# Stability Implies Renegotiation Proofness

Thomas W. L. Norman*

Magdalen College, Oxford

October 12, 2020

**Abstract**

In generic two-player finitely repeated games, if a particular outcome path is induced by a set of equilibria that is stable in the sense of Mertens (*Mathematics of Operations Research* 1989, 14:575–625), then a member of that set induces a renegotiation proof equilibrium of the continuation game after the first period. It follows that such a stable outcome path yields the payoff of a renegotiation proof equilibrium as the number of repetitions of the game goes to infinity. In this sense, any pure outcome consistent with Govindan and Wilson's (*Econometrica* 2012, 80:1639–1699) axioms—Undominated Strategies, Backward Induction and Invariance to Embedding—is generically renegotiation proof. *Journal of Economic Literature* Classification: C72.

*Key Words:* Repeated games; stability; renegotiation proofness.

*Email thomas.norman@magd.ox.ac.uk.

# 1 Introduction

The concept of "renegotiation proofness" is motivated by the idea that players should "let bygones be bygones"; an equilibrium that leads to Pareto-dominated play in some subgame should at that point be renegotiated away. By this account, although potent, the punishment threatened by a "grim-trigger" strategy is incredible, since both players would have an incentive to renegotiate continuation play. Strategic stability (Kohlberg and Mertens 1986), meanwhile, offers an apparently quite different equilibrium refinement; it highlights an equilibrium set's robustness to Selten's (1975) "trembles" as key to the satisfaction of a number of desirable properties of such a refinement. Whilst Kohlberg–Mertens stability unifies numerous resolutions of troublesome behavior—such as the intuitive criterion (Cho and Kreps 1987) and universal divinity (Cho and Sobel 1990)—it nonetheless fails to satisfy connectedness and backward induction, which led Mertens (1989, 1991) to develop a modified concept without these drawbacks. Moreover, Govindan and Wilson (2012) recently achieved Kohlberg and Mertens' stated aim of giving stability a decision-theoretic axiomatization, at least in finite two-player games; they showed that, if a refinement satisfies three axioms—Undominated Strategies, Backward Induction and Invariance to Embedding—then each solution of a two-player game with perfect recall and generic payoffs is a Mertens stable set. Since the converse holds in general, stability offers a characterization of these axioms in the relevant class of games.

In this paper I show that, in generic two-player finitely repeated games, if a particular outcome path is induced by a (Mertens) stable set of equilibria, then a member of that set induces a renegotiation proof equilibrium of the continuation game after the first period. Intuitively, under an appropriate choice of stability's embeddings (Govindan and Wilson 2012), a player may use a deviation to initiate Pareto-improving play and observation of such a deviation may be interpreted as a signal of such play. Thus, continuation play that fails renegotiation proofness is destabilized in a manner reminiscent of forward induction. Osborne (1990) shows forward induction to imply near efficiency of stable payoffs in certain repeated coordination games, but the scope of his result is limited to games that lend themselves to unambiguous signalling, in the sense that they admit a deviation with a unique profitable continuation. This problem does not arise here, since the relevant embedding game may be chosen to deliver the required unambiguous signalling. Indeed,

whilst Govindan and Wilson (2009) show that a broader notion of forward induction is implied by backward induction and a weak form of invariance, forward induction alone does not imply my result, with Invariance to Embedding performing an independent role. This is further illustrated by the insufficiency of Kohlberg–Mertens stability for the result, which I demonstrate with an example of a Kohlberg–Mertens (but not Mertens) stable set that fails to induce renegotiation proof continuations.

The restriction of renegotiation proofness to subgames after the first period is important, allowing the players an opportunity to make the signalling deviation in the first period. That the result cannot be extended to the full game is shown by Van Damme's (1988) two-period example of the inconsistency of renegotiation proofness with stability.[1] However, the renegotiation proofness of second-period continuation play is enough for a stable outcome path to approach a renegotiation proof payoff as the number of iterations of the game goes to infinity. Moreover, I argue that the concept of renegotiation proofness might reasonably be redefined without first-period Pareto efficiency, since it is not called "negotiation proofness"; a pure outcome's consistency with Govindan and Wilson's (2012) axioms would then imply renegotiation proofness in generic two-player repeated games of any finite length.

**Related literature** The Folk Theorem for finitely repeated games (Benoît and Krishna 1985) establishes an inherent unpredictability in games with multiple equilibrium payoffs for each player, and yet efficiency is often thought to be focal in such games. Kohlberg and Mertens (1986) argue that single-valued solution concepts such as Nash and subgame perfect equilibrium miss important aspects of rational decision-making in games that are captured by their set-valued notion of stability. As mentioned above, the refinement possibilities of stability have previously been highlighted in certain finitely repeated coordination games, where Osborne (1990) shows that, among the set of pure outcome paths that consist of sequences of one-shot Nash equilibria, only those with nearly Pareto efficient payoffs are stable.[2] But whilst this delivers sharp predictions in certain $2 \times 2$ stage games, his results lose force when each player has more than two actions, when equilibrium paths contain

[1] See Ferreira (1995), however, for an argument that the two concepts can be reconciled even in such cases. For a more applied exploration of the possible inconsistency of stability and renegotiation in a sequential signalling game, see Gale and Hellwig (1989).

[2] Van Damme (1989) also offers examples of the sometimes dramatic effects of stability in finitely repeated games.

outcomes that are not stage Nash equilibria, and when the game is asymmetric.[3]

Whilst there are numerous competing notions of renegotiation proofness in infinitely repeated games (Farrell and Maskin 1989; Bernheim and Ray 1989; Bergin and Macleod 1993; Pearce 1987; Abreu and Pearce 1991), the concept's definition has been uncontroversial in finitely repeated games (Farrell 1983; Bernheim and Ray 1989; Van Damme 1988; Benoît and Krishna 1993). It has important applications, such as Maskin and Tirole's (1988) renegotiation proof collusive outcome in a dynamic price-setting duopoly, and MacLeod and Malcomson's (1989) analysis of employment contracts when employee performance is unverifiable. Matsuyama (1990) uses it to eliminate an unappealing equilibrium of a trade liberalization game, whilst Pearce and Stacchetti (1997) and Matsuyama (1997) conduct renegotiation proof analyses of classic macroeconomic time inconsistency problems. The concept has also received some recent attention in the form of "renegotiation proof mechanism design" (Strulovici 2017; Silva 2019).

# 2 Example

Consider the example of Osborne (1990) where two players play the game $G_1$ in Figure 1 twice. This repeated game has a subgame perfect equilibrium outcome path $((A, A), (A, A))$ with a payoff of 2 for each player. Osborne makes the following forward induction argument for the instability of this path: If player 1 deviates by playing $B$ in period 1, the only second-period outcome yielding player 1 a payoff greater than 2 is $(B, B)$. Thus, player 2 can deduce from such a deviation that player 1 will play $B$ in period 2, so that it is better for player 2 to play $B$ in period 2. Hence, player 1 can get a payoff of 3 by deviating from the equilibrium path, destabilizing it.

Osborne's (1990) results generalize this example but with some important restrictions on the game, needed to afford player 1 an unambiguous signal of his intended continuation play following a deviation; namely, that there is only one continuation

[3]In infinitely repeated games, Aumann and Sorin (1989) select the optimal outcome of a two-player game of common interests using a tremble in which every strategy with finite memory is used with positive probability. Anderlini and Sabourian (1995) obtain a similar result for trembling-hand perfect equilibrium of two-player common interest games in computable pure strategies. Norman (2018) shows that strategies with inefficient stage Nash continuations are vulnerable to experimentation with efficient play, and hence are strategically unstable for two arbitrarily patient players.

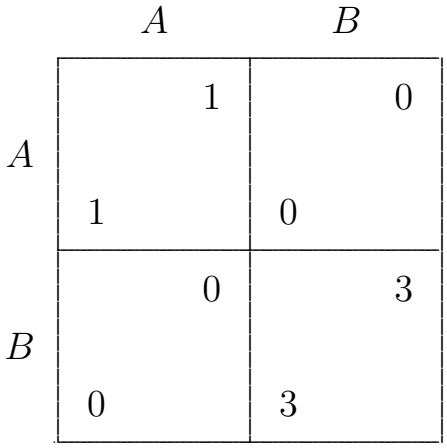


Figure 1: The game $G_1$

path yielding player 1 a higher payoff than the putative equilibrium path, and that player 2 cannot benefit from any outcome path resulting from her own deviation from the continuation. As a consequence, the results lose force when the players have more than two actions each period, when equilibrium paths contain outcomes that are not stage Nash equilibria, and when there is conflict over what is the best outcome (i.e. the game is asymmetric). Here I use a more general instability argument to remove these limitations.[4]

Consider any member $q^*$ of the component $Q^*$ of equilibria with outcome path $((A,A),(A,A))$.[5] Player $i$'s strategy $q_i^*$ plays $B$ with probability at most $1/4$ following the opponent $j \neq i$ playing $B$ in the first period. Let $p_i$ be a strategy that plays $B$ initially, then $B$ following the opponent's $A$. The strategy $q_i^*$ is a best reply to $q_j^*$, but not to $p_j$; the strategy that plays $A$ initially, then the first-period action of the opponent, is lexicographically optimal (Blume, Brandenburger, and Dekel 1991; Govindan

[4]However, a limitation of Osborne's results that remains in the present context arises when equilibrium paths involve randomization, since any completely mixed strategy equilibrium is automatically stable.

[5]The appeal of components is conferred by the Generic Finiteness Theorem (Kreps and Wilson 1982; Kohlberg and Mertens 1986; Govindan and Wilson 2001), which states that the Nash equilibria of a generic extensive-form game induce a finite number of outcome distributions over the terminal nodes. This implies that each equilibrium in a connected set generates the same outcome, so that connectedness provides a natural notion of equivalence in which equilibrium components are the unit. In this capacity, connectedness substitutes for the minimality requirement of Kohlberg–Mertens stability, which Mertens (2003, §4.2.6) shows to be inconsistent with ordinality.

and Klumpp 2002) against the sequence $(q_j^*, p_j)$. Hence, there is a deviation from equilibria in $Q^*$ against which elements of $Q^*$ are not lexicographically optimal, which provides the basis for establishing that $Q^*$ is not stable by Govindan and Wilson's (2012) Theorem 5.2. Their characterization of stability requires Invariance to Embedding the game in a range of larger games, with Backward Induction then requiring that a player's strategies must be lexicographically optimal against such sequences of opponent's strategies; here, the possibility of a player deviating to Pareto-improving play serves to destabilize the putative equilibria's off-equilibrium-path play.

But we need not stop here. Consider Osborne's (1990, §6a) coordination game $G(\alpha_1, \ldots, \alpha_m, \beta)$, where $m+1$ pure strategies $A_1, \ldots, A_m, B$ are associated with diagonal payoffs $\alpha_1, \ldots, \alpha_m, \beta$ for $\beta > \alpha_m \geq \cdots \geq \alpha_1 \geq 0$. His near-efficient stability result applies to this game only if $\alpha_1$ and $\alpha_m$ are sufficiently close, otherwise a deviation need not be an unambiguous signal of the deviator's intended continuation play (in the presence of multiple Pareto-improving continuations). But there is no reason why we may not apply the destabilizing argument from the last paragraph here, effectively disambiguating the deviator's signal with the choice of a particular Pareto-improving continuation under $p$ (and hence the relevant embedding of the game). For instance, any path $((A_k, A_k), (A_k, A_k))$, $k \in 1, \ldots, m$, is destabilized by a deviation $p_i$ that plays $B$ initially, then $B$ following the opponent's $A_k$. Thus, careful choice of stability's embeddings can serve to disambiguate signals of Pareto-improving play more generally than in Osborne (1990). In particular, it can do so whenever there exists a proper subgame in which there is a Pareto-improving continuation. The absence of such a subgame in a $T$-fold repeated game implies that continuation play after the first period is renegotiation proof; hence, as $T \to \infty$, stable outcomes approach renegotiation proofness.

# 3 The Model

The model is essentially that of Govindan and Wilson (2012), applied to the setting of finitely repeated games. Let $G = (A_1, A_2; u_1, u_2)$ be a two-player normal-form game, where $A_i$ is player $i$'s finite action set and $u_i : A_1 \times A_2 \to \mathbb{R}$ is his payoff function. Let $u \equiv u_1 \times u_2$ be the *payoffs* and $A \equiv A_1 \times A_2$ the set of *outcomes* of $G$. Let $G(T)$ be the game where $G$ is played $T$ times in succession. An *outcome path* $\vec{a} = (a^1, a^2, \ldots, a^T)$ yields the payoffs $\sum_{t=1}^{T} u(a^t)$. For $K < T$, let $h^K = (a^1, a^2, \ldots, a^K)$ be a $K$-period

*history* of play, belonging to the set of all $K$-period histories $H^K$, with $H \equiv \bigcup_{K=1}^{T} H^K$. Player $i$'s *pure strategy* $s_i$ is a function mapping each history in $H$ to an element of $A_i$. A strategy profile $s \equiv s_1 \times s_2$ induces an outcome $a^t(s)$ for each $t$, and payoffs $U(s) = \sum_{t=1}^{T} u(a^t(s))$. Let $s|_{h^K}$ be the pure strategy profile induced by $s$ on the subgame $G(T-K)$ following $h^K$. A strategy of player $i$ that follows the outcome path $\vec{a}$ so long as player $j \neq i$ does so is *consistent with* $\vec{a}$.

Denote $i$'s simplex of *mixed strategies* by $\Sigma_i$, the vertices of which constitute the set $S_i$ of $i$'s pure strategies, and extend each $U_i$ to be $i$'s multilinear expected payoff function on $\Sigma \equiv \Sigma_1 \times \Sigma_2$. Each mixed strategy $\sigma_i$ induces a *behavioral strategy* $b_i$, specifying a mixture over $i$'s actions following each history. Let $B_i$ be $i$'s set of behavioral strategies. Assume that the game $G(T)$ has perfect monitoring; hence, each behavioral strategy profile is induced by (and yields the same distribution of outcomes as) certain equivalent profiles of mixed strategies (Kuhn 1953).

*Renegotiation proofness* (Farrell 1983; Van Damme 1988; Bernheim and Ray 1989; Benoît and Krishna 1993) is defined inductively in finitely repeated games: A strategy profile is renegotiation proof in $G(1)$ if it is a Nash equilibrium of $G$ that is not (strictly) Pareto dominated by another equilibrium. The pure strategy profile $s$ is then renegotiation proof in $G(T+1)$ if:

i. it is a Nash equilibrium of $G(T+1)$;

ii. $s|_{h^1}$ is renegotiation proof in $G(T)$ for all $h^1 \in H^1$; and

iii. there does not exist $\hat{s}$ satisfying i and ii that strictly Pareto dominates $s$.

Clearly a renegotiation proof strategy profile is a subgame perfect equilibrium.

Let $X$ be the set of *nodes* in the extensive form of $G(T)$, and $X_i$ the set of nodes where $i$ moves (partitioned into his information sets). For any $x \in X$, let $h(x)$ be the history preceding $x$. For each $i$, $s_i \in S_i$ and $y \in X$, let $\beta_i(y, s_i)$ be the probability that $s_i$ *does not exclude* $y$—i.e. $\beta_i(y, s_i) = 1$ if $s_i$ plays in accordance with $h(y)$, and $\beta_i(y, s_i) = 0$ otherwise. The function $\beta_i$ may be extended to a function over mixed and behavioral strategy profiles in an obvious manner. Letting $Z \subset X$ be the set of *terminal nodes* in the extensive form of $G(T)$, for each $i$ define $\rho_i : \Sigma_i \to [0,1]^Z$ by the formula $\rho_i(\sigma_i) = (\beta_i(z, \sigma_i))_{z \in Z}$, and let $\rho \equiv \rho_1 \times \rho_2$. Then the space $P_i$ of $i$'s *enabling strategies* (Govindan and Wilson 2002) is the image of $\rho_i$ and the space $P \equiv P_1 \times P_2$ of enabling strategy profiles is the image of $\rho$. To each vertex of $P_i$

corresponds an equivalence class of $i$'s pure strategies that exclude the same outcome paths.[6] As Govindan and Wilson (2012, §4.3) note, enabling strategies are sufficient representations for extensive-form games of perfect recall.

A connected closed set of equilibria is *stable* (Mertens 1989, 1991) if the local projection map, from a connected closed neighborhood in the graph of equilibria over the space of players' strategies perturbed toward mixed strategies, is "homologically nontrivial". As a consequence: its image does not deform to a point in the tremble space (it is not "null-homotopic"); it is essential (it has a point of coincidence with every continuous map having the same domain and range); and it has nonzero degree (the number of times each tremble is covered by the projection map). Since a map that is not surjective has zero degree, it follows that there is a Nash equilibrium at each point along any continuous trajectory of perturbations in the relevant neighborhood, and hence a stable set is also strategically stable in the sense of Kohlberg and Mertens (1986). The converse does not apply, however, because a surjective map need not have nonzero degree, nor be null-homotopic. The latter prevents the application of Brouwer's theorem to a fixed point problem solved by proper equilibria (see Mertens 1989, Theorem 6), so that Kohlberg–Mertens stable sets need not contain such equilibria, and hence may fail backward induction (Van Damme 1984). Stability corrects this failure, at the cost of a slightly opaque definition for an economics audience—a defect that is alleviated by Govindan and Wilson's (2012) equivalent formulation in terms of lexicographic beliefs, which appears in Lemma 1 and for my purposes can be taken as definitional of stability.

A *component* of equilibria is a maximal closed connected set of equilibria in $\Sigma$. Membership of such a component provides a natural notion of equivalence of equilibria (see, e.g., Kohlberg and Mertens 1986, §2.8, Van Damme 1987, p. 271)—it generically implies a common payoff vector, with strategic differences occurring only off the equilibrium path (Kreps and Wilson 1982; Kohlberg and Mertens 1986; Govindan and Wilson 2001).[7] Let $\bar{\Sigma}^*$ be a component of the equilibria of $G(T)$ in mixed strategies (and the sets $\bar{B}^*$ and $\bar{P}^*$ the equivalent behavioral and enabling strategies). All

[6] These are the pure strategies in Mailath, Samuelson, and Swinkels's (1993) "pure reduced normal form".

[7] Whilst Kohlberg–Mertens stable sets need not be connected, there does always exist a stable set that belongs to a single component of equilibria (Kohlberg and Mertens 1986, p. 1027), and connectedness is satisfied by Mertens stability and by Govindan and Wilson's (2012) axiomatic solution.

equilibria in these sets generically induce the same outcome path, so that for each node $x \in X$ the probability $\beta_i(x, \bar{b}^*)$ of reaching $x$ in an equilibrium $\bar{b}^* \in \bar{B}^*$ is the same for all $\bar{b}^* \in \bar{B}^*$; let $\beta_i^*(x)$ be this probability. Now let $X_i^*$ be the collection of nodes $x \in X_i$ such that $\beta_i^*(x) > 0$ (i.e. that are not excluded by elements of $\bar{\Sigma}_i^*$), and let $A_i^*$ be the set of actions that are chosen with positive probability at nodes in $X_i^*$ by the equilibria in $\bar{B}^*$. Let $S_i^0 \subset S_i$ be the set of pure strategies $s_i^0$ with the property that, at each node $x_i \in X_i^*$ that $s_i^0$ does not exclude, $s_i^0$ prescribes an action in $A_i^*$. Let $S_i^1 \equiv S_i \backslash S_i^0$—i.e. each pure strategy $s_i$ in $S_i^1$ chooses a nonequilibrium action at some node $x_i \in X_i^*$ that it does not exclude.

For $k = 0, 1$, let $\Sigma_i^k$ be the set of mixed strategies whose support is contained in $S_i^k$. Note that the support of $i$'s strategy in every equilibrium in $\bar{\Sigma}^*$ is contained in $S_i^0$ and that every strategy in $S_i^0$ is a best reply to every equilibrium in $\bar{\Sigma}^*$; hence, $\bar{\Sigma}_i^* \subseteq \Sigma_i^0$. Now let $\Sigma^*$ be a component of the undominated equilibria in $\bar{\Sigma}^*$; then $\Sigma_i^*$ is a connected component of the intersection of $\bar{\Sigma}_i^*$ with the set of undominated strategies. Let $Q^*$ be the image of $\Sigma^*$ under $\rho$. Let $P_i^k$ be $i$'s set of enabling strategies in the image of $\Sigma_i^k$ for $k = 0, 1$, let $P^k \equiv P_1^k \times P_2^k$, and define $\mathbb{P} \equiv P^0 \times P^1$. Let $Z_i^1 \subset Z$ be the set of terminal nodes $z$ such that $h(z)$ contains an action profile with some $a \notin A_i^*$. Then $P_i^0$ is the set of $p_i \in P_i$ such that $p_i(z) = 0$ for all $z \in Z_i^1$ and thus $P_i^0$ (but not necessarily $P_i^1$) is a face of $P_i$.

Given $p_i \in P_i$, let $\psi_{Z_i^1}(p_i)$ be the projection of $p_i$ to $\mathbb{R}_+^{Z_i^1}$; then $\psi_{Z_i^1}(p_i) = 0$ if and only if $p_i \in P_i^0$. Fixing a point $\bar{p}_j$ in the interior of $P_j$, $j \neq i$, and defining $\eta_i : P_i \to \mathbb{R}$ by $\eta_i(p_i) = \sum_{z \in Z_i^1} \bar{p}_j(z) p_i(z)$, clearly $\eta(p_i) = 0$ if and only if $p_i \in P_i^0$. Choose $\varepsilon > 0$ such that $\eta_i(p_i) > \varepsilon$ for all $p_i \in P_i^1$, and let $\mathscr{H}_i$ be the hyperplane in $\mathbb{R}^{Z_i^1}$ with normal $(\bar{p}_j(z))_{z \in Z_i^1}$ and constant $\varepsilon$, which separates the origin from $\psi_{Z_i^1}(P_i^1)$. Let $\Pi_i^1$ be the intersection of $\mathscr{H}_i$ with $\psi_{Z_i^1}(P_i)$, and $\bar{\pi}_i^1$ be the function from $P_i \backslash P_i^0$ to $\Pi_i^1$ that maps each $p_i \in P_i^0$ to the point $\varepsilon(\eta_i(p_i))^{-1} \psi_{Z_i^1}(p_i)$.

Player $i$'s strategy $\sigma_i$ is *lexicographically optimal* (Blume, Brandenburger, and Dekel 1991; Govindan and Klumpp 2002) against a sequence $(\sigma_j^n)_{n=1,2,\ldots}$ of his opponent's strategies if any alternative strategy $\hat{\sigma}_i$ that is a better reply to $\sigma_j^n$ for some $n$ is a worse reply to $\sigma_j^m$ for some $m < n$. Given $Q^*$, let $\mathscr{Q}$ be the set of those $(q^*, (p^0, p^1), \pi^1) \in Q^* \times \mathbb{P} \times \Pi^1$ such that there exist $r^0, \tilde{p}^0 \in P^0$, and $r^1 \in P^1$, and for

each $i$, scalars $\lambda_i^0$, $\lambda_i^1$, and $\mu_i^1$ in the interval $(0,1]$ such that, if

$$\begin{aligned} q_i^0 &= \lambda_i^0 p_i^0 + (1-\lambda_i^0) r_i^0 \quad \text{and} \\ q_i^1 &= (1-\lambda_i^1)\tilde{p}_i^0 + \lambda_i^1 \left(\mu_i^1 p_i^1 + (1-\mu_i^1) r_i^1\right), \end{aligned} \tag{1}$$

then for each $i$,

i. $\bar{\pi}_i^1(q_i^1) = \pi_i^1$;

ii. $q_i^0$, and $r_i^0$ if $\lambda_i^0 < 1$, are lexicographically optimal replies against $(q_j^*, q_j^0, q_j^1)$;

iii. if $\mu_i^1 < 1$, then $r_i^1$ is an optimal reply against $q_j^*$ and lexicographically as good a reply against $(q_j^*, q_j^0, q_j^1)$ as other strategies in $P_i^1$.

The set $\mathscr{Q}$ is the graph of lexicographically optimal replies to possible deviations from equilibria in $Q^*$.[8]

Let $\psi : \mathscr{Q} \to \mathbb{P}$ be the natural projection, $\psi(q^*, (p^0, p^1), \pi^1) = (p^0, p^1)$, and $\partial\mathscr{Q} \equiv \psi^{-1}(\partial\mathbb{P})$; then $\psi$ is *essential* if, for every continuous map $\phi : \mathscr{Q} \to \mathbb{P}$ there exists some $q \in \mathscr{Q}$ such that $\phi(q) = \psi(q)$. Note that this corresponds to the concept of *essentiality in homotopy*, in the sense of not being null-homotopic (see Govindan and Wilson 2008, Lemmas A.3 and A.4, and Mertens 1991, lemma on p. 597).

**Lemma 1 (Govindan and Wilson 2012)** *$Q^*$ is stable if and only if the projection map $\psi : (\mathscr{Q}, \partial\mathscr{Q}) \to (\mathbb{P}, \partial\mathbb{P})$ is essential.*

The players' payoffs are given by a point $U$ in $\mathscr{U} = \mathbb{R}^2 \times Z$, where $U_i(z)$ is the payoff to player $i$ at terminal node $z \in Z$. I assume that payoffs are *generic* in the sense that there exists a lower-dimensional subset $\mathscr{U}_\circ$ of $\mathscr{U}$ such that the results are true for all repeated games in $\mathscr{U} \setminus \mathscr{U}_\circ$. Govindan and Wilson (2012, Theorem 5.1) show that, if a refinement satisfies three axioms—Undominated Strategies, Backward Induction and Invariance to Embedding—then each solution of a two-player game with perfect recall and generic payoffs is a stable set, and indeed an essential component of admissible equilibria. If a stable set induces a particular outcome path $\vec{a}$ with probability 1, call $\vec{a}$ a *stable outcome path*.

[8] For the case where $S_i^1$ is empty for some $i$ (which does not arise for Theorem 1 below) see Govindan and Wilson (2012, pp. 1655–7).

**Theorem 1** *If a stable set $\Sigma^*$ of a generic $G(T)$ induces a stable outcome path, then $\Sigma^*$ contains an equilibrium that induces a renegotiation proof equilibrium in $G(T-1)$.*

**Proof.** Suppose otherwise; then there exists a stable outcome path $\vec{a}$ induced by $\Sigma^*$. Moreover, for each $\sigma^* \in \Sigma^*$ there exists a $K < T$ and a nonempty history $h^K \in H^K$ such that either: (a) $\sigma^*|_{h^K}$ is not a Nash equilibrium of $G(T-K)$; or (b) there exists a subgame perfect equilibrium of $G(T-K)$ that strictly Pareto dominates $\sigma^*|_{h^K}$. Since a stable set satisfies backward induction, there exists some $\tilde{\sigma}^* \in \Sigma^*$ for which (b) holds with dominating equilibrium $\hat{s}^*$ following history $h^{\widetilde{K}}$. If any $\sigma^* \in \Sigma^*$ places probability 1 on $\hat{s}^*$ following $h^{\widetilde{K}}$, then there exists a subgame perfect equilibrium that also does so; hence, $\tilde{\sigma}^*$, $h^{\widetilde{K}}$ and $\hat{s}^*$ may be chosen such that no $\sigma^* \in \Sigma^*$ places probability 1 on $\hat{s}^*$ following $h^{\widetilde{K}}$. Moreover, if $\vec{a}$ is a continuation of $h^{\widetilde{K}}$, then there exists another $\widetilde{K}$-length history of which $\vec{a}$ is not a continuation and following which $\tilde{\sigma}^*$ is also dominated by $\hat{s}^*$; hence, we may suppose that $\vec{a}$ is not a continuation of $h^{\widetilde{K}}$.

Now let $p_i^0$ be a strategy that prescribes the same action mixture as $\tilde{q}_i^*$ at any node that $\tilde{q}_i^*$ does not exclude. Meanwhile, let $p_i^1$ be a strategy that plays according to $h^{\widetilde{K}}$, then $\hat{s}_i^*$ following $i$'s own adherence to $h^{\widetilde{K}}$. For any $q^* \in Q^*$, $r^0, \tilde{p}^0 \in P^0$, $r^1 \in P^1$ and $\lambda_i^0, \lambda_i^1, \mu_i^1 \in (0,1]$ satisfying iii on the previous page, the strategy $q_i^0$ is a best reply to both $q_j^*$ and $q_j^0$, but not to $p_j^1$ (and hence $q_j^1$, since any $r_j^1$ optimal against $q_i^*$ does not play according to $h^{\widetilde{K}}$ for generic payoffs) because $q_i^0$ does not induce $\hat{s}^*$ with probability 1 following $h^{\widetilde{K}}$. The strategy consistent with $\vec{a}$ that plays $\hat{s}_i^*$ following the opponent $j$'s adherence to $h^{\widetilde{K}}$ is a best reply against $q_j^*$ and $q_j^0$, and a better reply than $q_i^0$ against $p_j^1$ (and hence against $q_j^1$). Thus, $\psi^{-1}(p^0, p^1) \notin \mathscr{Q}$. Now, whilst maps that are essential in homotopy need not be surjective, (co)homologically essential maps must be so, and the two notions of essentiality are equivalent in this case by a theorem of Mertens (1989, pp. 704–5) and Govindan and Wilson's (2012, Theorem 5.2) result that the domain and codomain of $\psi$ are of equal dimension. Therefore, $\psi$ is not essential in homotopy, and the result follows by Lemma 1. ∎

**Corollary 1** *Any stable outcome path of a generic $G(T)$ yields payoffs approaching those from a renegotiation proof equilibrium as $T \to \infty$.*

Intuitively, if $\Sigma^*$ does not contain a strategy profile inducing renegotiation proof continuations, then there exists a possible deviation that initiates Pareto-improving play, against which none of $\Sigma^*$'s strategies is lexicographically optimal. This creates

a "hole" in the graph $\mathscr{Q}$, so that the projection map $\psi$ cannot be essential. The assumption that $\Sigma^*$ induces some outcome path with probability 1 is both necessary for $\Sigma^*$ to contain a renegotiation proof equilibrium (which is defined to be a pure strategy equilibrium), and sufficient for the existence of an off-equilibrium path first-period action, with which either player may signal Pareto-improving continuation play. Moreover, any completely mixed strategy equilibrium is automatically stable (the case of $S_i^1$ empty for both $i$), and clearly need not induce renegotiation proof continuations.

# 4 Discussion

## 4.1 The necessity of Mertens stability

A natural question to arise is whether Theorem 1 would fail under Kohlberg and Mertens' (1986) original notion of stability. Consider the two-fold repetition of the game $G_2$ in Figure 2, and in particular the equilibrium path $((U, L), (D, R))$.[9] Any equilibrium inducing this path must fail renegotiation proofness, as it must deter a first-period deviation to $M$ with second-period play that is strictly Pareto dominated by $(D, R)$. However, there is a Kohlberg–Mertens stable set that induces this path. To see this, suppose first that strategies are perturbed such that, following first-period play of $(M, L)$, the tremble (i.e. involuntary) probabilities $(\pi_U, \pi_M, \pi_D)$ on the actions $U$, $M$ and $D$ satisfy $\pi_M \geq \max\{\pi_U, 2\pi_D\}$. Then $C$ is a best response for Bob, deterring Ann's first-period deviation to $M$.

Alternatively, suppose that the strategy perturbations are such that $\pi_M < \max\{\pi_U, 2\pi_D\}$. Then if Bob mixes with probabilities $(1/2, 1/2)$ on actions $C$ and $R$ following $(M, L)$, Ann is indifferent between playing $U$ then $D$ unconditionally and playing $M$ both in the first period and following $(M, L)$. And if Ann mixes with probabilities $(\delta, 1-\delta)$ on these two strategies, then there exist constants $\varpi_1, \varpi_2, \varpi_3, \varpi_4 > 0$ such that Bob is indifferent between $C$ and $R$ following $(M, L)$ if and only if

$$\delta = \frac{(2(1-\varpi_4) - \pi_M)\varpi_2}{(1-\varpi_1)(1-\pi_U-\pi_M-\pi_D+2(1-\varpi_3-\varpi_4)) - (2(1-\varpi_4)-\pi_M)(1-\varpi_1-\varpi_2)}.$$

The numerator here tends to 0 from above as the strategy perturbations vanish, and

[9] I am grateful to an anonymous referee for suggesting this example.

| | $L$ | $C$ | $R$ |
|---|---|---|---|
| $U$ | 1<br>7 | 0<br>0 | 0<br>0 |
| $M$ | 0<br>9 | 1<br>2 | 0<br>4 |
| $D$ | 0<br>0 | 0<br>0 | 2<br>5 |

Figure 2: The game $G_2$

the denominator to 1, so $\delta$ tends to 0 from above. Bob's continuation payoff following $(M, L)$ approaches 1 in this limit, so he will not deviate to $L$.

Thus, the path $((U, L), (D, R))$ is a Kohlberg–Mertens stable outcome. Note, however, that it is not a Mertens stable outcome, as Bob's play is not lexicographically optimal against a sequence of equilibrium strategies that induce the equilibrium path, followed by a strategy that deviates to $M$ in the first period and then plays $D$.

## 4.2 Forward induction

The deviating strategies used in the proof of Theorem 1 are reminiscent of those used to make forward induction arguments, and indeed Osborne's (1990) Proposition 2 is implied by Theorem 1. The idea of forward induction is that "players assume, even if they see something unexpected, that the other players chose rationally in the past" (Hillas and Kohlberg 2002, §42.13.6), which motivates van Damme's (1989) requirement that a deviation be followed by a unique profitable continuation in order to allow forward induction, since only then is an unambiguous signal sent through the act of deviating. If we think back to Osborne's coordination game $G(\alpha_1, \ldots, \alpha_m, \beta)$ (discussed in Section 2 above), with its multiple Pareto-improving continuations following a deviation from $(A_1, A_1)$ for instance, it is clear that this narrow notion of forward induction will not suffice for renegotiation proof continuations.

But nor will Govindan and Wilson's (2009) broader forward induction requirement that the players believe their opponents to use only relevant optimal strategies (as opposed to optimal continuation play). To see this, note that the path $((A_1, A_1), (A_1, A_1))$ in the twice-repeated game $G(1, 3, 4)$ is consistent with forward induction, enforced by a belief that the opponent will play $A_2$ with probability $4/7$ and $B$ with probability $3/7$ after observing a first-period deviation. Indeed, for generic two-player games, Govindan and Wilson show their notion of forward induction to be a consequence of invariance to redundant strategies and backward induction, and hence of both Mertens stability and the original Kohlberg–Mertens stability, the latter of which was shown above to be insufficient for renegotiation proofness.[10]

Al-Najjar (1995) provides another forward induction concept that allows multiple Pareto-improving continuations, and which is sufficient for renegotiation proof continuations. However, this concept is implied by neither Kohlberg–Mertens nor Mertens stability, as is clear from its existence problems in games such as the repeated Battle of the Sexes. The proof of Theorem 1 avoids the problem of ambiguous signalling in games like $G(\alpha_1, \ldots, \alpha_m, \beta)$ by careful choice of the deviation $p^1$, and hence of the implicit embedding of the game to which a Mertens stable set must be robust. Thus, Invariance to Embedding plays a role over and above (invariance to redundant strategies and) forward induction, by coordinating play off the equilibrium path.

## 4.3 A redefinition

Given the limitation of stable renegotiation proofness to continuation play after the first period, perhaps renegotiation proofness should be redefined with this in mind; it is, after all, not called "negotiation proofness". In particular, a strategy profile could be considered renegotiation proof in $G(1)$ if it is a Nash equilibrium of $G$. The pure strategy profile $s$ would then be renegotiation proof in $G(T+1)$ if:

i. it is a Nash equilibrium of $G(T+1)$;

ii. $s|_{h^1}$ is renegotiation proof in $G(T)$ for all $h^1 \in H^1$; and

iii. there does not exist $\hat{s}$ satisfying i and ii and $h^1 \in H^1$ such that $\hat{s}|_{h^1}$ strictly Pareto dominates $s|_{h^1}$.

[10] Whilst Kohlberg–Mertens stability does not satisfy backward induction in general, it does so in two-player games (Govindan and Wilson 2006).

With this alternative definition, Theorem 1 would yield a stable renegotiation proof equilibrium in $G(T)$, rather than $G(T-1)$.

# References


Abreu, D., and D. Pearce. 1991. "A Perspective on Renegotiation in Repeated Games". In *Game Equilibrium Models*, ed. by R. Selten, 2:44–55. Berlin Heidelberg: Springer-Verlag.

Al-Najjar, N. 1995. "A Theory of Forward Induction in Finitely Repeated Games". *Theory and Decision* 38:173–193.

Anderlini, L., and H. Sabourian. 1995. "Cooperation and Effective Computability". *Econometrica* 63:1337–1369.

Aumann, R. J., and S. Sorin. 1989. "Cooperation and Bounded Recall". *Games and Economic Behavior* 1:5–39.

Benoît, J.-P., and V. Krishna. 1985. "Finitely Repeated Games". *Econometrica* 53:905–922.

— . 1993. "Renegotiation in Finitely Repeated Games". *Econometrica* 61:303–323.

Bergin, J., and W. B. Macleod. 1993. "Efficiency and Renegotiation in Repeated Games". *Journal of Economic Theory* 61:42–73.

Bernheim, B. D., and D. Ray. 1989. "Collective Dynamic Consistency in Repeated Games". *Games and Economic Behavior* 1:295–326.

Blume, L., A. Brandenburger, and E. Dekel. 1991. "Lexicographic Probabilities and Equilibrium Refinements". *Econometrica* 59:81–98.

Cho, I.-K., and D. M. Kreps. 1987. "Signalling Games and Stable Equilibria". *Quarterly Journal of Economics* 102:179–221.

Cho, I.-K., and J. Sobel. 1990. "Strategic Stability and Uniqueness in Signaling Games". *Journal of Economic Theory* 50:381–413.

Farrell, J. 1983. "Credible Repeated Game Equilibrium". Unpublished manuscript.

Farrell, J., and E. Maskin. 1989. "Renegotiation in Repeated Games". *Games and Economic Behavior* 1:327–360.

Ferreira, J. L. 1995. "On the Possibility of Stable Renegotiation". *Economics Letters* 47:269–274.

Gale, D., and M. Hellwig. 1989. "Repudiation and Renegotiation: The Case of Sovereign Debt". *International Economic Review* 30:3–31.

Govindan, S., and T. Klumpp. 2002. "Perfect Equilibrium and Lexicographic Beliefs". *International Journal of Game Theory* 31:229–243.

Govindan, S., and R. Wilson. 2001. "Direct Proofs of Generic Finiteness of Nash Equilibrium Outcomes". *Econometrica* 69:765–769.

— . 2002. "Structure Theorems for Game Trees". *Proceedings of the National Academy of Sciences* 99:9077–9080.

— . 2006. "Sufficient Conditions for Stable Equilibria". *Theoretical Economics* 1:167–206.

— . 2008. "Metastable Equilibria". *Mathematics of Operations Research* 33:787–820.

— . 2009. "On Forward Induction". *Econometrica* 77:1–28.

— . 2012. "Axiomatic Equilibrium Selection for Generic Two-Player Games". *Econometrica* 80:1639–1699.

Hillas, J., and E. Kohlberg. 2002. "Conceptual Foundations of Strategic Equilibrium". In *Handbook of Game Theory*, ed. by R. J. Aumann and S. Hart, 3:1597–1663. Amsterdam: Elsevier.

Kohlberg, E., and J.-F. Mertens. 1986. "On the Strategic Stability of Equilibria". *Econometrica* 54:1003–1037.

Kreps, D. M., and R. Wilson. 1982. "Sequential Equilibria". *Econometrica* 50:863–894.

Kuhn, H. W. 1953. "Extensive Games and the Problem of Information". In *Contributions to the Theory of Games II,* Annals of Mathematics Study 28, ed. by H. W. Kuhn and A. W. Tucker, 193–216. Princeton, NJ: Princeton University Press.

MacLeod, W. B., and J. M. Malcomson. 1989. "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment". *Econometrica* 57:447–480.

Mailath, G. J., L. Samuelson, and J. M. Swinkels. 1993. "Extensive Form Reasoning in Normal Form Games". *Econometrica* 61:273–302.

Maskin, E., and J. Tirole. 1988. "A Theory of Dynamic Oligopoly, II: Price Competition, Kinked Demand Curves, and Edgeworth Cycles". *Econometrica* 56:571–599.

Matsuyama, K. 1990. "Perfect Equilibria in a Trade Liberalization Game". *American Economic Review* 80:480–492.

— . 1997. "Credibility and Intertemporal Consistency: A Note on Strategic Macroeconomic Policy Models". *Macroeconomic Dynamics* 1:658–665.

Mertens, J.-F. 1989. "Stable Equilibria: A Reformulation Part I. Definition and Basic Properties". *Mathematics of Operations Research* 14:575–625.

— . 1991. "Stable Equilibria: A Reformulation Part II. Discussion of the Definition and Further Results". *Mathematics of Operations Research* 16:694–753.

— . 2002. "Stochastic Games". In *Handbook of Game Theory*, ed. by R. J. Aumann and S. Hart, 3:1809–1832. Amsterdam: Elsevier.

— . 2003. "Ordinality in Non Cooperative Games". *International Journal of Game Theory* 32:387–430.

Norman, T. W. L. 2018. "Inefficient Stage Nash is not Stable". *Journal of Economic Theory* 178:275–293.

Osborne, M. J. 1990. "Signalling, Forward Induction, and Stability in Finitely Repeated Games". *Journal of Economic Theory* 50:22–36.

Pearce, D. 1987. "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation". Cowles Foundation Discussion Paper No. 855.

Pearce, D., and E. Stacchetti. 1997. "Time Consistent Taxation by a Government with Redistributive Goals". *Journal of Economic Theory* 72:282–305.

Selten, R. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games". *International Journal of Game Theory* 4:25–55.

Silva, F. 2019. "Renegotiation-Proof Mechanism Design with Imperfect Type Verification". *Theoretical Economics* 14:971–1014.

Strulovici, B. 2017. "Contract Negotiation and the Coase Conjecture". *Econometrica* 85:585–616.

Van Damme, E. 1984. "A Relation between Perfect Equilibria in Extensive Form Games and Proper Equilibria in Normal Form Games". *International Journal of Game Theory* 13:1–13.

— . 1987. *Stability and Perfection of Nash Equilibria*. Berlin: Springer Verlag.

— . 1988. "The Impossibility of Stable Renegotiation". *Economics Letters* 26:321–324.

— . 1989. "Stable Equilibria and Forward Induction". *Journal of Economic Theory* 48:476–496.