligned} \hat{\pi} - \hat{\pi}_1^{\text{seq}} &= (1-\delta)\{\pi_2\widehat{c}_1 + \delta\pi_1(\widehat{c}_2 - \widehat{c}_1) + \delta^2\pi_2(\widehat{c}_3 - \widehat{c}_2) + \delta^3\pi_1(\widehat{c}_4 - \widehat{c}_3) + \ldots\} \\ &= (1-\delta)\widehat{c}_1\{\pi_2\widehat{c}_1 + \delta\pi_1a\widehat{c}_1 + \delta^2\pi_2a^2\widehat{c}_1 + \delta^3\pi_1a^3\widehat{c}_1 \ldots\} \\ &= (1-\delta)\widehat{c}_1[\pi_2(1 + \delta^2a^2 + \delta^4a^4 + \ldots) + \delta a\pi_1(1 + \delta^2a^2 + \delta^4a^4 + \ldots)] \\ &= \frac{(1-\delta)\widehat{c}_1}{1-\delta^2a^2}[\pi_2 + \delta a\pi_1] < (1-\delta)\frac{\widehat{c}_1\pi_2}{1-(\pi_1/\pi_2)^2}. \end{aligned}$$

So, as $\widehat{c}_1 < c^*$ for all δ, $\hat{\pi} - (1-\delta)\theta < \hat{\pi}_1^{\text{seq}}$, for some constant $\theta > 0$. Consequently, for any $\varepsilon > 0$, $\hat{\pi} - \varepsilon < \hat{\pi}_1^{\text{seq}}$ for all $\delta \geq \delta(\varepsilon) = 1 - \varepsilon/\theta$, as required. A similar argument applies for $i = 2$. ∥

21. That is, $\hat{\pi} = (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\pi(\hat{c}_t, \hat{c}_t)$.

Consequently there is no limiting inefficiency due to the sequential structure as far symmetric payoffs are concerned. We can also show[22] that one point on the efficiency frontier of payoffs in the simultaneous move game is attained in the sequential game.

7. FURTHER EXTENSIONS AND CONCLUSIONS

In an earlier version of this paper (Lockwood and Thomas (1999))[23], we also extended the base model to allow for a small amount of reversibility of actions, so that any player can reduce his or her cooperation level by some (small) fixed percentage. This has two countervailing effects. The first is to make *deviation more profitable*; the deviator at t can lower his cooperation level below last period's, rather than just keeping it constant. The second effect is to make *punishment more severe*; the worst possible perfect equilibrium punishment is for the punisher to reduce his cooperation over time, rather than just not increase it. We are able to show that for a small amount of reversibility the second effect dominates, and in the linear kinked case it dominates for any degree of reversibility. In our model, then, reversibility is desirable in that it allows more cooperative equilibria to be sustained.[24]

This paper has studied a simple dynamic game where the level of cooperation chosen by each player in any period is irreversible. We have shown that irreversibility causes *gradualism*: any (subgame-perfect) sequence of actions involving partial cooperation cannot involve an immediate move to full cooperation, and we have refined and extended this basic insight in various ways. First, we showed that if payoffs are differentiable in actions, then (for a fixed discount factor) the level of cooperation asymptotes to a limit strictly below full cooperation, and this limit value is easily characterized. For the case where payoffs are linear up to some joint cooperation level, and decreasing thereafter, the results are different—above some critical discount factor equilibrium cooperation can converge asymptotically to the fully efficient level, but below this critical discount factor, no cooperation is possible. The basic model is then extended in several directions.

However, throughout, we have continued to assume that the underlying model is symmetric. This is somewhat restrictive; in many situations where irreversibility arises naturally, for example in Coasian bargaining without enforceable contracts but where actions are irreversible, payoffs will be asymmetric. Another limitation of the model is that players only have a scalar action variable; in many applications, players have several action variables, as in, for example, capacity reduction games, where firms control both capacity and output. Extending the model in these directions is a project for the future.

APPENDIX

Proof of Lemma 2.1. (i) Suppose to the contrary that $c_t \geq c^*$ for some $t > 0$, with $c_{t-1} < c^*$. From the assumptions on $\pi(\cdot,\cdot)$, $\pi(c_{t-1}, c_t) > \pi(c^*, c^*)$, and since $\pi(c_\tau, c_\tau) \leq \pi(c^*, c^*)$ all τ by definition of c^*, it is clear that (2.2) is violated, a contradiction.

(ii) If this is not the case, then on some equilibrium path, $c_t > c_{t-1}$ for some $t > 0$, and there exists a $T \geq t$ with $c_\tau < \widetilde{c}$ for $\tau < T$, and $c_\tau = \widetilde{c}$ for all $\tau \geq T$. Then, player 1, by deviating at T, would receive $\pi(c_{T-1}, \widetilde{c})/(1-\delta) > \pi(\widetilde{c}, \widetilde{c})/(1-\delta)$, where $\pi(\widetilde{c}, \widetilde{c})/(1-\delta)$ is 1's equilibrium continuation payoff at T, and where the inequality follows from π decreasing in its first argument. Thus a deviation is profitable, contradicting the equilibrium assumption. ‖

22. This is proved in Lockwood and Thomas (1999); the common point is one end of the linear section of the frontier of $\Pi_E(\delta)$, discussed in Section 5, and depicted in Figure 2 as point A.

23. Available online at `http://www.warwick.ac.uk/fac/soc/Economics/research/twerps.htm`

24. For any positive degree of reversibility, if players are sufficiently patient then c^* can be attained immediately. On the other hand, at a *fixed* discount factor, introducing a small amount of reversibility will not undo the gradualism result.

Proof of Lemma 2.2. (a) To prove existence, consider the product space of sequences $C^* := [0, c^*]^\infty$ endowed with the product topology, and let $\mathbf{c} = \{c_t\}_{t=1}^\infty$ denote a typical element. Now let $\overline{\pi}$ be the supremum of the set of present value payoffs generated by sequences $\mathbf{c} \in C_{SE}$. By definition, there must be a sequence $\{\mathbf{c}^n\}_{n=1}^\infty$ with the property that each $\mathbf{c}^n \in C_{SE}$, and moreover, $\lim_{n\to\infty} \sum_{t=1}^\infty \delta^{t-1}\pi(c_t^n, c_t^n) = \overline{\pi}$. By Lemma 2.1(i), $C_{SE} \subseteq C^*$, and C^* is sequentially compact in the sense that any sequence has a convergent subsequence (*e.g.* Jameson (1974), Theorem 11.6 and 14.6). Let $\{\mathbf{c}^{n_k}\}_{k=1}^\infty$ be a convergent subsequence of $\{\mathbf{c}^n\}_{n=1}^\infty$ with limit $\mathbf{c}^\infty \in C^*$. By $\mathbf{c}^{n_k} \in C_{SE}$, $\pi(c_{t-1}^{n_k}, c_t^{n_k})/(1-\delta) \leq \sum_{\tau=t}^\infty \delta^{\tau-t}\pi(c_\tau^{n_k}, c_\tau^{n_k})$ for all $t \geq 1$, and consequently by the continuity of the discounted payoff sum, $\pi(c_{t-1}^\infty, c_t^\infty)/(1-\delta) \leq \sum_{\tau=t}^\infty \delta^{\tau-t}\pi(c_\tau^\infty, c_\tau^\infty)$ for all $t \geq 1$. Moreover, $\mathbf{c}^\infty$ is non-decreasing. From these two facts, we have $\mathbf{c}^\infty \in C_{SE}$. Finally, by the continuity of the discounted payoff sum, $\sum_{t=1}^\infty \delta^{t-1}\pi(c_t^\infty, c_t^\infty) = \overline{\pi}$. So, the supremum can be achieved by an equilibrium path; consequently, $\mathbf{c}^\infty$ must be an efficient equilibrium path.

(b) We refer to (2.2) holding at t as the *t-constraint*. To show that all the t-constraints hold with equality, suppose to the contrary that for some t, $\pi(\widehat{c}_{t-1}, \widehat{c}_t)/(1-\delta) < \sum_{\tau=t}^\infty \delta^{\tau-t}\pi(\widehat{c}_\tau, \widehat{c}_\tau)$. Let $\tau \geq t$ be the first integer greater than or equal to t such that *either* $\widehat{c}_\tau < \widehat{c}_{\tau+1}$ *or* that the $\tau+1$-constraint holds with equality. There must exist such a τ. For suppose not: then $\widehat{c}_s = \widehat{c}_t$ for all $s > t$, in which case the $t+1$-constraint holds with equality, a contradiction. Moreover, note that as τ exists, the τ-constraint always holds with a strict inequality. Thus, there are two possibilities at τ.

1. $\widehat{c}_\tau = \widehat{c}_{\tau+1}$, and the $\tau+1$-constraint holds with equality. In this case, we establish a contradiction. Note that

$$\sum\nolimits_{s=\tau}^\infty \delta^{s-\tau}\pi(\widehat{c}_s, \widehat{c}_s) > \frac{\pi(\widehat{c}_{\tau-1}, \widehat{c}_\tau)}{(1-\delta)} \geq \frac{\pi(\widehat{c}_\tau, \widehat{c}_{\tau+1})}{(1-\delta)} = \sum\nolimits_{s=\tau+1}^\infty \delta^{s-\tau-1}\pi(\widehat{c}_s, \widehat{c}_s),$$

where the first inequality follows from the τ-constraint holding with inequality and the second inequality follows from $\widehat{c}_{\tau-1} \leq \widehat{c}_\tau = \widehat{c}_{\tau+1}$. Noting that the first term on the left is $\pi(\widehat{c}_\tau, \widehat{c}_\tau) + \delta \sum_{s=\tau+1}^\infty \delta^{s-\tau-1}\pi(\widehat{c}_s, \widehat{c}_s)$, we have $\pi(\widehat{c}_\tau, \widehat{c}_\tau) > (1-\delta)\sum_{s=\tau+1}^\infty \delta^{s-\tau-1}\pi(\widehat{c}_s, \widehat{c}_s)$, which is impossible given that $\pi(\widehat{c}_\tau, \widehat{c}_\tau) \leq \pi(\widehat{c}_s, \widehat{c}_s)$ for all $s \geq \tau+1$, due to $\widehat{c}_s$ being a non-decreasing sequence bounded above by c^*.

2. $\widehat{c}_\tau < \widehat{c}_{\tau+1}$. In this case, we also establish a contradiction. Consider a small increase in $\widehat{c}_\tau$ to $\widehat{c}_\tau + \varepsilon$, holding $\widehat{c}_s$, $s \neq \tau$ fixed. As the τ-constraint holds with strict inequality, by continuity, this increase does not violate the τ-constraint for ε sufficiently small. Moreover, (i) the t-constraints, $t < \tau$, are relaxed by an increase in $\widehat{c}_\tau$, holding $\widehat{c}_{\tau-1}, \widehat{c}_{\tau-2}, \ldots, \widehat{c}_1$ fixed since the only effect of an increase in $\widehat{c}_\tau$ is to increase the R.H.S. of these constraints; (ii) the $\tau+1$-constraint is relaxed by an increase in $\widehat{c}_\tau$, holding $\widehat{c}_{\tau+1}, \widehat{c}_{\tau+2}, \ldots$ fixed, as π is decreasing in its first argument; (iii) all t-constraints with $t > \tau+1$ are unaffected. So, the path $\{\widehat{c}_1, \ldots, \widehat{c}_{\tau-1}, \widehat{c}_\tau + \varepsilon, \widehat{c}_{\tau+1} \ldots\}$ is *also* an equilibrium path which, moreover, yields each player a higher payoff than $\{\widehat{c}_t\}_{t=1}^\infty$, contradicting the assumed efficiency of $\{\widehat{c}_t\}_{t=1}^\infty$. ‖

Proof of Lemma 2.3. It suffices to prove the upper envelope property, as there cannot be more than one such envelope. Suppose to the contrary there exists a $\{c'_t\}_{t=1}^\infty$ in C_{SE} with $c'_t > \widehat{c}_t$ for some t. Define for all $t \geq 0$, $\tilde{c}_t = \max\{\widehat{c}_t, c'_t\}$. It is clear from Assumption A1 and Lemma 2.1 (i) that $\pi(\tilde{c}_t, \tilde{c}_t) \geq \pi(\widehat{c}_t, \widehat{c}_t)$, all t, with at least one strict inequality, so that $\{\tilde{c}_t\}_{t=1}^\infty$ gives both agents a higher present-value payoff than $\{\widehat{c}_t\}_{t=1}^\infty$. So, if we can show that $\{\tilde{c}_t\}_{t=1}^\infty$ is an equilibrium path, this will contradict the assumed efficiency of $\{\widehat{c}_t\}_{t=1}^\infty$ and the result is then proved.

Say the sequences $\{\widehat{c}_t\}_{t=1}^\infty$, $\{c'_t\}_{t=1}^\infty$ have *a crossing point at* τ if $c'_{\tau-1} \leq \widehat{c}_{\tau-1}$, $c'_\tau \geq \widehat{c}_\tau$ with at least one strict inequality, or $c'_{\tau-1} \geq \widehat{c}_{\tau-1}$, $c'_\tau \leq \widehat{c}_\tau$ with at least one strict inequality. Also, define $S_t = \sum_{\tau=t}^\infty \delta^{\tau-t}\pi(c_\tau, c_\tau)$, so that $\tilde{S}_t \geq \widehat{S}_t, S'_t$ by the definition of $\tilde{c}_t$, for all t. There are then two possibilities at any time τ for the sequences $\{\widehat{c}_t\}_{t=1}^\infty, \{c'_t\}_{t=1}^\infty$.

(i) No crossing point at τ. Then, either $(\tilde{c}_{\tau-1}, \tilde{c}_\tau) = (\widehat{c}_{\tau-1}, \widehat{c}_\tau)$ or $(\tilde{c}_{\tau-1}, \tilde{c}_\tau) = (c'_{\tau-1}, c'_\tau)$. Without loss of generality, assume the former. As $\{\widehat{c}_t\}_{t=1}^\infty$ is an equilibrium path, we have $\pi(\widehat{c}_{\tau-1}, \widehat{c}_\tau)/(1-\delta) \leq \widehat{S}_\tau$, so that $(\tilde{c}_{\tau-1}, \tilde{c}_\tau) = (\widehat{c}_{\tau-1}, \widehat{c}_\tau)$ and $\tilde{S}_\tau \geq \widehat{S}_\tau$ together imply $\pi(\tilde{c}_{\tau-1}, \tilde{c}_\tau)/(1-\delta) \leq \tilde{S}_\tau$, *i.e.* the τ-constraint is satisfied for $\{\tilde{c}_t\}_{t=1}^\infty$.

(ii) A crossing point at τ. Assume w.l.o.g. that

$$c'_{\tau-1} \leq \widehat{c}_{\tau-1}, \; c'_\tau \geq \widehat{c}_\tau. \tag{A.1}$$

Then as $\{c'_t\}_{t=1}^\infty$ is an equilibrium sequence, $\pi(c'_{\tau-1}, c'_\tau)/(1-\delta) \leq S'_\tau$. Also, $\tilde{S}_\tau \geq S'_\tau$ and from (A.1), $\tilde{c}_\tau = c'_\tau$. Consequently, $\pi(c'_{\tau-1}, \tilde{c}_\tau)/(1-\delta) \leq \tilde{S}_\tau$. Finally, again from (A.1), $c'_{\tau-1} \leq \widehat{c}_{\tau-1} = \tilde{c}_{\tau-1}$. Using this fact, plus π decreasing in its first argument, we have $\pi(\tilde{c}_{\tau-1}, \tilde{c}_\tau) \leq \pi(c'_{\tau-1}, \tilde{c}_\tau)$, so we conclude $\pi(\tilde{c}_{\tau-1}, \tilde{c}_\tau)/(1-\delta) \leq \tilde{S}_\tau$, *i.e.* the τ-constraint holds for $\{\tilde{c}_t\}_{t=1}^\infty$.

So in either case (i) or (ii), all τ-constraints hold for the sequence $\{\tilde{c}_t\}_{t=1}^\infty$, so it is an equilibrium path, as required. ‖

Proof of Lemma 2.4. *Necessity.* Equation (2.3) can be written $\pi(c_{t-1}, c_t)/(1-\delta) = S_{t+1}$, $t \geq 1$, where we again write $S_t := \pi(c_t, c_t) + \delta\pi(c_{t+1}, c_{t+1}) + \ldots$. Advancing by one period, we get $\pi(c_t, c_{t+1}) = S_{t+1}$. Also,

$S_t = \pi(c_t, c_t) + \delta S_{t+1}$ by definition. So, combining these equations, we get

$$\frac{\pi(c_{t-1}, c_t)}{1-\delta} = \pi(c_t, c_t) + \frac{\delta\pi(c_t, c_{t+1})}{1-\delta}, \qquad t \geq 1. \tag{A.2}$$

Rearrangement of (A.2) gives the difference equation (2.4). Moreover, since $\{c_t\}_{t=1}^{\infty}$ is non-decreasing, it satisfies the irreversibility conditions (2.1), and since (2.3) implies (2.2), $\{c_t\}_{t=1}^{\infty}$ is an equilibrium sequence and thus by Lemma 2.1, $\{c_t\}_{t=1}^{\infty}$ must converge to $c_\infty \leq c^*$, and so must be a bounded solution to (2.4).

Sufficiency. As just shown above, (2.4) is equivalent to (A.2). By successive substitution using (A.2), we get

$$\frac{\pi(c_{t-1}, c_t)}{1-\delta} = \pi(c_t, c_t) + \cdots + \delta^{n-1}\pi(c_{t+n-1}, c_{t+n-1}) + \frac{\delta^n\pi(c_{t+n-1}, c_{t+n})}{1-\delta}. \tag{A.3}$$

Now, as $\{c_t\}_{t=1}^{\infty}$ converges by assumption, we must have $\lim_{n\to\infty} \delta^n\pi(c_{t+n-1}, c_{t+n})/(1-\delta) = 0$. So, taking the limit in (A.3), we recover (2.3). Finally, if $c_{t-1} \leq c_t$, the term in square braces in (2.4) is nonnegative, as π is decreasing in its first argument. So, we have $\pi(c_t, c_{t+1}) \geq \pi(c_t, c_t)$, implying $c_{t+1} \geq c_t$, as π is increasing in its second argument. So, by induction, the solution to (2.4) is non-decreasing given $c_1 \geq c_0$. ‖

Proof of Lemma 2.5. (i) Lemma 2.4 implies that $\{c_t(c_1; \delta)\}_{t=1}^{\infty}$ is non-decreasing and solves (2.3), which in turn implies that it is an equilibrium path. (ii) From Lemma 2.2 and Lemma 2.4, the efficient path exists, solves (2.4) with initial conditions $c_0 = 0$, $c_1 \geq 0$ and must also converge. Consequently, $\{\widehat{c}_t\}_{t=1}^{\infty} = \{c_t(\widehat{c}_1; \delta)\}_{t=1}^{\infty}$ for some $\widehat{c}_1 \in C_1(\delta)$. Now suppose that there exists another $c_1' \in C_1(\delta)$ with $c_1' > \widehat{c}_1$. In this case, $\{c_t(c_1'; \delta)\}_{t=1}^{\infty}$ is an equilibrium (by part (i)), and by construction, $c_1(c_1'; \delta) > c_1(\widehat{c}_1; \delta)$ which contradicts Lemma 2.3. So, $c_1 > c_1'$, all $c_1' \in C_1(\delta)$. Finally, as an efficient path exists, $C_1(\delta)$ must contain its supremum, so $\hat{c}_1$ in the Lemma is well-defined. ‖

Acknowledgements. We are grateful for comments from participants at the ESRC Game Theory meeting in Kenilworth, September 1998, and also at seminars at Royal Holloway College, St Andrews University, Southampton University and the Centre for Globalisation and Regionalisation, University of Warwick. We are also particularly grateful for many helpful discussions with Carlo Perroni, for valuable comments and suggestions from Martin Cripps, and also to Daniel Seidmann, Norman Ireland, Steve Matthews, Anthony Heyes and William Walker, and to two anonymous referees for a number of suggestions. Both authors acknowledge the financial support of the ESRC Centre for Globalisation and Regionalisation, University of Warwick.

REFERENCES

ADMATI, A. R. and PERRY, M. (1991), "Joint Projects without Commitment", *Review of Economic Studies*, **58**, 259–276.

COMPTE, O. and JEHIEL, P. (1995), "On the Role of Arbitration in Negotiations" (Mimeo, C.E.R.A.S., Paris).

DEVEREUX, M. (1997), "Growth, Specialization and Trade Liberalization", *International Economic Review*, **38**, 565–585.

FERSHTMAN, C. and NITZAN, S. (1991), "Dynamic Voluntary Provision of Public Goods", *European Economic Review*, **35**, 1057–1067.

FURUSAWA, T. and LAI, L.-C. (1999), "Adjustment costs and gradual trade liberalization", *Journal of International Economics*, **49**, 333–361.

GALE, D. (2000), "Monotone Games with Positive Spillovers" (Mimeo, Department of Economics, New York University).

GHEMAWAT, P. and NALEBUFF, B. (1990), "The Devolution of Declining Industries", *Quarterly Journal of Economics*, **105**, 167–186.

JAMESON, G. J. O. (1974), *Topology and Normed Spaces* (London: Chapman and Hall).

LOCKWOOD, B. and THOMAS, J. P. (1999), "Gradualism and Irreversibility" (Warwick Economic Research Paper 550, University of Warwick).

MARX, L. and MATTHEWS, S. A. (1998), "Dynamic Voluntary Contributions to a Public Project", *Review of Economic Studies*, **67**, 327–358.

SALANT, L. and WOROCH, G. A. (1992), "Trigger Price Regulation", *RAND Journal of Economics*, **23**, 29–51.

SCHELLING, T. C. (1960), *The Strategy of Conflict* (Cambridge: Harvard University Press).

STAIGER, R. (1995), "A Theory of Gradual Trade Liberalization", in J. Levinsohn, A. V. Deardorff and R. M. Stern (eds.) *New Directions in Trade Theory* (Ann Arbor, MI: University of Michigan Press).