RESEARCH PAPER NO. 2008

Axiomatic Theory of Equilibrium Selection for Games with Two Players, Perfect Information, and Generic Payoffs

Srihari Govindan

Robert Wilson

February 2009

This work was partially funded by a National Science Foundation grant.

ABSTRACT

Three axioms from decision theory are applied to refinements that select connected subsets of the Nash equilibria of games with perfect recall. The first axiom requires all equilibria in a selected subset to be admissible, i.e. each player's strategy is an admissible optimal reply to other players' strategies. The second axiom invokes backward induction by requiring a selected subset to contain a sequential equilibrium. The third axiom requires a refinement to be immune to embedding a game in a larger game with additional strategies and players, provided the original players' strategies and payoffs are preserved, viz., selected subsets must be the same as those induced by the selected subsets of any larger game in which it is embedded. These axioms are satisfied by refinements that select subsets that are stable as defined by Mertens (1989).

For a game with two players, perfect information, and generic payoffs, we prove the converse that the axioms require a selected set to be stable. In the space of mixed strategies of minimal dimension, the stable set is unique and consists of all admissible equilibria with the same outcome as the unique subgame-perfect equilibrium obtained by backward induction. Each other admissible equilibrium with this outcome is the profile of players' strategies in an admissible sequential equilibrium of a larger game in which the original game is embedded, so the third axiom requires it to be included.

AXIOMATIC THEORY OF EQUILIBRIUM SELECTION FOR GAMES WITH TWO PLAYERS, PERFECT INFORMATION, AND GENERIC PAYOFFS

SRIHARI GOVINDAN AND ROBERT WILSON

ABSTRACT. Three axioms from decision theory are applied to refinements that select connected subsets of the Nash equilibria of games with perfect recall. The first axiom requires all equilibria in a selected subset to be admissible, i.e. each player's strategy is an admissible optimal reply to other players' strategies. The second axiom invokes backward induction by requiring a selected subset to contain a sequential equilibrium. The third axiom requires a refinement to be immune to embedding a game in a larger game with additional strategies and players, provided the original players' strategies and payoffs are preserved, viz., selected subsets must be the same as those induced by the selected subsets of any larger game in which it is embedded. These axioms are satisfied by refinements that select subsets that are stable as defined by Mertens (1989).

For a game with two players, perfect information, and generic payoffs, we prove the converse that the axioms require a selected set to be stable. In the space of mixed strategies of minimal dimension, the stable set is unique and consists of all admissible equilibria with the same outcome as the unique subgame-perfect equilibrium obtained by backward induction. Each other admissible equilibrium with this outcome is the profile of players' strategies in an admissible sequential equilibrium of a larger game in which the original game is embedded, so the third axiom requires it to be included.

Date: 14 February 2009.

Key words and phrases. extensive-form game, perfect information, equilibrium, refinement, admissibility, backward induction, small worlds, stability.

JEL subject classification: C72.

This work was funded in part by a grant from the National Science Foundation of the United States. We thank the outsiders in our metagame.

Contents

List of Figures

1. Introduction

Kohlberg and Mertens [12] propose that Nash's [21, 22] criterion of equilibrium in a non-cooperative game should be refined by applying principles from decision theory.[1] Here we apply three axioms from decision theory adapted to games with perfect recall. In brief, these axioms require a refinement that selects connected closed subsets of equilibria to satisfy:

- Admissibility: Players' strategies are admissible optimal replies.
- Backward Induction: Selected subsets contain sequential equilibria.
- Small Worlds: Selected subsets are not affected by embedding the game within larger games that preserve players' strategies and payoffs.

These are among criteria proposed by Kohlberg and Mertens [12] and Mertens [20], although we invoke a stronger version of small worlds. Our version excludes dependence on outsiders whose presence and actions enable new pure strategies equivalent to mixed strategies in the original game. Small worlds excludes framing effects that could occur if a refinement were sensitive to the wider context in which a game is embedded.

We apply these axioms to the class of games with two players, perfect information, and generic payoffs. We prove that a refinement must select stable subsets of equilibria as defined by Mertens [18, 19]. Mertens establishes for general games the converse that a refinement that selects stable sets satisfies the axioms. Hence the axioms characterize stability as a solution concept for games with two players, perfect information, and generic payoffs.[2] Any refinement that satisfies admissibility and backward induction but is more restrictive than stability must therefore violate small worlds, e.g. by restricting the class of larger games in which a game can be embedded.

Section 2 establishes notation for Section 3, which specifies Axioms A (admissibility), B (backward induction), and S (small worlds), including a precise definition of embedding a game in a larger game. The axioms are stated for general games in extensive form with perfect recall. Section 4 summarizes implications of the axioms for games with two players, perfect information, and generic payoffs, and provides an example. Section 5 establishes notation for this class of games and states some useful properties, including a key technical proposition proved in Appendix A. Section 6 states and proves the main theorem. The proof is constructive in that each equilibrium in a stable set is shown to be induced by an admissible sequential equilibrium of a particular larger game in extensive form with perfect

[1] Also see Kohlberg [11]. Hillas and Kohlberg [10] and van Damme [25] survey subsequent developments.

[2] In [8] we prove similarly that the axioms imply that a refinement selects stable sets of signaling games with two players and generic payoffs.

recall (but with imperfect information and nongeneric payoffs) in which the given game is embedded. Section 7 provides concluding remarks and another example.

2. Notation

A typical game in extensive form is denoted Γ. Its specification includes a set N of players, a game tree that has perfect recall for each player, and a real-valued payoff $u_n(z)$ to each player n at each node z in the set Z of terminal nodes of the tree. The tree can include a specified mixed strategy of Nature. As usual, payoffs are assumed to be von Neumann-Morgenstern utilities. We assume throughout the standard epistemic conditions that the game is common knowledge and players' rationality is common knowledge.

2.1. **Strategies and Expected Payoffs.** In the normal-form representation of the game, a player's pure strategy specifies the action chosen at each of his information sets in the game tree. However, outcomes are not affected by a strategy's actions at information sets excluded by his previous actions. Thus it suffices to specify a pure strategy by the terminal nodes that are not excluded by his actions.

This specification is formalized as follows [4]. A pure strategy of a player does not exclude a terminal node z from being reached if at each of his information sets that intersect the path to z it chooses his unique action on that path. Alternatively, the player might randomize over his pure strategies, or he might use a behavioral strategy that randomizes over actions at each of his information sets.[3] A randomized strategy of either kind induces a probability distribution over the terminal nodes that are not excluded. Here we take the set $P_n \subset [0,1]^Z$ of these probability distributions as player n's set of strategies, called his *mixed* strategies.[4] If $p_n \in P_n$ then $p_n(z)$ is the probability that his actions do not exclude z, and these probabilities uniquely determine a corresponding behavioral strategy at his information sets that his prior actions do not exclude.

Player n's set P_n of mixed strategies is a closed convex polyhedron. Its vertices are obtained from profiles of pure strategies of the normal form. Let $P = \prod_n P_n$ be the set of profiles of players' mixed strategies. Note that P depends only on the game tree and summarizes its essential features.

[3]Kuhn [17] shows that these yield the same distributions of outcomes when the game has perfect recall. A randomization over pure strategies induces a unique behavior at each information set it does not exclude, and for every behavioral strategy there exist randomizations over pure strategies that, for each profile of others' strategies, yields the same probability distribution over terminal nodes.

[4]P_n is called n's set of *enabling* strategies in [4, 8]. Mertens [20, p. 554] introduces the technique of mapping randomized strategies to their induced probability distributions on terminal nodes. Koller and Megiddo [14] call them realization plans, and Koller, Megiddo, and von Stengel [15] use them for efficient computation.

If $p \in P$ then the probability that terminal node z is the outcome of the game is $\pi(z|p) = p_*(z) \prod_n p_n(z)$, where $p_*(z)$ is the probability that Nature's actions do not exclude z, because Nature and the players randomize independently. Hence player n's expected payoff is $G_n(p) = \sum_z \pi(z|p)u_n(z)$. Thus the extensive-form game Γ is summarized by the multilinear function $G : P \to \mathbb{R}^N$ that to each profile of players' mixed strategies assigns their expected payoffs. This summary specification is called the *strategic form* of the game.

2.2. **Equilibria and Refinements.** Adapting Nash's [21, 22] definition, an *equilibrium* of a game in strategic form is a profile $p \in P$ of players' mixed strategies such that each player's strategy is an optimal reply to others' strategies. That is, for each player n, $G_n(p) \geqslant G_n(p'_n, p_{-n})$ for every $p'_n \in P_n$. Note that each equilibrium by this definition corresponds to a family of equivalent equilibria, represented by either behavioral strategies or randomizations over normal-form pure strategies, that have the same distribution over outcomes.

A *refinement* is a correspondence that assigns to each game a nonempty collection of nonempty closed connected subsets of its equilibria. Each selected subset is called a *solution*. We assume that solutions are sets because Kohlberg and Mertens [12, pp. 1015, 1019, 1029] show that there need not exist a single equilibrium that satisfies weaker assumptions than the axioms invoked here. The technical requirement that a solution is connected excludes the trivial refinement that always selects the set of all equilibria. If payoffs are generic then all equilibria in a connected subset yield the same probability distribution over terminal nodes, and thus the same paths of equilibrium play in the extensive form.[5] In this case, connectedness associates solutions with selections of probability distributions over outcomes.

3. Axioms

This section presents the three axioms. The first two invoke principles of rational decisions by individual players. The third axiom requires that a refinement is not affected by extraneous features of contexts in which a game is presented.[6]

3.1. **Admissibility.** For a game with two players, a player's strategy is admissible iff it is not weakly dominated in terms of expected payoffs by another strategy. In this case admissibility is the same as in decision theory. We consider games with more than two players, however, so we assume the stronger property that a strategy is an admissible reply.

[5]Kreps and Wilson [16, Theorem 2]. We use here the stronger characterization in [3] that nongeneric payoffs lie in a lower dimensional subset.

[6]The axioms are stated for the strategic form. They have equivalent statements using, instead of players' polyhedra of mixed strategies, their simplices of randomizations over normal-form pure strategies.

Definition 3.1 (Admissible Reply)**.** A player's strategy is an *admissible reply* to a profile $p \in P$ if it is an optimal reply to each profile in some sequence in the interior of P for which p is a limit point.

An equivalent decision-theoretic specification is obtained by Blume, Brandenburger, and Dekel [1] and Govindan and Klumpp [2]. They use randomizations over normal-form pure strategies but their results apply also to the strategic form of a game. A player's strategy is an admissible reply to p iff it is a lexicographically optimal reply to a representation of other players' strategies by a lexicographic probability system $\hat{p}^0, \hat{p}^1, \hat{p}^2, \ldots$, where $\hat{p}^0 = p$ and the interior of P intersects the convex hull of the profiles $\hat{p}^k$. For a game in extensive form this condition requires that, at each information set his own strategy does not exclude, continuation of his strategy is a lexicographically optimal reply to the profile of others' strategies in the sequence $\hat{p}^k, \hat{p}^{k+1}, \ldots$ where $\hat{p}^k$ is the first profile in the system that does not exclude that information set from being reached.

Say that a profile $p \in P$ of players' strategies is admissible if each player n's strategy p_n is an admissible reply to p. When there are more than two players, this is much weaker than requiring that p results from a perfect equilibrium, which requires that the justifying sequence in Definition 3.1 is the same for all players.

Axiom A (Admissibility): Each equilibrium in a solution is admissible.

3.2. **Backward Induction.** The second axiom invokes consistent beliefs and sequential equilibria as defined by Kreps and Wilson [16, p. 872].

Definition 3.2 (Consistent Beliefs)**.** A player's *belief* assigns to each of his information sets a probability distribution over the nodes at this information set. Players' beliefs are *consistent* with an equilibrium if they are limits of conditional probabilities induced by a sequence of profiles of completely mixed strategies converging to the equilibrium.

This definition of consistent beliefs appears to depart from standard decision theory because it invokes perturbed strategies, but Kohlberg and Reny [13] show that consistency of beliefs can be derived from primitive axioms appropriate for a frequency interpretation of probabilities. We adhere to Kreps and Wilson [16] definition of sequential equilibrium in terms of behavioral strategies.

Definition 3.3 (Sequential Equilibrium)**.** An equilibrium in behavioral strategies is *sequential* if there exists a profile of consistent beliefs such that, conditional on a player's belief at an information set, continuation of his behavioral strategy is an optimal reply to other players' strategies.

Govindan and Klumpp [2, Section 5] observe that a sequential equilibrium can be represented by a lexicographic probability system. The optimality property in Definition 3.3 is called sequential rationality. If continuation is required to be optimal only at his information sets that the player's own strategy does not exclude then it is called weak sequential rationality by Reny [23].

The second axiom requires that some equilibrium in a solution is sequential.

Axiom B (Backward Induction): Each solution contains an equilibrium implied by a sequential equilibrium.

That is, a solution must contain an equilibrium p such that, for some sequential equilibrium, each $p_n(z)$ is the product of player n's behavioral probabilities of choosing his actions on the path to z.

For the games with perfect information and generic payoffs studied later, Axiom B requires that a solution contains a subgame-perfect equilibrium constructed by backward induction, which is a special case of a sequential equilibrium. For more general games we interpret sequential equilibrium as the relevant generalization of backward induction.

3.3. **Small Worlds.** Equilibria of a game depend only on its strategic form. The analogous property of a refinement is called *invariance*. As in decision theory, invariance requires that it is irrelevant whether a randomization over pure strategies is treated as an additional pure strategy. Similarly, equilibria are not affected by adding dummy players, i.e. 'outsiders' whose actions do not affect strategies and payoffs of 'insiders' who are the players in the given game. The analogous property of a refinement is called 'small worlds' by Mertens [20]. This property too is familiar in decision theory where one excludes dependence on payoff-irrelevant events (Savage [24]). When invariance and small worlds are adopted as axioms, they require that a refinement is not affected by two particular presentation effects, i.e. embeddings of the given game in larger games with redundant pure strategies or dummy players.

The axiom adopted here excludes a refinement from depending on more general presentation effects. We use the same name, *small worlds*, but consider more general embeddings. For notational simplicity, we use the strategic form of a game to state the axiom. Thus, as in Section 2, a game Γ is summarized by a multilinear function $G : P \to \mathbb{R}^N$ that to each profile of players' mixed strategies assigns expected payoffs to the players in N.

As mentioned, the purpose of the axiom is to prevent refinements from depending on wider contexts in which a game is played, provided a context does not alter players' feasible strategies and payoffs. By a context we mean here a 'larger' game $\tilde{G} : \tilde{P} \times P_o \to \mathbb{R}^{N \cup o}$ in

which game G is embedded, subject to certain restrictions specified below. The larger game $\tilde{G}$ has outsiders in a set o, in addition to insiders who are the players in N, and there can be additional moves by Nature. Also, an insider n can have additional pure strategies in $\tilde{G}$ that are not pure strategies in G.

The basic requirement is that an embedding should not alter the game among insiders, conditional on any specific strategies of outsiders. Restrictions on an embedding should therefore ensure that outsiders' strategies are not payoff-relevant for insiders, and that insiders' additional pure strategies are redundant—although translation from a pure strategy in $\tilde{G}$ to a mixed strategy in G might depend on outsiders' strategies.[7]

These restrictions have a technical formulation. There should exist a multilinear map $f : \tilde{P} \times P_o \to P$ that is surjective and such that $\tilde{G}_n = G_n \circ f$ for each insider n. Moreover, to exclude an embedding from enabling insiders to coordinate their strategies, f should factor into separate multilinear maps $(f_n)_{n\in N}$, where each component is a map $f_n : \tilde{P}_n \times P_o \to P_n$ such that $f_n(\cdot, p_o)$ maps $\tilde{P}_n$ surjectively onto P_n for each mixed strategy $p_o \in P_o$ of outsiders.

Admittedly, a statement of the axiom that uses this technical language could contain unsuspected implications. However, after stating the formal definition, we provide in Proposition 3.5 an equivalent formulation that is more detailed and more transparent, and that verifies the requisite properties. Also, Proposition 3.6 applies a precise test of whether the axiom is correctly stated—a refinement that satisfies the axiom should be immune to the same embeddings that equilibria are.

Definition 3.4 (Embedding)**.** A game $\tilde{G} : \tilde{P} \times P_o \to \mathbb{R}^{N\cup o}$ and a collection of multilinear maps $f_n : \tilde{P}_n \times P_o \to P_n$, one for each player $n \in N$, *embed* a game $G : P \to \mathbb{R}^N$ if

(a) for each $p_o \in P_o$, $f_n(\cdot, p_o)$ maps $\tilde{P}_n$ surjectively onto P_n, and
(b) $\tilde{G}_n = G_n \circ f$, where $f = (f_n)_{n\in N}$.

Condition (a) ensures that embedding has no net effect on an insider's set of mixed strategies, conditional on outsiders' strategies, and condition (b) ensures that there is no net effect on any insider's payoffs. Proposition 3.5 below elaborates this interpretation in terms of pure strategies.

Hereafter, if $\tilde{G}$ embeds G via maps $f = (f_n)$ then we say that $(\tilde{G}, f)$ embeds G and that $\tilde{G}$ is a *metagame* for G. We omit description of f for metagames in extensive form that embed

[7]For example, an insider might condition his choices on which actions he observes an outsider takes, but if other insiders do not observe this outsider's actions then this is equivalent to the insider using the outsider's actions as a randomization device. More generally, outsiders' strategies can affect how an insider's redundant pure strategies in $\tilde{G}$ are mapped into mixed strategies in G. Proposition 3.5 below states the general form of this map.

a game in extensive or strategic form. An elaborate example of a metagame in extensive form that embeds a game in extensive form is constructed in proving Theorem 6.1.

Invariance uses the special case in which o and P_o are singletons and each f maps pure strategies of $\tilde{G}$ to equivalent pure or mixed strategies of G. Mertens' small worlds criterion uses the special case in which $\tilde{P} = P$ and each f_n is the projection map to P_n. In these two cases, $\tilde{P} \supseteq P$, but embedding allows more general versions that are identified precisely in Proposition 3.5 below.

A multilinear map $f_n : \tilde{P}_n \times P_o \to P_n$ is completely specified by its values at vertices of $\tilde{P}_n \times P_o$, which recall are images of pure strategies of the normal form. Let $\tilde{P}_n^\circ$ and P_o° be the sets of vertices of $\tilde{P}_n$ and P_o, and let f_n° be the restriction of f_n to $\tilde{P}_n^\circ \times P_o^\circ$.

Proposition 3.5. *$\tilde{G}$ embeds G via a collection of multilinear maps $f = (f_n)_{n\in N}$ if and only if for each player n there exists $\tilde{T}_n \subseteq \tilde{P}_n^\circ$ and a bijection $h_n : \tilde{T}_n \to P_n^\circ$ such that for each $(\tilde{p}^\circ, p_o^\circ) \in \tilde{P}^\circ \times P_o^\circ$ and $\tilde{t}_n \in \tilde{T}_n$:*

(1) $f_n^\circ(\tilde{t}_n, p_o^\circ) = h_n(\tilde{t}_n)$,

(2) $\tilde{G}_n(\tilde{p}^\circ, p_o^\circ) = G_n(f^\circ(\tilde{p}^\circ, p_o^\circ))$, *where* $f^\circ = (f_n^\circ)_{n\in N}$.

Property (1) assures that each vertex $p_n^\circ \in P_n^\circ$ is equivalent to some vertex $\tilde{t}_n = h_n^{-1}(p_n^\circ) \in \tilde{T}_n$, independently of the outsiders' profile p_o°. Property (2) assures that players' payoffs from vertices of G are preserved by the metagame $\tilde{G}$.

Vertices in $\tilde{P}_n^\circ \setminus \tilde{T}_n$ are redundant because payoffs from profiles in $\prod_n \tilde{T}_n$ exactly replicate payoffs from corresponding profiles in $\prod_n P_n^\circ$ for the embedded game G. In particular, if $f_n^\circ(\tilde{p}_n^\circ, p_o^\circ) = p_n \notin P_n^\circ$ then, conditional on p_o°, the vertex $\tilde{p}_n^\circ$ is equivalent for insiders to the mixed strategy p_n in P_n. Thus, conditional on each profile p_o° of outsiders' vertices, embedding preserves the strategic form of the game among insiders.

Proof of Proposition. Suppose we have a game $\tilde{G} : \tilde{P} \times P_o \to \mathbb{R}^{N\cup o}$ and a collection of multilinear maps $f_n : \tilde{P}_n \times P_o \to P_n$, one for each $n \in N$, such that conditions (1) and (2) of the proposition are satisfied. Then, by condition (1) and multilinearity of f_n for each n, for each fixed p_o, $f_n(\cdot, p_o)$ is surjective because it maps the convex hull of $\tilde{T}_n$ onto P_n. Also, condition (2) and multilinearity of each f_n imply that $\tilde{G} = G \circ f$. According to Definition 3.4, therefore, $(\tilde{G}, f)$ embeds G.

Now suppose that $(\tilde{G}, f)$ embeds G. Let p_o be a profile of completely mixed strategies for outsiders. Because f_n is multilinear it induces a linear mapping $f_n(\cdot, p_o)$ from $\tilde{P}_n$ to P_n that is surjective by the definition of embedding. Hence, for each $p_n^\circ \in P_n^\circ$ there exists a vertex $\tilde{t}_n(p_n^\circ)$ in $\tilde{P}_n^\circ$ that is mapped to p_n° by this linear map. We claim that $f_n(\tilde{t}_n(p_n^\circ), p_o^\circ) = p_n^\circ$

for all $p_o^\circ \in P_o^\circ$. Indeed, since p_o is in the interior of P_o, we can express it as a convex combination $\sum p_o(p_o^\circ)p_o^\circ$, where for each vertex p_o°, $p_o(p_o^\circ) > 0$ is the weight on the vertex p_0°. Then, $f_n(\tilde{t}_n(p_n^\circ), p_o) = \sum_{p_o^\circ} f_n(\tilde{t}_n(p_n^\circ), p_o^\circ)p_o(p_o^\circ)$. Therefore, if $f_n(\tilde{t}_n(p_n^\circ), p_o^\circ) \neq p_n^\circ$ for some p_o° then $f_n(\tilde{t}_n(p_n^\circ), p_o)$, which is an average of values at vertices of P_o°, cannot be p_n°. Thus, $f_n(\tilde{t}_n(p_n^\circ), p_o^\circ) = p_n^\circ$ for all p_o°. Let $\tilde{T}_n \subset \tilde{P}_n^\circ$ be a collection comprising a different vertex $\tilde{t}_n(p_n^\circ)$ for each $p_n^\circ \in P_n^\circ$ and let h_n be the associated bijection. Define $f_n^\circ : \tilde{P}_n^\circ \times P_o^\circ \to P_n$ by $f_n^\circ(\tilde{p}_n^\circ, p_o^\circ) = f_n(\tilde{p}_n^\circ, p_o^\circ)$. Then conditions (1) and (2) of the proposition are satisfied. □

Now we apply the aforementioned test and verify that equilibria are not affected by embedding in a metagame.

Proposition 3.6. *If $(\tilde{G}, f)$ embeds G then the equilibria of G are the f-images of the equilibria of $\tilde{G}$.*

Proof. Suppose $(\tilde{p}, p_o)$ is an equilibrium of $\tilde{G}$ and let $p = f(\tilde{p}, p_o)$. For any insider n and his strategy $p_n' \in P_n$ there exists $\tilde{p}_n' \in \tilde{P}_n$ such that $f_n(\tilde{p}_n', p_o) = p_n'$ because $f_n(\cdot, p_o)$ is surjective by condition (a) of Definition 3.4 an embedding. Using condition (b),

$$G_n(p_n', p_{-n}) = G_n(f(\tilde{p}_n', \tilde{p}_{-n}, p_o)) = \tilde{G}_n(\tilde{p}_n', \tilde{p}_{-n}, p_o) \leqslant \tilde{G}_n(\tilde{p}, p_o) = G_n(f(\tilde{p}, p_o)) = G_n(p)\,,$$

where the inequality obtains because $(\tilde{p}, p_o)$ is an equilibrium of $\tilde{G}$. Hence p is an equilibrium of G.

Conversely, suppose p is an equilibrium of G. For each n, express p_n as a convex combination $\sum \alpha(p_n^\circ)p_n^\circ$ of the vertices p_n° of P_n. For each n, let h_n be the bijection given by Proposition 3.5. Let $\tilde{p}_n$ be the strategy for insider n in $\tilde{G}$ given by $\sum \alpha(p_n^\circ)h^{-1}(p_n^\circ)$. Since f_n is multilinear, by condition (1) of Proposition 3.5, $f_n(\tilde{p}_n, \cdot) = p_n$ and thus $f(\tilde{p}, \cdot) = p$. Hence, it suffices to show that there exists a strategy profile p_o for outsiders such that $(\tilde{p}, p_o)$ is an equilibrium of $\tilde{G}$. By fixing the profile of insiders' strategies to be $\tilde{p}$ one induces a game among outsiders. Let p_o be an equilibrium of this induced game among outsiders. To see that $(\tilde{p}, p_o)$ is an equilibrium of $\tilde{G}$, observe that for each vertex $\tilde{p}_n^\circ$ of an insider n:

$$\tilde{G}_n(\tilde{p}_n^\circ, \tilde{p}_{-n}, p_o) = G_n(f_n(\tilde{p}_n^\circ, p_o), p_{-n}) \leqslant G_n(p) = G_n(f(\tilde{p}, p_o)) = \tilde{G}_n(\tilde{p}, p_o)\,,$$

where the first and second equalities use the property $f(\tilde{p}, \cdot) = p$ established above, and the inequality obtains because p is an equilibrium of G. □

A corollary of Proposition 3.6 is that embedding does not introduce correlation among insiders' strategies.

Using Definition 3.4 of embedding, the small worlds axiom is the following.

Axiom S (Small Worlds): If $(\tilde{G}, f)$ embeds G then the f-images of the solutions that a refinement selects for $\tilde{G}$ are the solutions selected for G.

In view of Proposition 3.6, this axiom is an instance of the general principle that a refinement should inherit invariance properties of equilibria.

3.4. **Summary of the Axioms.** We study refinements that are independent of embeddings in metagames that, for each profile of outsiders' strategies, preserve the strategic form of the game among insiders. And, we require that their solutions are closed connected subsets of admissible equilibria that contain sequential equilibria. In particular, a solution of a metagame must contain an admissible sequential equilibrium whose image is in the corresponding solution of the embedded game.

Mertens [18, 19] proves for general games that stable sets of equilibria satisfy Axiom A, Axiom B, invariance, and his version of small worlds. A modification of his proof extends this conclusion to Axiom S.

4. Games with Perfect Information

The remainder of this paper applies Axioms A, B, and S to the class of games with two players, perfect information, and generic payoffs. A game in this class is called a *PI game* for simplicity.

In this section we summarize immediate implications of the axioms for PI games, and describe the main theorem that is stated and proved in Section 6. We also provide a simple example that illustrates the content of the theorem and the key property that is the focus of the proof.

4.1. **Implications of the Axioms for PI Games.** A PI game has special features. It has a unique sequential equilibrium. This is the subgame-perfect equilibrium obtained by backward induction, and it uses only pure strategies. Moreover, this equilibrium is included in the unique essential component [6] of the game's equilibria. Within this component is the unique essential component of admissible equilibria.

The theorem in Section 6 shows that a refinement satisfying the axioms selects a unique solution for each PI game. This solution is the entire component of admissible equilibria that contains the subgame-perfect equilibrium. In particular, Axiom A requires that a solution contains only admissible equilibria—which for two-player games are the weakly undominated strategies—and Axiom B requires that the unique subgame-perfect equilibrium is included in each solution. A solution must therefore be a connected closed subset of the component of admissible equilibria that contains the subgame-perfect equilibrium.

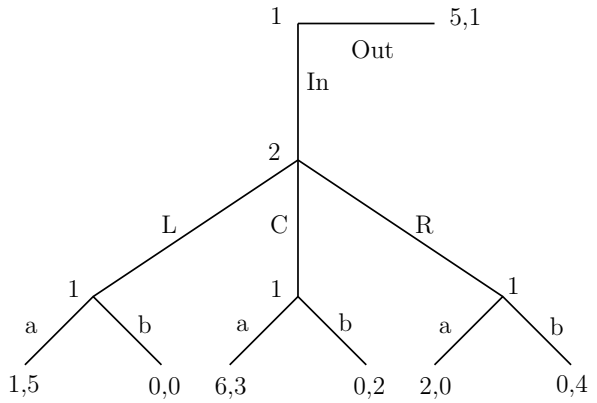

FIGURE 1. A game Γ with perfect information

Extreme Point	$\Pr(L \mid In)$	$\Pr(C \mid In)$	$\Pr(R \mid In)$
L:	1	0	0
LC:	.2	.8	0
CR:	0	.75	.25
R:	0	0	1

TABLE 1. Extreme points of 2's behavioral strategies in the component of admissible equilibria

The remarkable aspect of Theorem 6.1 is that *every* equilibrium in the component of admissible equilibria must be included in a solution. This is necessary to account for all the metagames in which the PI game can be embedded. In other words, the theorem shows that stability against every perturbation of players' strategies is equivalent to immunity to embeddings in metagames, as required by Axiom S.

4.2. **An Example.** We use an example to illustrate what is required for a proof. Figure 1 shows an example of a PI game Γ. There are two components of its equilibria. One component is inessential and all its equilibria are inadmissible. On the equilibrium path, 1 chooses *In*, then 2 chooses *C*, and then 1 chooses *a*. This equilibrium path is sustained by 1's inadmissible strategies that choose *b* with sufficiently high probability after 2's choice of *L*. Either Axiom A or B excludes a solution from residing in this component.

The other component is essential and its equilibrium path is sustained by admissible equilibria. It contains the subgame-perfect equilibrium in which 1 chooses *Out*, anticipating that after *In* she would choose *a* after each choice by 2, which optimally for 2 is to choose *L*. This component's four extreme points are identified by 2's strategies labeled L, R, CR, LC in Table 1.

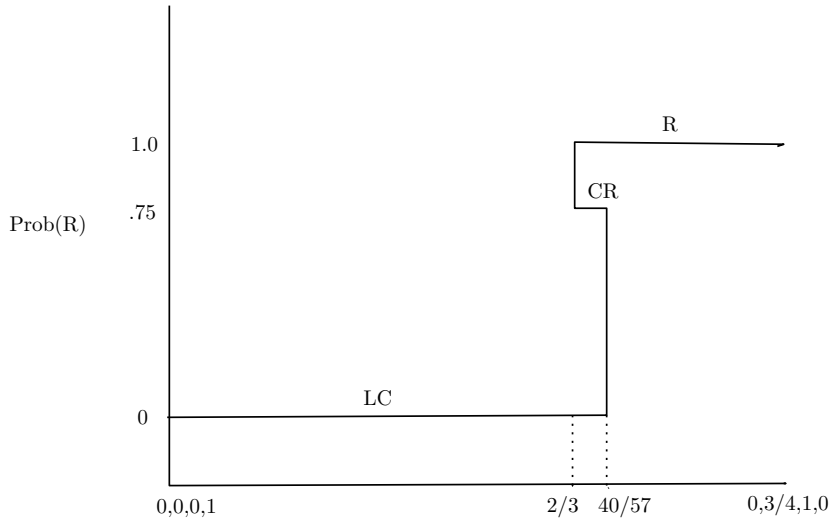

FIGURE 2. Graph of 2's strategies in admissible equilibria over the interval of perturbations of 1's strategies between (0,0,0,1) and (0,3/4,1,0).

To illustrate requirements for stability, we show examples of equilibria of nearby games obtained by perturbing player 1's strategies. Figure 2 shows the graph of admissible equilibria over an interval of perturbations, constructed as follows. Represent a behavioral strategy for player 1 by a vector $b_1 = [b_1(Out), b_1(a|In, L), b_1(a|In, C), b_1(a|In, R)]$ in the 4-dimensional unit cube. Each perturbed game is obtained by assuming that for each strategy b_1 player 1 might choose, what actually happens is that with arbitrarily small probability $\varepsilon > 0$ her choice is superseded by implementation of another strategy b_1^x, where x ranges over the interval $0 \leqslant x \leqslant 1$ in Figure 2. To construct the figure we assume that $b_1^0 = (0, 0, 0, 1)$ and $b_1^1 = (0, 3/4, 1, 0)$. The figure implies that, besides 2's choice of L in the subgame-perfect equilibrium, a stable set must also include each extreme point LC, CR, and R, since each is the limit of admissible equilibria of perturbed games.

Theorem 6.1 below shows that the axioms imply that indeed all four of 2's extreme points and their mixtures must be included in a solution. The method of proof is to show that if some point in the convex hull of these four extreme points is not included in a proposed solution, then there exists a metagame $(\tilde{\Gamma}, f)$ in extensive form that embeds Γ and for which the f-images of admissible sequential equilibria lie outside this proposed solution—thus Axioms B and S require that the solution includes the entire convex hull.

A key step of the proof modifies the game Γ by adding the redundant strategy for player 1 that is shown in Figure 3. In this expanded game, after player 1 initially rejects Out but before committing to In, she can choose Reconsider, which implements the strategy that

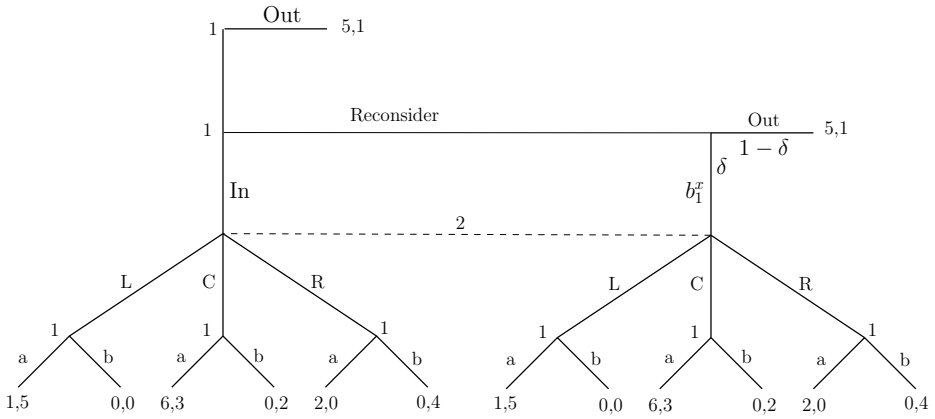

FIGURE 3. Game Γ augmented with 1's redundant strategy Reconsider that with probabilities $1-\delta$ and δ implements either Out, or In followed by the behavioral strategy b_1^x

with probability $1-\delta$ chooses Out and with probability δ chooses In and then implements the behavioral strategy b_1^x at her information sets that follow player 2's responses to In. The information set for player 2 indicates that he cannot know whether 1 chose In or Reconsider. When δ is sufficiently small, in any equilibrium of the subgame that follows 1's initial rejection of Out, player 1 must choose Reconsider with positive probability because it is nearly as advantageous as Out. The expanded game therefore simulates the effect of perturbing the strategies of player 1 (other than Out, which is her equilibrium strategy in Γ). The proof in Section 6 also introduces outsiders whose strategies determine which behavioral strategy b_1^x is implemented if player 1 chooses Reconsider and the outcome again rejects Out. This behavioral strategy determines which admissible equilibrium results in the expanded game. Player 2 is also provided options to reconsider his choices.

5. Notation and Properties of PI Games

In this section we establish notation and properties specific to PI games.

We now use Γ to denote a typical PI game. The set of players is $N=\{1,2\}$. Represent the game tree as $(X,\prec)$, where X is the set of nodes and $\prec$ is the relation of precedence. As before, $Z\subset X$ is the set of terminal nodes and payoffs are given by a point u in $U=\mathbb{R}^{N\times Z}$, where $u_n(z)$ is the payoff to player $n\in N$ at terminal node $z\in Z$. We assume throughout that payoffs are generic, i.e. $u\in U\setminus U_\circ$ where the excluded set $U_\circ$ is a lower-dimensional set of payoffs derived in [3].

5.1. **Derivation of the Strategic Form.** For completeness, we first derive the strategic form from the normal form of the game. Let X_0 be the set of nodes where Nature moves.

Assume that all of Nature's strategies have positive probability. For each player n, let X_n be the set of nodes where player n moves. For each node $x \in X_n$, let $A_n(x)$ be the set of actions available to player n at x. Assuming actions at all nodes are labeled differently, let A_n be the set of all actions of player n. Then the set S_n of n's normal-form pure strategies is the set of functions $s_n : X_n \to A_n$ such that $s_n(x) \in A_n(x)$ for each $x \in X_n$. Let Σ_n be the simplex of randomizations over S_n.

For each player n, his pure strategy $s_n \in S_n$, and any node $y \in X$, let $\beta_n(y, s_n)$ be the probability that s_n does not exclude y, i.e. $\beta_n(y, s_n) = 1$ if $s_n(x) = a$ for every $(x, a) \prec y$ such that $x \in X_n$ and $a \in A_n(x)$, and otherwise $\beta_n(y, s_n) = 0$. Extend $\beta_n(y, \cdot)$ to a function over n's simplex Σ_n of randomized strategies via $\beta_n(y, \sigma_n) = \sum_{s_n \in S_n} \beta_n(y, s_n)\sigma_n(s_n)$. Similarly, let $\beta_*(y)$ be the probability that Nature does not exclude y. Then the probability that a profile $s \in S$ of pure strategies does not exclude y is $\beta(y, s) = \beta_*(y)\beta_1(y, s_1)\beta_2(y, s_2)$. Because Nature and players move independently, the function β extends similarly to profiles of randomized strategies via $\beta(y, \sigma) = \beta_*(y)\beta_1(y, \sigma_1)\beta_2(y, \sigma_2)$. Player n's expected payoff from a profile $s \in S$ of players' pure strategies is $\sum_{z \in Z} \beta(z, s)u_n(z)$, and from a profile $\sigma \in \Sigma$ of randomized strategies it is $\sum_{z \in Z} \beta(z, \sigma)u_n(z)$.

Define maps $\rho = (\rho_n)_{n \in N}$ where for each player n, $\rho_n : \Sigma_n \to [0, 1]^Z$ and $\rho_n(\sigma_n) = (\beta_n(z, \sigma_n))_{z \in Z}$. Let $P_n = \rho_n(\Sigma_n)$ be the image of ρ_n and let $P = \prod_n P_n$. Then P_n is the set of n's mixed strategies of the strategic form as defined in Section 2. Given a profile $\sigma \in \Sigma$, if $p_n = \rho_n(\sigma_n)$ for each player n then player n's payoff from σ is $G_n(p) = \sum_{z \in Z} \beta_*(z)p_1(z)p_2(z)u_n(z)$ since $p_n(z)$ is the probability that n's strategy does not exclude z. As in Section 2, the multilinear map $G : P \to \mathbb{R}^N$ is the strategic form of Γ.

5.2. **Stable Sets.** Recall that for a two-player game a strategy is admissible iff it is not weakly dominated. Also, because Γ has perfect information and payoffs are generic, there is a unique subgame-perfect equilibrium s^*, and all equilibria in the same component as s^* induce the same distribution of outcomes. Therefore, let Σ^* be the unique component of admissible equilibria that contains s^*. Every stable set of Γ is contained in Σ^*. Moreover, Σ^* is itself stable [5]. Let P^* be the image of Σ^* under ρ, and let P_n^* be its projection into P_n.

5.3. **Subgames after Deviations.** For each node $x \in X$, $\beta(x, \sigma)$ is the same, say $\beta^*(x)$, for all $\sigma \in \Sigma^*$. Let X^* be the subset of nodes such that $\beta^*(x) > 0$, and for each n let $X_n^* = X_n \cap X^*$. Similarly, let $Z^* \subset Z$ be the set of terminal nodes for which $\beta^*(z) > 0$. By genericity, at each node $x \in X_n^*$ player n chooses the same action $a^*(x) \in A_n(x)$ in all equilibria in Σ^*. Therefore, for each $z \in Z^*$ both players choose all their actions on the

path to z with probability one, i.e. for each player n, $p_n^*(z) = 1$ for all $p_n^* \in P_n^*$ and thus $\beta^*(z) = \beta_*(z)$.

Given $z \notin Z^*$, let $x \in X \setminus X_0$ be the last node preceding z such that $\beta^*(x) > 0$. Then $x \in X_m^*$ for some player m and z follows x by m's choice of some action $a \in A_m(x)$, $a \neq a^*(x)$. In the subgame following a, if player n has no move then by genericity a is an inferior action for m against all equilibria in P^* and thus $p_m^*(z) = 0$ and $p_n^*(z) = 1$ for all z following a—but if player n does have a move following a then $p_n^*(z)$ might differ among equilibria in P^*.

To summarize the preceding paragraph, for each player n and each $p_n^* \in P_n^*$, $p_n^*(z) = 1$ if $z \in Z^*$; $p_n^*(z) = 0$ if the last node preceding $z \notin Z^*$ that belongs to $X_m^* \cup X_n^*$ belongs to X_m^*; and $p_n^*(z) = 1$ if the last node x preceding z that belongs to $X_m^* \cup X_n^*$ belongs to X_m^* and n has no move following m's choice a at x that leads to z. Thus the only indeterminacy is when in the latter case player n has a move after player m chooses the non-equilibrium action a at x. This motivates the following constructions.

Because Γ has perfect information, each node $y \in X \setminus Z$ initiates a subgame that we denote Γ^y. For each player n, let X_n° be the set of nodes $y \in X \setminus X^*$ such that the immediate predecessor x of y belongs to X_n^*, and in the subgame Γ^y that starts at y, player n has some node where he moves. (Note that y need not belong to X_n: it merely has the property that its predecessor belongs to X_m^* and the action there leading to y is a non-equilibrium action.) Let $X^\circ = X_1^\circ \cup X_2^\circ$. For each $y \in X^\circ$, let S_n^y, Σ_n^y, B_n^y, and P_n^y be the sets of n's pure strategies, randomizations over pure strategies, behavioral strategies, and mixed strategies in the subgame Γ^y, with ρ_n^y being the map from randomizations over pure strategies to mixed strategies, and let $P^y = P_1^y \times P_2^y$.

Let $W_m^{y,*}$ be the continuation payoff to player m at her node $x \in X_m^*$ preceding y when she chooses $a^*(x)$ and subsequent play adheres to an equilibrium in P^*.

By construction, for each $y \in X_n^\circ$ player n does not exclude y in an equilibrium $p^* \in P^*$, and in particular $\beta_n(y, p_n^*) = 1$. Hence, for each $p_n^* \in P_n^*$ the projection of p_n^* to the set of z that follow y is a mixed strategy $p_n^{y,*} \in P_n^y$ of the subgame Γ^y. Let $P_n^{y,*} \subset P_n^y$ be the collection of n's mixed strategies for Γ^y that are projections of n's mixed strategies in P_n^*.

Proposition 5.1. *$P^* = P_1^* \times P_2^*$. Moreover, for each n, the projection from P_n^* to $\prod_{y \in X_n^\circ} P_n^{y,*}$ is a homeomorphism.*

Proof. Given $(p_1^*, p_2^*) \in P_1^* \times P_2^*$, observe that for each n, p_n^* is an admissible strategy with the property that $p_n^*(z) = \beta_n(z, s_n^*)$ for all terminal nodes z that do not succeed a node $y \in X_n^\circ$; and in the subgame Γ^y at each such y, the projection of p_1^* to P_n^y is such that player m's continuation payoff at her node x immediately preceding y is no more than $W_m^{y,*}$ by

leading play into the subgame Γ^y. Thus, (p_1^*, p_2^*) is an admissible equilibrium inducing the same outcome as s^*. Also, by definition P_n^* is a connected set of strategies that includes s^*. Therefore, (p_1^*, p_2^*) belongs to the connected set of admissible equilibria that contains s^*. This last set is, by definition, P^*. Hence, $P_1^* \times P_2^* \subseteq P^*$; the reverse inclusion being obvious, the first statement is proved.

As for the second statement, since strategies in P_n^* vary only across terminal nodes that follow some $y \in X_n^\circ$, the projection from P_n^* to $\prod_y P_n^{y,*}$ is injective. To prove that it is surjective, take a strategy p_n^y in $P_n^{y,*}$ for each $y \in X_n^\circ$. Construct a strategy $p \in P_n$ by letting $p_n(z)$ be $p_n^y(z)$ if z succeeds $y \in X_n^\circ$; otherwise let it equal $\beta_n(z, s_n^*)$. Clearly p_n belongs to P_n^*. □

5.4. **The Pseudo-Manifold Property.** We conclude the setup by stating a key technical property that enables the Hopf extension theorem to be invoked in the proof of Theorem 6.1.

Let $A^y \subset P^y \times P^y$ be the closure of the set of pairs (p^y, q^y) of profiles of mixed strategies for the subgame Γ^y such that p^y is in the interior of P^y, $q_n^y \in P_n^{y,*}$, and there exist $\lambda_m, \lambda_n \in [0, 1)$ and a profile $r_n^y \in P^y$ such that:[8]

(i) $q_m^y = (1 - \lambda_m)p_m^y + \lambda_m r_m^y$.

(ii) if $\lambda_m > 0$ then r_m^y yields payoff $W_m^{y,*}$ against q_n^y in Γ^y and there exists a sequence of ε's converging to zero such that r_m^y is a weakly sequentially rational strategy for m in Γ^y against beliefs induced by the corresponding sequence of n's mixed strategy

$$(1 - \varepsilon)q_n^y + \varepsilon((1 - \lambda_n)p_n^y + \lambda_n r_n^y)\,.$$

(iii) $(1 - \lambda_n)q_n^y + \lambda_n r_n^y$ is an admissible best reply for n against m's strategy q_m^y .

Let $\pi^y : A^y \to P^y$ be the projection map to the first factor. Also, let ∂A^y be the inverse image of ∂P^y under π^y.

Proposition 5.2. *$(A^y, \partial A^y)$ is a pseudo-manifold with boundary and has the same dimension as P^y. Moreover, the projection map $\pi^y : (A^y, \partial A^y) \to (P^y, \partial P^y)$ has degree one.*

The proof is in Appendix A.

[8]An interpretation is that the profile p^y represents players' initial beliefs about each other's strategy after the deviation at y, and q^y is the updated profile obtained by anticipating that with some probabilities λ they will voluntarily choose strategies r^y that are optimal replies to each other's initial and updated beliefs. The conditions for m and n are asymmetric because m's voluntary part r_m^y replies optimally mainly to n's equilibrium strategy $q_n^y \in P_n^{y,*}$ because only with arbitrarily small probability ε will m's initial deviation at x be followed by a second deviation by n at y or later in the subgame Γ^y.

6. Statement and Proof of the Theorem

Theorem 6.1. *Axioms A, B, S imply that the unique solution of the strategic form of a PI game is the component of admissible equilibria that contains the subgame-perfect equilibrium.*

Proof. Suppose $\bar{P}$ is a solution in terms of mixed strategies that is selected by a refinement satisfying the axioms. Then by Axioms A and B, $\bar{P}$ is contained in P^*. By Proposition 5.1, it is sufficient to prove for each pair of collections, one for each player n, of $(q_n^{y,*})_{y\in X_n^\circ}$, with $q_n^{y,*} \in P_n^{y,*}$ for each y, that there exists $q \in \bar{P}$ whose projection for each n and $y \in X_n^\circ$ is the given $q_n^{y,*}$. For each player n and $y \in X_n^\circ$, let V^y be an arbitrary neighborhood of $q_n^{y,*}$ in P_n^y. We construct a metagame in which every sequential equilibrium has player n using a mixed strategy in V^y for each $y \in X_n^\circ$. Since $\bar{P}$ is a closed set, and V^y is an arbitrary neighborhood of $q_n^{y,*}$ for each n and $y \in X_n^\circ$, this proves the theorem.

6.1. **Preliminary Constructions.** For each player n and each node $y \in X_n^\circ$, pick a mixed strategy $q_n^{y,*} \in P_n^{y,*}$. Let m be the other player. By admissibility, there exists $p_m^{y,*}$ in the interior of P_m^y against which $q_n^{y,*}$ is a best reply, and there exists $p_n^{y,*}$ in the interior of P_n^y such that m's choice of $a^*(x)$ is the only optimal reply in the continuation from x. Therefore, $(p^{y,*}, q^{y,*}) \in A^y \setminus \partial A^y$, where $p^{y,*} = (p_m^{y,*}, p_n^{y,*})$ and $q^{y,*} = (p_m^{y,*}, q_n^{y,*})$. Let U^y be a neighborhood of $(p^{y,*}, q^{y,*})$ that is a simplex of the same dimension as P^y, is contained in $A^y \setminus \partial A^y$, and has a projection onto the last factor that is contained in V^y.

Since π^y has degree one, so does its restriction $\pi^y_{\partial A^y} : \partial A^y \to \partial P^y$. Define $\tilde{\pi}_{\partial A^y} : \partial A^y \to \partial P^y$ as follows: for each $(p^y, q^y) \in \partial A^y$ $\tilde{\pi}_{\partial A^y}(p^y, q^y)$ is the unique point on the boundary that belongs to the line from p^y through $p^{y,*}$, i.e. it is the unique point in ∂P^y of the form $\lambda p^y + (1-\lambda)p^{y,*}$ with $\lambda \neq 1$. $\tilde{\pi}_{\partial A^y}$ is the composition of π^y with an "antipodal" map from ∂P^y to itself; thus, it is a degree-one mapping as well. Moreover it has no point of coincidence with $\pi^y_{\partial A^y}$. $(A^y \setminus (U^y \setminus \partial U^y), \partial A^y \cup \partial U^y)$ is a pseudo-manifold with boundary. Therefore, we can now construct a map $\tilde{f}^y_{\partial U^y}$ from ∂U^y to ∂P^y with degree one such that, using the Hopf extension theorem, we can extend the two maps $\tilde{\pi}_{\partial A^y}$ and $\tilde{f}_{\partial U^y}$ to a map $\tilde{f}^y$ from $A^y \setminus (U^y \setminus \partial U^y)$ to ∂P^y. Finally, we can extend $\tilde{f}^y$ to a map from A^y to P^y as follows: map $(p^{y,*}, q^{y,*})$ to $p^{y,*}$ and map all other points in U^y by linear interpolation, i.e. $\tilde{f}^y(\lambda(p^{y,*}, q^{y,*}) + (1-\lambda)(p^y, q^y)) = \lambda p^{y,*} + (1-\lambda)\tilde{f}^y(p^y, q^y)$ for all $\lambda \in [0,1]$ and $(p^y, q^y) \in \partial U^y$. The only point of coincidence between $\tilde{f}^y$ and π^y is $(p^{y,*}, q^{y,*})$. Extend $\tilde{f}^y$ to map from $P^y \times P^y$ to P^y, denoting it still by $\tilde{f}^y$. Replacing $\tilde{f}^y$ with a small perturbation of it, we can assume

that the image of $\tilde{f}^y$ is contained in the interior of P^y and that $(p^{y,*}, q^{y,*})$ is still the only point of coincidence between π^y and the restriction of $\tilde{f}^y$ to A^y.

Choose $\alpha > 0$ such that for $(p^y, q^y) \in P^y \times P^y$, the distance between $\tilde{f}^y(p^y, q^y)$ and ∂P^y is strictly greater than α and, furthermore, $\|\tilde{f}^y(p^y, q^y) - p^y\| > \alpha$ if $(p^y, q^y) \in A^y \setminus (U^y \setminus \partial U^y)$. Take simplicial subdivisions $\mathcal{K}^y_m$ of P^y_m and $\mathcal{K}^y_n$ of P^y_n such that ∂P^y_m and ∂P^y_n are full subcomplexes and the diameter of each multisimplex $K^y_m \times K^y_n$ of $\mathcal{K}^y \equiv \mathcal{K}^y_m \times \mathcal{K}^y_n$ is at most $\alpha/2$. Take subdivisions $\mathcal{L}^y_m$ and $\mathcal{L}^y_n$ of P^y_m and P^y_n such that, letting $\mathcal{L}^y$ be the multisimplicial complex $\mathcal{L}^y_m \times \mathcal{L}^y_n$, for each player j, the j-th coordinate $\tilde{f}^y_j$ of $\tilde{f}^y$ has a multisimplicial approximation f^y_j from $\mathcal{L}^y \times \mathcal{L}^y \to \mathcal{K}^y_j$. (See [6, Appendix B] for the multisimplicial approximation theorem.) We use f^y to denote $f^y_1 \times f^y_2$.

We emphasize two properties of the multisimplicial approximation. (1) For each j, no vertex of $\mathcal{K}_j$ is the image of a vertex of $\mathcal{L}$ under f^y_j. (2) For $(p^y, q^y) \in A^y$, if there exists a simplex K that contains its image under both π^y and f^y, then it belongs to U^y (and hence q^y belongs to V^y). To see these two claims, observe that, since f^y is a multisimplicial approximation of $\tilde{f}^y$, for any point (p^y, q^y), there exists a multisimplex $\tilde{K}$ that contains its image under both $\tilde{f}^y$ and f^y; hence, $\|f^y(p^y, q^y) - \tilde{f}^y(p^y, q^y)\| \leqslant \alpha/2$. Now, if a point (p^y, q^y) that represents a vertex of $\mathcal{L}$ gets mapped to a point in ∂P^y by f^y, then the distance between $\tilde{f}^y(p^y, q^y)$ and ∂P^y would be at most $\alpha/2$, which is impossible, thus proving (1). As for (2), if p^y and $f^y(p^y, q^y)$ belong to a multisimplex K for $(p^y, q^y) \in A^y$, then $\|p^y - f^y(p^y, q^y)\| \leqslant \alpha/2$, implying that $\|\tilde{f}^y(p^y, q^y) - p^y\| \leqslant \alpha$, which is impossible unless $(p^y, q^y) \in U^y$, proving (2).

Take a further polyhedral subdivision of $\mathcal{L}^y \times \mathcal{L}^y$ and let $\mathcal{T}^y$ be the set of its full-dimensional polyhedra. Let γ^y be the function generated by $\mathcal{T}^y$, i.e. a piecewise-linear convex function that is linear on and only on each polyhedron in $\mathcal{T}^y$. The construction of such a function is specified in [9, Theorem B.2].

Next we construct a family of metagames $\tilde{\Gamma}^\delta$ in extensive form parameterized by $0 < \delta < 1$ that embed Γ.

6.2. **A Game with Redundant Strategies.** First we construct an extensive-form game $\Gamma(\delta, p^0)$, given $0 < \delta < 1$ and a collection $p^0 = (p^y)_{y \in X^\circ}$ of mixed strategy profiles, where each $p^y \in P^y$. For each player m and each non-equilibrium action a at x that leads to a node $y \in X^\circ_n$, just after m chooses a, she has the option of reconsidering her decision. If m revises her decision, then Nature steps in and with probability $1 - \delta$ implements m's equilibrium action $a^*(x)$ at x and the following continuation in the subgame following $a^*(x)$: For any node $x' \in X^*_m$, it chooses the prescription given by the subgame-perfect equilibrium; if $x' \in X^\circ_m$, then in the subgame $\Gamma_{x'}$, it prescribes the mixture $p^{x'}_m$; and with probability δ

Nature continues with a, thus leading into Γ^y, and implements the strategy prescribed by p^y_m. If m does not revise her decision, then after n moves she makes choices in Γ^y as in the original game.[9]

If m chooses to play into the subgame or Nature does so, then next player n gets to move, knowing only that play is now in the subgame Γ^y, i.e. n knows that at each predecessor $x' \in X_m$ of x, m chose the subgame-perfect equilibrium action $a^*(x')$, and then at x, m chose a, after which possibly she revised her strategy, in which case with probability δ Nature chose to lead play into Γ^y.

Player n chooses one of his pure strategies in S^y_n in a sequential process. First, he provisionally chooses some s^y_n. Then he too gets to reconsider his choice, that is, he can choose to implement s^y_n or not. If he chooses to persist with s^y_n then that strategy is automatically implemented in Γ^y. Or if he chooses to revise his choice then for each pure strategy $t^y_n \in S^y_n$ he can pick a redundant strategy that plays t^y_n with probability $1-\delta$ and with probability δ plays p^y_n, and in either case this mixture is also automatically implemented. As with m's choice and revision, n's choices and revisions are not observed by m, who observes only which nodes of the original subgame Γ_x are reached.

The resulting game $\Gamma(\delta, p^0)$ has the same reduced normal form as Γ because for either player a revision implements a redundant strategy that with probabilities $(1-\delta, \delta)$ chooses one of two continuation strategies available in Γ.

6.3. **The Game Tree for Metagames.** Now we describe the metagame $\tilde{\Gamma}^\delta$ for each $0 < \delta < 1$. The game begins with a collection of seven outsiders o^y_0, $o^y_{m,i}$, and $o^y_{n,i}$ for $i = 1, 2, 3$ for each player n and each $y \in X^\circ_n$, all of whom move simultaneously. Outsider o^y_0 chooses a full-dimensional polyhedron T^y in $\mathcal{T}^y$. For $i = 1, 3$, the pure-strategy sets of outsiders $o^y_{m,i}$ and $o^y_{n,i}$ are the vertex sets V^y_m and V^y_n of $\mathcal{K}^y_m$ and $\mathcal{K}^y_n$, respectively. For $j = m, n$, outsider $o^y_{j,2}$'s pure-strategy set is a finite set $S^{y,\delta}_j$ of points in P^y_j chosen such that every point in P^y_j is within δ of some point in $S^{y,\delta}_j$. For $j = 1, 2$, a pure strategy v^y_j of $o^y_{j,1}$ corresponds to a point $p^y_j(v^y_n)$ in P^y_j; hence, a mixed strategy $\sigma^y_{j,1}$ of $o^y_{j,1}$ induces a mixed strategy that is a point $p^y_j(\sigma^y_{j,1})$ in P^y_j. Likewise, a randomized strategy $\sigma^y_{j,2}$ induces a point $q^y_j(\sigma^y_{j,2})$ in P^y_j.

For outsiders $o^y_{j,1}$ and their choices v^y_j for $j = 1, 2$ and $y \in X^\circ$, let $p^0(v^0)$ be the collection of strategies $p^y_j(v^y_j)$. After each strategy profile of the outsiders in which these particular outsiders choose the profile given by the v^y_j's, there ensues a copy of $\Gamma(\delta, p^0(v^0))$. In the metagame, no insider (a player in N) is informed about choices of outsiders, so an information

[9]Examples of game trees induced by m's reconsideration of a deviation from her subgame-perfect strategy are displayed in Figure 3 and in the lower panel of Figure 4.

set of an insider is the union of the corresponding information sets in the games $\Gamma(\delta, p^0(v^y))$, where the only difference among them is the parameter $p^0(v^0)$.

6.4. **Payoffs in the Metagames.** Each terminal node of Γ^δ is a copy of a terminal node of Γ, and the insiders' payoffs are the same as in Γ. We now describe the payoffs of the outsiders. Fix $y \in X^\circ$. Suppose $y \in X^\circ_n$, payoffs of the y-outsiders are as follows.

The convex function γ^y is linear over each polyhedron T^y in the subdivision and has a unique linear extension over $P^y \times P^y$ denoted $\gamma^y_{T^y}$. The payoffs of o^y_0 depend on the choices of $o^y_{m,i}$ and $o^y_{n,i}$ for $i = 1, 2$ as follows. Each profile of mixed strategies of these players induces a point (p^y, q^y) in $P^y \times P^y$ and the payoff to o^y_0 from choosing T^y is $\gamma^y_{T^y}(p^y, q^y)$.

For $j = 1, 2$, outsider $o^y_{j,1}$ wants to mimic $o^y_{j,3}$: the payoff to $o^y_{j,1}$ if he chooses vertex v^y_j and $o^y_{j,2}$ chooses w^y_j is 1 if $v^y_j = w^y_j$ and zero otherwise.

For $j = 1, 2$, the payoff of $o^y_{j,3}$ depends on all the other y-outsiders and is defined as follows. For each pure strategy T^y of o^y_0, there exists a unique multisimplex $L^y \times \tilde{L}^y$ of $\mathcal{L} \times \mathcal{L}$ that contains it. For each pure strategy v^y_j of $o^y_{j,3}$ and each vertex w of $L^y \times L^y$, define $u^\delta_{j,2}(T^y, v^y_j, w)$ to be 1 if v^y_j is the image of w under the j-th coordinate f^y_j of f^y and zero otherwise. The function $u^\delta_{j,2}(T^y, v^y_m, \cdot)$ extends uniquely to a multilinear function over $P^y \times P^y$ since $L^y \times \tilde{L}^y$ is full-dimensional. Now when $o^y_{j,2}$ plays v^y_j, o^y_0 plays T^y, and for $k = 1, 2$ and $l = 1, 2$, $o^y_{k,l}$ plays a mixed strategy $\sigma_{k,l}$, the payoff of $o^y_{j,2}$ is $u^\delta_{j,2}(T^y, v^y_j, (p, q))$ where for $k = 1, 2$, $p^y_k = p^y_k(\sigma_{k,1})$ and $q^y_k = q^y_k(\sigma_{k,2})$.

Finally, we describe payoffs to the second set of y-outsiders. The ambient space of P^y_m and P^y_n is the space $\mathbb{R}^{Z_y}$, where Z_y is the set of terminal nodes of the subgame Γ^y. Let $\varphi^y : \mathbb{R}^{Z_y} \to \mathbb{R}$ be the function given by $\varphi^y(r^y) = -\sum_{z \in Z_y} r^2_z$. For each $r \in \mathbb{R}^{Z_y}$, let $\xi(r; \cdot)$ be the affine approximation to φ^y at r, i.e. for each r', $\xi(r, r') = \sum_{z \in Z_y}(-r^2_z + 2(r_z - r'_z))$. For outsider $o^y_{n,2}$, his payoff depends on the choices of the insiders and player $o^y_{n,1}$. His payoffs are uniformly zero unless the play in the game has the following history. The original players choose all the moves leading to x, player m chooses a leading to y, and then (regardless of whether he chooses to revise his strategy or not), play leads to the subgame at y. In this case, if $o^y_{n,2}$ chooses a pure strategy $s^{\delta,y}_{n,2}$, his payoffs are defined as follows: when player n chooses s^y_n without revision, his payoff is $\xi(s^{y,\delta}_{n,2}, s^y_n)$. If player n revises his choice to the redundant strategy that uses t^y_n, then if $o^y_{n,1}$ had chosen p^y_n, his payoff is $\xi(s^{y,\delta}_{n,2}, (1-\delta)t^y_n + \delta p^y_n)$. Thus $o^y_{n,2}$ wants to mimic the actual choice of n, in the sense that if the final strategy of n that gets implemented is q^y_n, then $o^y_{n,2}$'s best replies are the points in $S^{y,\delta}_n$ that are closest to q^y_n (under the l_2 distance) and thus are all within δ of q^y_n.

The payoff to outsider $o^y_{m,2}$ is more complicated because player m does not choose a normal-form strategy in Γ^y. His payoff depends on the choices of $o^y_{n,1}$, $o^y_{m,1}$ and the insiders. His payoffs are uniformly zero unless the play of the game has the following choices: (i) $o^y_{n,1}$ chooses a vertex v^y_n that does not belong to the boundary of P^y_n; (ii) the original players choose the actions at nodes preceding x that lead to x; m chooses a, which leads to y, and then either chooses not to revise her choice or Nature's choice leads back into Γ^y; (iii) player n chooses to revise his choice, and Nature implements the p^y_n-part of the mixture, i.e. the history has Nature not implementing the t^y_n part (the part implemented with probability $(1-\delta)$). In the exceptional cases satisfying these three conditions, if $o^y_{m,2}$ chose a pure strategy $s^\delta_{m,2}$, then his payoff at the terminal node z is $\xi(q^y_m(s^\delta_{m,2}),z)/p^y_n(v^y_n;z)\beta_*(z)$, where $\beta_*(z)$ is the probability in Γ that Nature does not exclude z and $p^y_n(v^y_n;z)$ is the probability that $p^y_n(v^y_n)$ does not exclude z in Γ^y.

The resulting game $\tilde\Gamma^\delta$ is a metagame that embeds the original game Γ. As in Proposition 3.5, the players retain in $\tilde\Gamma^\delta$ equivalent versions of all their strategies and payoffs available in Γ. Additional strategies obtained upon reconsideration of a choice introduce only redundant strategies. In particular, Nature's action after m's deviation implements m's redundant strategy that is a $(1-\delta,\delta)$ mixture of her subgame-perfect strategy s^*_m and this same strategy up to x that chooses a at x and then follows with $p^y_m(v^y_m)$. Similarly, after n's reconsideration rejects his provisional choice of some s^y_n, Nature's action implements his redundant strategy that is a $(1-\delta,\delta)$ mixture of t^y_n and $p^y_n(v^y_n)$ in the continuation from y. Outsiders' actions affect insiders' payoffs only via effects on the availability of these redundant strategies. Thus, the players' have larger sets of strategies in the metagame $\tilde\Gamma^\delta$ than in the original game Γ, but only because they can opt for redundant strategies determined by outsiders' actions.

6.5. **Equilibrium Strategies of the Outsiders.** Axioms B and S require that any solution of the metagame $\tilde\Gamma^\delta$ contains a sequential equilibrium, say $\tilde b^\delta$ represented in behavioral strategies, whose equivalent profile of mixed strategies has an image in P that is contained in the solution $\bar P$. For each n and $y\in X^\circ_n$, use $\sigma^{\delta,y}_i$ to denote the strategy of outsider i. For $j=m,n$, the strategies of outsiders $(j,1)$ and $(j,2)$ induce points $p^{\delta,y}_j \equiv p_j(\sigma^{\delta,y}_{j,1})$ and $\tilde q^{\delta,y}_j \equiv q_j(\sigma^{\delta,y}_{j,2})$, respectively. Let $p^{\delta,y}=(p^{\delta,y}_m,p^{\delta,y}_n)$ and $\tilde q^{\delta,y}=(\tilde q^{\delta,y}_m,\tilde q^{\delta,y}_n)$. Let $\alpha^{\delta,y}_m$ be the probability that m chooses not to revise her decision to play into the subgame Γ^y, and let $r^{\delta,y}_m\in P^y_m$ be the mixed strategy adopted by m after this choice. Let $V^{\delta,y}_m$ and $V^{\delta,y}_n$ be the supports of the strategies of outsiders $(m,3)$ and $(n,3)$. Because player n observes only the outcome of m's consideration of revising her choice of a, from his perspective, m's mixed

strategy in the subgame Γ^y is the induced average

$$q_m^{\delta,y} \equiv (1-\beta_m^{\delta,y})p_m^{\delta,y} + \beta_m^{\delta,y} r_m^{\delta,y}\,,$$

where

$$\beta_m^{\delta,y} = \alpha_m^{\delta,y}/[(1-\alpha_m^{\delta,y})\delta + \alpha_m^{\delta,y}]\,.$$

Similarly, let $q_n^{\delta,y}$ be the mixed strategy implemented by n in the subgame Γ^y.

Lemma 6.2. *The equilibrium strategies of the outsiders satisfy the following properties.*

1. *For $j = m, n$, suppose the vertices in $V_j^{\delta,y}$, which is the support of outsider $o_{j,3}^y$'s strategies, span a simplex $K_j^{\delta,y}$ of $\mathcal{K}_j$. Then $p_j^{\delta,y}$ belongs to $K_j^{\delta,y}$.*
2. *If every polyhedron in the support of $\tilde{\sigma}_0^{\delta,y}$, which is outsider o_0^y's equilibrium strategy, contains $(p^{\delta,y}, \tilde{q}^{\delta,y})$, then for $j = 1, 2$, the vertices in $V_j^{\delta,y}$ span a simplex $K_j^{\delta,y}$ that does not have a vertex in ∂P_j^y; moreover, in this case, $f(p^{\delta,y}, \tilde{q}^{\delta,y}) \in K_m^{\delta,y} \times K_n^{\delta,y}$.*
3. *Every polyhedron in the support $\tilde{\sigma}_0^{\delta,y}$ contains $(p^{\delta,y}, \tilde{q}^{\delta,y})$.*
4. *For $j = m, n$, $\tilde{q}_j^{\delta,y}$ is within δ of $q_j^{\delta,y}$.*

Proof of Lemma. For $j = m, n$, outsider $o_{j,1}^y$ wants to mimic outsider $o_{j,3}^y$. So, if the vertices of $V_j^{\delta,y}$ span a simplex $K_j^{\delta,y}$, then the payoff to $o_{j,1}^y$ from choosing a vertex w_j^y is positive if it belongs to $V_j^{\delta,y}$ and zero otherwise. Point (1) follows.

Let $\hat{L} = ((L_m \times L_n) \times (\tilde{L}_m \times \tilde{L}_n))$ be the unique multisimplex that contains $(p^{\delta,y}, \tilde{q}^{\delta,y})$ in its interior. For each polyhedron T^y in the support of o_0^y's strategy, there exists now a full-dimensional multisimplex $\bar{L}$ of $\mathcal{L} \times \mathcal{L}$ that contains T^y. Obviously $\bar{L}$ has $\hat{L}$ as a face. For $j = m, n$, by construction, his payoff from choosing a strategy w_j^y if o_0^y chooses such a T^y and given the strategies of the other outsiders, is positive if it is the image of a vertex of $\hat{L}$ under f and zero otherwise. Also, since no vertex in ∂P_j^y is the image of a vertex of $\mathcal{L}$, no such vertex can be a best reply. Therefore, point (2) follows.

For each polyhedron T^y of of $\mathcal{T}^y$, the payoff from T^y is $\gamma_{T^y}^y(p^y, q^y)$ and by construction, $\gamma_{T^y}^y(p^y, \tilde{q}^y) \leqslant \gamma(p^y, \tilde{q}^y)$ with the inequality being strict iff $(p^y, \tilde{q}^y)$ does not belong to T^y, which proves (3).

Admissibility of $o_{n,2}^y$'s strategy requires that it be a best reply to $q_n^{\delta,y}$. By construction, $\tilde{q}_n^{\delta,y}$ is within δ of $q_n^{\delta,y}$.

As for outsider $o_{m,2}^y$, his strategy $\tilde{q}_m^y$ has to be an admissible best reply against the equilibrium. Let $\hat{\tau}$ be a completely mixed strategy of the others. Observe first that $o_{m,2}^y$'s choice of a reply to $\hat{\tau}$ depends only on the following: the insiders adhere to equilibrium play up to x; at x, m chooses a leading to y and then either Nature or player m leads play to Γ^y; the

point that $o^y_{n,1}$'s strategy induces in P^y_n; the total probability that n revises his strategy in the subgame Γ^y; the actual mixed strategy for m that gets implemented in Γ^y. Specifically, let $(1-\hat{\alpha}^y_m)$ be the probability under $\hat{\tau}$ of player m opting out of the subgame and let $\hat{r}^y_m$ be the strategy under $\hat{\tau}$ that m employs in Γ^y if he does not opt out. Let $\hat{\alpha}^y_n$ be the total probability of n revising his choice under $\hat{\tau}$, and let $\hat{p}_m = p^y_m(\tilde{\tau}^y_{o_{m,1}})$. Also, let $\beta_i(x;\hat{\tau})$ be the probability that node x is enabled by insider i under the strategy $\hat{\tau}$, and let $W^{y,\delta}_n$ be the set of vertices of $\mathcal{K}_n$ that do not belong to ∂P^y_n. The expected payoff of $o^y_{m,2}$ from a strategy $s^{y,\delta}_m$ is then $\beta_m(x;\hat{\tau})\beta_n(x;\hat{\tau})$ times

$$\sum_{z\in Z^y}\beta_*(z)\sum_{v^y_n\in W^{y,\delta}_n}\hat{\tau}^y_{o^y_{n,1}}(v^y_n)\hat{\alpha}^y_n\delta p^y_n(v^y_n;z)[(1-\hat{\alpha}^y_m)\delta\hat{p}_m(z)+\hat{\alpha}^y_m\hat{r}^y_m(z)]\xi(s^{y,\delta}_m,z)/p^y_n(v^y_n;z)\beta_*(z)\,,$$

which equals

$$\beta_m(x;\hat{\tau})\beta_n(x;\hat{\tau})\sum_{v^y_n\in W^{y,\delta}_n}\tau_{o^y_{n,1}}(v^y_n)\hat{\alpha}^y_n\delta[(1-\hat{\alpha}^y_m)\delta+\hat{\alpha}^y_m]\xi(s^{y,\delta}_m,\hat{q}^y_m)\,,$$

where

$$\hat{q}^y_m=[(1-\hat{\alpha}^y_m)\delta\hat{p}_m+\hat{\alpha}^y_m\hat{r}^y_m]/[(1-\hat{\alpha}^y_m)\delta+\hat{\alpha}^y_m]$$

is the average strategy of m that is implemented in the subgame Γ^y. By construction, the best replies for $o^y_{m,2}$ against $\hat{\tau}$ are those points that are within δ of $\hat{q}^y_m$. Thus if we have a sequence of such completely mixed strategies $\hat{\tau}$ converging to our equilibrium, then the corresponding sequence $\hat{q}^y_m$ converges to $q^{\delta,y}_m$, which is the strategy of m that gets implemented in Γ^y under our equilibrium. Thus $\tilde{q}^{\delta,y}_m$ is within δ of $q^{\delta,y}_m$. □

6.6. **Final Step of the Proof.** Take a sequence of δ's converging to zero and a corresponding sequence of sequential equilibria $\tilde{b}^\delta$ in behavioral strategies. By points (1) and (2) of Lemma 6.2, $p^{\delta,y}$ and $f^y(p^{\delta,y},\tilde{q}^{\delta,y})$ belong to the same multisimplex $K^{\delta,y}$. Along a subsequence, this multisimplex is the same, say K^y. By point (3) of Lemma 6.2, $\tilde{q}^{\delta,y}$ and $q^{\delta,y}$ have the same limit, say $q^{0,y}$. Let $p^{0,y}$ be the limit of $p^{\delta,y}$. Then $p^{0,y}$ and $f^y(p^{0,y},q^{0,y})$ belong to K^y. If we show that $(p^{0,y},q^{0,y})$ belongs to A^y, then $(p^{0,y},q^{0,y})$ belongs to U^y, thus $q^{0,y}_n$ belongs to V^y, and the theorem is proved.

Therefore all that remains is to show that $(p^{0,y},q^{0,y})$ belongs to A^y. By point (2) of Lemma 6.2, K^y does not have a vertex in ∂P^y; therefore, for all δ, including $\delta=0$, $p^{\delta,y}$ belongs to the interior of P^y. Recall that $q^{\delta,y}_m=(1-\beta^{\delta,y}_m)p^{\delta,y}_m+\beta^{\delta,y}_m r^{\delta,y}_m$. Hence $q^{0,y}_m$ can be expressed as $(1-\beta^{0,y}_m)p^{0,y}_m+\beta^{0,y}_m r^{0,y}_m$. Obviously $q^{\delta,y}_n$ is a best reply against $q^{\delta,y}_m$. Therefore, $q^{0,y}_n$ is a best reply against $q^{0,y}_m$. For all small δ, the strategy for m of letting Nature play yields nearly the continuation payoff $W^{y,*}_m$ from choosing $a^*(x)$ at x. Therefore, $\beta^{0,y}_m$ is positive only if $r^{0,y}_m$

yields a payoff of $W_m^{y,*}$ against $q_n^{0,y}$. If $\beta_m^{0,y} = 0$ then all strategies in P_n^y yield no more than m's equilibrium continuation payoff $W_m^{y,*}$ against $q_n^{0,y}$; and $q^{0,y}$ is a best reply against $p_m^{0,y}$, which is in the interior of P_m^y. Thus $(p^{0,y}, q^{0,y})$ belongs to A^y and we are done.

Suppose now that $\beta_m^{0,y}$ is positive. Let $b_m^{\delta,y}$ be a sequence of behavioral strategies in Γ^y corresponding to the sequence $r_m^{\delta,y}$ and let $b_m^{0,y}$ be its limit. For each δ, let $\tilde{b}^{\delta,\varepsilon(\delta)}$ be a sequence of completely mixed behavioral strategies converging to $\tilde{b}^\delta$ (the originally specified sequential equilibrium in behavioral strategies of $\tilde{\Gamma}^\delta$) such that $b_m^{\delta,y}$ is a sequentially rational strategy against the beliefs induced by the sequence.

Let $S_n^{0,y}$ be the set of pure strategies for n in Γ^y that are best replies against $q_m^{0,y}$. By the optimality property of a sequential equilibrium, for each small δ, player n avoids choosing a strategy that is not in $S_n^{0,y}$ both when he makes a provisional choice and then at the node where he has an option to revise his strategy, where he would strictly prefer to play one of the duplicates that chooses a strategy in $S_n^{0,y}$ with probability $(1-\delta)$. Since these duplicates result in implementing the completely mixed strategy $p_n^{\delta,y}$ with positive probability, for all such small δ, the beliefs in Γ^y can be obtained from replacing the sequence $\tilde{b}^{\delta,\varepsilon(\delta)}$ with the corresponding sequence of the induced conditional distributions over the strategies in $S_n^{0,y}$ as well as the duplicates. In terms of mixed strategies of the original strategy space, this corresponds to a sequence $(1-\varepsilon(\delta))\hat{q}_n^{\delta,y,\varepsilon(\delta)} + \varepsilon(\delta)p_n^{\delta,y}$ where $\hat{q}_n^{\delta,y,\varepsilon(\delta)}$ has its support in $S_n^{0,y}$ and converges to $q_n^{\delta,y}$. Therefore, $\hat{q}_n^{\delta,y,\varepsilon(\delta)}$ itself can be written as $(1-\lambda(\varepsilon(\delta)))q_n^{\delta,y} + \lambda(\varepsilon(\delta))r_n^{\delta,y,\varepsilon(\delta)}$, for a suitable sequence of $\lambda(\varepsilon(\delta))$ converging to zero and where the support of $r_n^{\delta,y,\varepsilon(\delta)}$ is contained in $S_n^{0,y}$. Rewriting the sequence, we can express it as $(1-\hat{\varepsilon}(\delta))q_n^{\delta,y} + \hat{\varepsilon}(\delta)(\hat{\lambda}(\hat{\varepsilon}(\delta))r_n^{\delta,y,\hat{\varepsilon}(\delta)} + (1-\hat{\lambda}(\hat{\varepsilon}(\delta),\delta))p_n^{\delta,y})$ for a sequence of $\hat{\varepsilon}$ converging to zero and a corresponding sequence of $\hat{\lambda}(\varepsilon(\delta))$. Let $\hat{\lambda}(\delta)r_n^{\delta,y} + (1-\hat{\lambda}(\delta))p_n^{\delta,y}$ be the limit of the sequence $(\hat{\lambda}(\hat{\varepsilon}(\delta),\delta)r_n^{\delta,y,\hat{\varepsilon}(\delta)} + (1-\hat{\lambda}(\hat{\varepsilon}(\delta),\delta))p_n^{\delta,y})$. Then $b_m^{\delta,y}$ is sequentially rational against the beliefs induced by the sequence $(1-\hat{\varepsilon})q_n^{\delta,y} + \hat{\varepsilon}(\hat{\lambda}(\delta)r_n^{\delta,y} + (1-\hat{\lambda}(\delta))p_n^{\delta,y})$. Taking the limit of $\lambda(\delta)$, $r_n^{\delta,y}$ and $p_n^{\delta,y}$ as δ goes to zero, and denoting them by λ, $r_n^{0,y}$ and $p_n^{0,y}$, respectively, we get that $b_m^{0,y}$ is sequentially rational against the beliefs induced by $(1-\varepsilon)q_n^{0,y} + \varepsilon(\hat{\lambda}r_n^{0,y} + (1-\hat{\lambda})p_n^{0,y})$ for a sequence of ε's converging to zero.

It remains to prove that $r_n^{0,y}$ is an admissible best reply against $q_m^{0,y}$ if $\hat{\lambda} > 0$. Observe that by construction, the support of $r_n^{0,y}$ is contained in $S_n^{0,y}$, the set of strategies that are best replies against $q_m^{0,y}$. Furthermore, $q_m^{0,y}$, which equals $\beta_m^{0,y}r_m^{0,y} + (1-\beta_m^{0,y})p_m^{0,y}$, is completely mixed, since $p_m^{0,y}$ is completely mixed: indeed, otherwise $\beta_m^{0,y}$ is zero, which implies that there exists a continuum of equilibria where player m randomizes at x between his equilibrium

play and choosing x followed with $r_m^{0,y}$, which is impossible because Γ is generic. Thus $q_m^{0,y}$ is completely mixed, so $r_n^{0,y}$ is admissible, and we are done. □

7. Concluding Remarks

Axiom S excludes refinements from depending on embeddings in metagames. The motivation for this axiom is to prevent refinements from being sensitive to presentation effects. Yet Theorem 6.1 shows that for PI games this axiom requires a solution to contain all admissible equilibria in the same component as the subgame-perfect equilibrium. Each equilibrium in this solution is included because it could occur as the insiders' strategies in a sequential equilibrium of a metagame in which the PI game is embedded.

One could argue that this conclusion contradicts the motivation for the axiom since, other than the subgame-perfect equilibrium, the admissible equilibria in the stable set are included precisely because they are occur as sequential equilibria of a metagame, which is a particular presentation effect. The implication we see is that the selection of a particular equilibrium can stem from an associated class of embeddings, but if the PI game is specified in isolation, without restricting the possible embeddings, then a refinement cannot exclude any equilibrium in the solution. Sufficiently rich detail about how the game is embedded might select a unique equilibrium of the game among insiders, but absent such context, one needs more information to select any proper subset of the solution.

One could also argue that the theorem is uninteresting because all equilibria in the solution of a PI game have the same outcome, namely the outcome of the subgame-perfect equilibrium. In this view, players' strategies after one deviates are irrelevant except that they must sustain players' incentives to stay on paths of equilibrium play. Our view is that it is important to understand how rational behavior is conditioned by one player's interpretation of the other's deviation, that is, by the beliefs that sustain the equilibrium in the ensuing subgame.

To illustrate, we repeat here an example in [7] that invokes only invariance, which is a weaker restriction than Axiom S. Figure 4 shows at the top a PI game Γ in which players 1 and 2 alternate moves. In the subgame-perfect equilibrium each player chooses down at each opportunity, which we represent by the pure strategy D, ignoring his subsequent choice were the player to deviate. There is a single component of the Nash equilibria in which 1 uses D and 2 uses any mixed strategy for which the probability of D is $\geqslant 2/3$. The component of admissible equilibria requires further that 2's probability of a is zero.

Figure 4 shows at the bottom the metagame $\tilde{\Gamma}^\delta$ in which player 1 can reject D and then upon reconsideration choose either the redundant strategy $x(\delta)$, which is a mixture

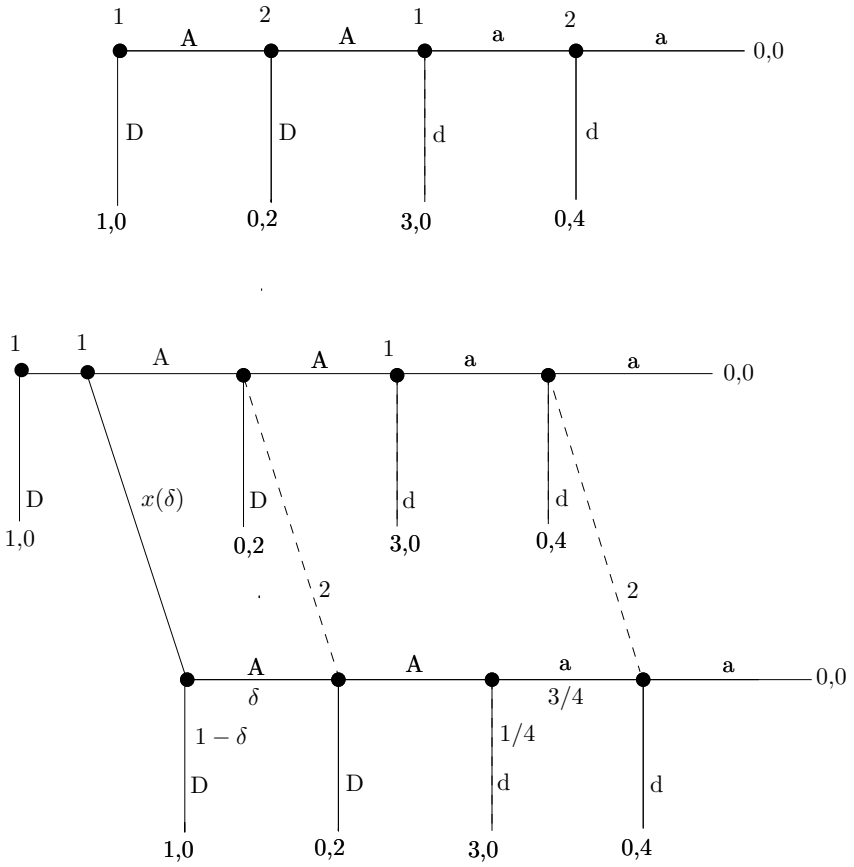

FIGURE 4. Top: A game Γ between players 1 and 2. Bottom: The metagame obtained by allowing player 1 to commit to the redundant strategy $x(\delta)$ after rejecting D.

$(1-\delta, \delta/4, 3\delta/4)$ of D, d, a, where $0 < \delta < 1$, or continue into the subgame by choosing A and then later d or a if 2 chooses A. The two information sets of player 2 indicate that 2 cannot know whether 1 chose A or $x(\delta)$. The reduced normal form of the metagame is the same as the original, since $x(\delta)$ is a redundant strategy.

There is a unique sequential equilibrium in the metagame in which 1 chooses D and 2 randomizes between D and d with probabilities $\alpha(\delta)$ and $1-\alpha(\delta)$, where $\alpha(\delta) = [8+\delta]/[12-3\delta]$. This is sustained by 2's belief at his first information set that the conditional probability that 1 chose $x(\delta)$ given that she rejected D is $\beta(\delta) = 2/[2+\delta]$. By Bayes' Rule, the conditional probability that 1 chose $x(\delta)$ given that A occurred is $p = 2/3$.

A refinement that includes the sequential equilibrium of each metagame $\tilde{\Gamma}^\delta$ must therefore include every profile $(D; \alpha(\delta), 1-\alpha(\delta), 0)$ as δ varies between zero and one. Since $\alpha(0) = 2/3$ and $\alpha(1) = 1$ this requires the refinement to select the entire component of admissible equilibria, which is the stable set.

This example illustrates that each equilibrium in the stable set is sustained by 2's belief derived from a particular hypothesis, namely, embedding in the metagame $\tilde{\Gamma}^\delta$ for some particular value of the parameter δ.

Appendix A. Proof of the Pseudo-Manifold Property

This appendix proves Proposition 5.2.

Proposition 5.2: For each player n and node $y \in X_n^\circ$, $(A^y, \partial A^y)$ is a pseudo-manifold with boundary and has the same dimension as P^y. Moreover, the projection map $\pi^y : (A^y, \partial A^y) \to (P^y, \partial P^y)$ has degree one.

Proof. Fix $y \in X_n^\circ$. Fix a pair $T = (T_m^y, T_n^y)$ of subsets of pure strategies such that T_n^y is nonempty, while T_m^y could be empty, and with the following additional properties: (i) for each player i, the strategies in T_i span a face of P_i^y, say $P_i^y(T_i^y)$, and all points on this face are admissible strategies; (ii) there exist points in $P_n^y(T_n^y)$ against which the strategies in T_m^y all give a payoff of $W_m^{y,*}$ and strategies in $S_m^y \setminus T_m^y$ give no more than $W_m^{y,*}$.

If T_m^y is empty then let $C_n(T)$ be the set of points in $P_n^{y,*} \cap P_n^y(T_n^y)$; by genericity, all other strategies yield m strictly less than $W_m^{y,*}$ against every point in the relative interior of $C_n(T)$. If T_m^y is nonempty then let $C_n(T)$ be the set of points in $P_n(T_n^y)$ against which all the strategies in T_m^y are optimal in the subgame Γ^y and yield $W_m^{y,*}$. Let $\tilde{T}_n^y$ be the vertices of the maximal face of $P_n^y(T_n^y)$ whose interior intersects $C_n(T)$. Using (ii) above and the genericity of payoffs, then, there exist points in the interior of $P_n^y(T_n^y)$ against which all the strategies in P_m^y yield strictly less than $W_m^{y,*}$, and hence $\tilde{T}_n^y = T_n^y$ if T_m^y is empty. Let $\tilde{T}_m^y$ be a minimal subset of T_m^y with respect to the following property: for each strategy $p_n \in P_n^y(\tilde{T}_n^y)$, the probability distribution induced by the strategies in $T_m^y \setminus \tilde{T}_m^y$ are affine combinations of those induced by strategies in $\tilde{T}_m^y$. By genericity of payoffs, the dimension of $C_n(T)$ is then $c_n(T) = l_n(\tilde{T}_n^y) - |\tilde{T}_m^y|$, where $l_n(\tilde{T}_n^y)$ is the dimension of $P_n^y(\tilde{T}_n^y)$ and $|\tilde{T}_m^y|$ is zero if T_m^y is empty.

If T_m^y is empty then let $X_n(T_m^y)$ be the set of points in P_n^y against which all the strategies in P_m^y yield no more than $W_m^{y,*}$. If T_m^y is nonempty then let $\bar{T}_m^y$ be the set of strategies s_m^y in $S_m^y \setminus T_m^y$ that are equally as good replies as strategies in T_m^y against all points in $C_n(T)$. Since the strategies in T_m^y are admissible, the set $X_n(T_m^y)$ consisting of the points q_n against which the strategies in T_m^y are equally good replies, yield $W_m^{y,*}$, and are at least as good as strategies in $\bar{T}_m^y$, has a nonempty intersection with the interior of P_n^y. By genericity of payoffs, $X_n(T_m^y)$ has dimension $d_n(T_m^y) \equiv l_n^y - l_m^y(T_m^y) - 1$, where l_n^y and $l_m^y(T_m^y)$ are respectively the dimensions of P_n^y and $P_m(T_m^y)$.

Let $B_n(T)$ be the set of points in P_n^y of the form $\lambda q_n + (1-\lambda) r_n$, where $q_n \in X_n(T_m^y)$, $r_n \in P_n(T_n^y)$, and $\lambda \geqslant 1$. Observe that if T_m^y is empty then $B_n(T)$ is P_n^y. Indeed, in this case pick a point r_n in the interior of $C_n(T)$. All strategies in P_m^y yield strictly less than $W_m^{y,*}$ against r_n. Therefore, for each $p_n \in P_n^y$ and each small $0 < \lambda < 1$, these strategies still yield less than $W_m^{y,*}$ against $(1-\lambda) r_n + \lambda p_n$; thus $p_n \in B_n(T)$. The following lemma characterizes the nature of the set $B_n(T)$ when T_m^y is nonempty. Here $l_n^y(T_n^y)$ is the dimension of the face $P_n^y(T_n^y)$.

Lemma A.1. *Suppose T_m^y is nonempty. $B_n(T)$ is a nonempty polyhedron of dimension $d_n(T_m^y) + l_n^y(T_n^y) - c_n(T)$. Each maximal face $B_n'(T)$ of $B_n(T)$ satisfies exactly one of the following:*

(i) *Its relative interior is contained in the relative interior of a maximal proper face of P_n^y.*

(ii) *There exists a strategy $s_m^y \in \bar{T}_m^y$ such that s_m^y is as good a reply against every $q_n \in X_n(T_m^y)$ for which $(1-\lambda) r_n + \lambda q_n$ belongs to $B_n'(T)$ for some $\lambda > 1$ and $r_n \in P_n^y(T_n y)$; moreover, in this case, $P_m^y(T_m^y)$ is a maximal proper face of the smallest face of P_m^y that includes the strategies T_m^y and s_m^y.*

(iii) *There exists a maximal proper face of $P_n^y(T_n^y)$ such that for each $q_n \in X_n(T_m^y)$, $r_n \in P_n^y(T_n^y)$ and $\lambda > 1$, if $\lambda q_n + (1-\lambda) r_n$ belongs to $B_n'(T)$, then r_n belongs to this face; moreover in this case, letting T' be the vertices of this face, $C_n(T) = C(T')$.*

Proof of Lemma. $B_n(T)$ is a subset of the affine space generated by $X_n(T)$ and $P_n^y(T_n^y)$. This affine space has dimension $d_n(T) + l_n^y(T_n^y) - c_n^y(T)$, since the intersection of $X_n(T)$ and $P_n^y(T_n^y)$ is exactly $C_n(T)$. By admissibility, there exists a point $q_n \in X_n(T) \setminus \partial P_n^y$. There now exist points p_n arbitrarily close to such a q_n such that $\lambda p_n + (1-\lambda) r_n$ belongs to $X_n(T)$ for some $0 < \lambda < 1$ and $r_n \in P_n^y(T_n^y)$. Clearly p_n has a neighborhood in the affine space generated by $X_n(T)$ and $P_n^y(T_n^y)$ that is contained entirely in $B_n(T)$. Hence $B_n(T)$ is nonempty and has dimension $d_n(T) + l_n^y(T_n^y) - c_n(T)$, as asserted. We now show that $B_n(T)$ is actually a polyhedron whose maximal faces satisfy the properties of the lemma.

Choose a basis $q_{n,i}$, $i = 0, \ldots, d_n(T)$ for the affine space spanned by $X_n(T_m^y)$ such that $q_{n,i}$ belongs to $X_m(T) \setminus C_n(T))$ for $i \leqslant d_n(T) - c_n(T) - 1$ and it belongs to $C_n(T)$ otherwise. Choose vectors $q_{n,i}$ for $i = d_n(T) + 1, \ldots, d_n(T) + l_n(T_n^y) - c_n(T)$ such that the vectors $q_{n,i}$ for $i \geqslant d_n(T) - c_n(T)$ span the affine space generated by $P_n^y(T_n^y)$. Let $\bar{B}_n(T)$ be the set of pairs $(\lambda, \mu) \in \mathbb{R}^{d_n(T)} \times \mathbb{R}^{l_n^y(T_n^y)}$ such that $\bar{\lambda} - \bar{\mu} = 1$, $\bar{\lambda} \geqslant 1$, where $\bar{\lambda} = \sum_i \lambda_{n,i}$ and $\bar{\mu} = \sum_i \mu_{n,i}$. For each $(\lambda, \mu) \in \bar{B}_n(T)$, let $q_n(\lambda) = \sum_{i \leqslant d_n(T))} \lambda_{n,i} q_{n,i}$, $r_n(\mu) = \sum_i \mu_{n,i} r_{n,i}$, let $h(\lambda, \mu) = q_n(\lambda) - r_n(\lambda)$. Let $\tilde{B}_n(T)$ be the set of (λ, μ) such that $r_n(\mu)$ and $h_n(\lambda, \mu)$ belong

to the convex cone spanned by $P_n^y(T_n^y)$ and P_n^y respectively (which immediately implies that $q_n(\mu)$, as a linear combination of $r_n(\lambda,\mu)$ and $h_n(\lambda,\mu)$ also belongs to the convex cone of P_n^y); the strategies in T_m^y yield $W_m^{y,*}$ against $\bar{\lambda}^{-1}q_n(\mu)$ and other strategies yield no more than $W_m^{y,*}$. $B_n(T)$ is now the image of $\tilde{B}_n(T)$ under h. $\tilde{B}_n(T)$ and $B_n(T)$ are now easily seen to be polyhedra.

For a maximal proper face $B_n'(T)$ of $B_n(T)$, exactly one of the following holds uniformly for all p_n in the interior of $B_n'(T)$ and for each (λ,μ) in $h^{-1}(p_n)$: (i) p_n is on the boundary; (ii) one or more of the strategies in $\bar{T}_m^y$ is a best reply against $q_n(\mu)$; (iii) $r_n(\mu)$ belongs to a face of $P_n(T_n^y)$; or (iv) $\sum_i \lambda_i = 1$. For condition (i), genericity of payoffs implies that the boundary has to be a maximal proper face. In the case of (ii) the strategies s_m^y in $\bar{T}_m^y$ that are now equally good replies are such that their union with T_m^y spans a face of P_m^y of which $P_m(T_m^y)$ is a maximal proper face, since otherwise the intersection of the sets where they all yield $W_m^{y,*}$ has a dimension that is at least two less than that of $X_n(T_m^y)$. As for (iii), suppose $C_n(T') \subsetneq C_n(T)$, $q_n(\lambda)$ is in the relative interior of $X_n(T_m^y)$, $\bar{\lambda} \equiv \sum_i \lambda_i > 1$ and $r_n(\mu) \in \partial P_n(T_n^y)$. Pick an r_n^* in $C_n(T) \setminus C_n(T')$; $q_n' \equiv (1-\alpha)q_n(\mu) + \alpha r_n^* = (1-\alpha)((\bar{\lambda})^{-1}(p_n + (1-\bar{\lambda})r_n) + \alpha r_n^*$ belongs to $X_n(T_m^y)$. Since $r_n^* \notin P_n^y(T_n')$, we could not have a boundary point in case (iii) if $C_n(T') \subsetneq C_n(T)$. Condition (iv) implies that $B_n(T) = X_n(T_m^y)$. Let $X_n^+(T_m^y)$ (resp. $X_n^-(T_m^y)$) be the set of points where strategies in T_m are equally good replies, are all better than strategies in $\bar{T}_m$, and yield at least (resp. no more than) $W_m^{y,*}$. The relative interior of $X_n(T_m^y)$ is contained in the relative interior of their union. Moreover, each set intersects $P_n^y(\tilde{T}_n^y)$. For a point $p_n^+ \in X_n^+(T)$, we can choose a point r_n^- in $X_n^-(T) \cap P_n(\tilde{T}_n)$ and then some convex combination of p_n^+ and r_n^- belongs to $X_n(T_m^y)$; thus $p_n^+ \in B_n(T)$. Likewise $X_n^-(T)$ is contained in $B_n(T)$ as well. Since the relative interior of $X_n(T_m^y)$ is in the relative interior of their union, we could not have that $B_n(T') = X_n(T_m^y)$. Thus case (iv) is impossible. □

We need one more lemma concerning these sets $B_n(T)$. Let C' be a maximal face of $C(T)$. Let $\mathcal{T}_n$ be the set of T_n' such that $P_n^y(T_n')$ is a maximal proper face of $P_n^y(T_n^y)$ and $C_n(T') = C'$, where $T' = (T_m^y, T_n')$. Likewise, let $\mathcal{T}_m$ be the set of T_m''s such that $P_m(T_m^y)$ is a maximal proper face of $P_m^y(T_m')$ and $C_n(T') = C'$, where now $T' = (T_m', T_n^y)$. Let S_m' be the set of strategies t_m' such that the face spanned by T_m^y and t_m' equals $P_m'(T_m')$ for some $T_m' \in \mathcal{T}_m$. Let $\mathcal{T}$ be the set of T' of the form (T_m', T_n^y) or (T_m^y, T_n') for $T_m' \in \mathcal{T}_m$ and $T_n' \in \mathcal{T}_n$.

Lemma A.2. *Each $B_n(T')$ is a full-dimensional subset of $B_n(T)$; $B_n(T) = \cup_{T' \in \mathcal{T}} B_n(T')$; and the intersection of $B_n(T') \cap B_n(T'')$ for $T', T'' \in \mathcal{T}$ is a proper face of each. Thus, the $B_n(T')$'s form a polyhedral subdivision of $B_n(T)$.*

Proof of Lemma. If $T' = (T'_m, T^y_n)$, then $d_n(T') = d_n(T) - 1$, $c_n(T') = c_n(T) - 1$, and by the previous lemma, $B_n(T')$ and $B_n(T)$ have the same dimension. If $T' = (T^y_m, T'_n)$, then $d_n(T') = d_n(T)$, $c_n(T')$ = $c_n(T)$ - 1, $l_n(T'_n) = l_n(T^y_n) - 1$, and again the result follows.

We now show that $B_n(T) = \cup_{T'} B_n(T')$. Obviously for each $T' \in \mathcal{T}$, $B_n(T')$ is contained in $B_n(T)$ and $\cup_{T'} B_n(T') \subseteq B_n(T)$. To prove the reverse inequality, given $p_n \in B_n(T)$ expressed as $\lambda q_n + (1-\lambda) r_n$ for some $\lambda \geqslant 1$, $q_n \in X_n(T^y_m)$ and $r_n \in P^y_n(T^y_n)$, suppose there exists $t'_m \in S'_m$ such that t'_m yields at least $W^{y,*}_m$ against q_n. Then, we claim that $p_n \in B_n(T')$ for some $T' = (T'_m, T^y_n)$. For this claim, let r^*_n be a point in the interior of $C_n(T)$. Strategies in S'_m yield strictly less than $W^{y,*}_m$ against r^*_n. For each $q_n(\alpha) \equiv \alpha r^*_n + (1-\alpha) q_n$, let $v(\alpha)$ be the highest payoff from strategies $t'_m \in S'_m$ against $q_n(\alpha)$. $v(0) < W^{y,*}_m$ and $v(1) \geqslant W^{y,*}_m$. There now exists $0 < \alpha \leqslant 1$ such that $v(\alpha) = W^{y,*}_m$. Let t'_m be a strategy in S'_m that achieves $W^{y,*}_m$ against $q_n(\alpha)$; then p_n belongs to $B_n(T'_m, T^y_n)$ where T'_m is the face spanned by T^y_m and t'_m.

Thus, it remains to consider the case where for this p_n and any expression of p_n in the form $\lambda q_n + (1-\lambda) r_n$, the payoff from each $t'_m \in S'_m$ is strictly smaller than $W^{y,*}_m$. In this case, we claim that $\mathcal{T}_n$ is nonempty. Indeed, to see this claim, suppose that $\mathcal{T}_n$ is empty. Then, since C' is a maximal proper face of $C_n(T)$, its interior lies in the interior of $P^y_n(\tilde{T}^y_n)$. Expressing p_n as $\lambda q_n + (1-\lambda) r_n$ in $B_n(T)$, by assumption, the payoff from every strategy in S^y_m is smaller than $W^{y,*}_m$ against q_n. There now exists a point r'_n in the interior of $P^y_n(\tilde{T}^y_n)$ against which the strategies in T^y_m still yield $W^{y,*}_m$ but some strategy in S'_m yields a higher payoff. Let $q_n(\alpha) \equiv \alpha r'_n + (1-\alpha) q_n$ for $0 \leqslant \alpha \leqslant 1$ and let $v(\alpha)$ be the highest payoff from the strategies in S'_m against q^α_n. $v(0) < W^{y,*}_m < v(1)$ and now there exists $0 < \alpha < 1$ such that some strategy in S'_m yields $W^{y,*}_m$ against $q_n(\alpha)$, which by assumption is impossible, since such a $q_n(\alpha)$ is expressible as a convex combination of p_n and a point in $P^y_n(T^y_n)$. Thus $\mathcal{T}_n$ is nonempty.

Since $\mathcal{T}_n$ is nonempty, there exists at least one maximal proper face $P^y_n(T'_n)$ such that $C_n(T') = C'$. And, $C_n(T)$ is not contained in any such face. Choose now r^*_n in the interior of $C_n(T)$. r^*_n does not belong to any $P^y_n(T'_n)$ for $T'_n \in \mathcal{T}_n$. For the given p_n, choose an expression $p_n = \lambda q_n + (1-\lambda) r_n$. We can assume without loss of generality that q_n is completely mixed, if necessary by replacing p_n with a point that is arbitrarily close to it in $B_n(T)$ and proving that this p_n now belongs to $B_n(T^y_m, T'_n)$ for some $T'_n \in \mathcal{T}_n$. For each $0 < \alpha < 1$, now let $q_n(\alpha) = (1-\alpha) r'_n + \alpha q_n$ where r'_n is some point in the interior of C'. Since q_n is completely mixed, $q_n(\alpha)$ is in the interior of P^y_n for all α. Therefore, for each α, there exists a unique $\lambda(\alpha) > 1$ such that $r_n(\alpha) \equiv \lambda(a) q_n(\alpha) + (1 - \lambda(a)) r^*_n$ belongs to the boundary of P^y_n. For α close to zero, $q_n(\alpha)$ is very close to r'_n, which belongs to the boundary of P^y_n (as it belongs to C' which belongs to a face of proper face of $P^y_n(T^y_n)$). Therefore, for such α, $r_n(\alpha)$ belongs

to a face Q_n^y of P_n^y that contains r_n' in in its interior (and hence also C', by virtue of the fact that r_n' belongs to the interior of C'). Q_n^y is then a proper face of $P_n^y(T_n^y)$. And, it cannot contain r_n^*: indeed if it did then $q_n(\alpha)$, as a convex combination of $r_n(\alpha)$ and r_n^*, would belong to Q_n^y as well, which is impossible, since it belongs to the interior of P_n^y. Thus, Q_n^y is a proper face of $P_n^y(T_n^y)$ that contains C' but not $C_n(T)$. Let T_n' be any maximal proper face of $P_n^y(T_n^y)$ that contains Q_n^y but not $C_n(T)$. Obviously p_n now belongs to $B_n(T_m^y, T_n')$. Thus, we have shown that $B_n(T) \subseteq \cup_{T'} B_n(T')$ and in fact that the two sets are equal.

To show that the intersection of two sets $B_n(T')$ is a face of each, it is sufficient to show that for each $B_n(T')$ every maximal proper face of $B_n(T')$ either belongs to the boundary of $B_n(T)$ or is a maximal proper face of exactly one other $B_n(T'')$. Suppose $T' = (T_m', T_n^y)$ and B' is a maximal proper face of $B_n(T')$. By the previous lemma, there are three possibilities. Under case (i) there, B' belongs to the boundary of $B_n(T)$. Under case (ii), let $\hat{T}_m'$ be the vertices of the set spanned by T_m' and this strategy t_m^y identified under (ii). Then $P_n^y(\hat{T}_m')$ has $P_m^y(T_m')$ as a maximal proper face and it in turn has $P_m^y(T_m^y)$ as a maximal proper face. Therefore there exists a subset T_m'' of $\hat{T}_m^y$ such that $P_m^y(T_m'')$ is a maximal proper face of $P_m^y(\hat{T}_m')$ that is different from $P_m^y(T_m')$ and that still contains $P_m^y(T_m^y)$ as a maximal proper face. Let $\hat{T}' = (\hat{T}_m', T_n^y)$ and let $T'' = (T_m'', T_n^y)$. Then, since $C_n(\hat{T}') = C'$, there are two possibilities. Either $C_n(T'') = C_n(T)$ or $C_n(T'') = C'$. In the former case, B' belongs to the boundary of $B_n(T)$. In the latter case, B' is a maximal proper face of $B_n(T'')$. Case (iii) implies that for the face $P_n(T'')$ of $P_n(T_n^y)$ such that for every $p_n \in B_n(T')$ expressed as some $\lambda q_n + (1-\lambda) r_n$, $r_n \in P_n(T_n'')$, $C_n(T_n'') = C_n(T') = C'$. Thus, $T_n'' \in \mathcal{T}_n$ and B' is a face of $B_n(T_m^y, T_n'')$. If B_n' is a face of $B_n(T_m^y, T_n')$, the proof is analogous to the above arguments and hence omitted. □

For each T, let $A_n(T) \equiv B_n(T) \times C_n(T)$. Then $A_n(T)$ is a polyhedron of dimension $l_n^y - l_m^y(T_m^y) + l_n(T_n^y) - 1$.

We turn now to an equivalent analysis of P_m. Again fix the sets (T_m^y, T_n^y) with the same properties as above. Let $X_m(T_n^y)$ be the set of points in P_m^y against which the strategies in T_n^y are best replies. The dimension of $X_m(T_n^y)$ is $d_m = l_m - l_n(T_n^y)$.

If T_m^y is empty, let A_m be the set of (p_m, p_m) such that $p_m \in X_m(T_n^y)$. Otherwise, let $A_m(T)$ be the set of $(p_m, q_m) \in P_m^y \times X_m(T_n^y)$ such that there exists $\lambda \geqslant 1$ such that $(1-\lambda)p_m + \lambda q_m$ belongs to $P_m(T_m^y)$. Observe that this λ is unique unless p_m (and hence also q_m) belongs to $P_m(T_m^y)$. The following is analogous to the previous lemma.

Lemma A.3. *The set $A_m(T)$ is a convex polyhedron of dimension $l_m + l_m(T_m^y) - l_n^y(T_n^y) + 1$. On a maximal proper face A_m' of $A_m(T)$ exactly one of the following inequalities holds*

uniformly for all (p_m, q_m) in A'_m: (i) p_m belongs to a maximal proper face of P^y_m; (ii) there exists $s_n \notin T^y_n$ such that s_n is a best reply against (p_m, q_m); moreover, $P_n(T^y_n)$ is a maximal proper face of P^y_n spanned by T^y_n and this strategy s_n; (iii) there exists a maximal proper face of $P^y_m(T^y_m)$ (which is empty if T^y_m is a singleton) such that if $q_m = (1-\lambda)p_m + \lambda r_m$, then r_m belongs to this face (and if T^y_m is a singleton, then $q_m = p_m$).

Lemma A.4. *(p, q) belongs to A^y iff there exists T as above such that $((p_m, q_m), (p_n, q_n)) \in A_m(T) \times A_n(T)$.*

Proof of Lemma. Suppose (p, q) belongs to $A_m(T) \times A_n(T)$ for some T. We will show that it belongs to A^y. It is sufficient to show this when (p, q) belongs to the interior of $A_m(T) \times A_n(T)$, since A^y is closed. By this assumption, p belongs to the interior of P^y. The support of q_n is in T^y_n; the strategies in T^y_n are best replies against q_m; and also, there exists $r_n \in P_n(T^y_n)$, $\lambda_n \in (0, 1]$ and $q'_n \in X_n(T^y_m)$ such that $q'_n = (1-\lambda_n)r_n + \lambda_n p_n$. Since the support of r_n and q_n are contained in T^y_n, which are all best replies against q_m, point (iii) of the definition of A^y is satisfied. Thus there remains point (ii). If $p_m = q_m$, there is nothing more to prove. Suppose now that $q_m = (1-\lambda_m)p_m + \lambda_m r_m$ for some $\lambda > 0$ and the support of r_m is in T^y_m. Fix a point r'_n in the relative interior of $C_n(T)$ and consider for fixed $0 < \delta < 1$ (which is to be specified later), the sequence $q(\varepsilon) = (1-\varepsilon)q_n + \varepsilon((1-\delta)r'_n + \delta q'_n)$. By the construction of $A_m(T)$, the support of r_m is contained in T^y_m, and the strategies T^y_m are optimal against q_n and r'_n, both of which belong to $C_n(T)$. Also they do equally well against q'_n and hence against $q_n(\varepsilon)$ for all ε. If s^y_n belongs to $\bar{T}^y_m$, then it does as well as strategies in T^y_m against both q_n and r'_n but no better against q'_n. Finally, if s^y_n does not belong to $\bar{T}^y_m$, then it does no better than T^y_m against q_n and strictly worse than those strategies against r'_n, and thus worse against $q_n(\varepsilon)$ for small ε if δ is sufficiently close to 1. Thus we have shown that (p, q) belongs to A^y.

Suppose now that (p, q) belongs to A^y. We will show that it belongs to $A_m(T) \times A_n(T)$ for some T. Again, it is sufficient to assume that p is in the interior of P^y and (p, q) satisfies the conditions (i)-(iii) of the definition of A^y. Let $q_m = (1-\lambda_m)p_m + \lambda_m r_m$ and let $q'_n = (1-\lambda_n)p_n + \lambda_n r_n$. Let T_n be the support of q_n if $\lambda_n = 0$ and otherwise let it be the union of the supports of q_n and r_n. If $\lambda_m = 0$ then letting T_m be the empty set we see that (p, q) belongs to $A_m(T) \times A_n(T)$.

Suppose now that $\lambda_m \neq 0$. Let T_m be vertices of the face of P^y_m that contains r_m in its interior. Since $(1-\lambda_n)q_n + \lambda_n r_n$ is a best reply against $(1-\lambda_m)p_m + \lambda_m r_m$, which, like p_m is in the interior of P^y_m, it is an admissible best reply against that strategy and thus $(p_m, q_m) \in A_m(T)$. We now have to show that (p_n, q_n) belongs to $A_n(T)$. Let $q_n(\varepsilon) \equiv$

$(1-\varepsilon)q_n + \varepsilon((1-\lambda_n)p_n + \lambda_n r_n)$ be a sequence satisfying condition (ii). Then by weak sequential rationality, the strategies in T_m^y are best replies against q_n and thus q_n belongs to $C_n(T)$. Thus there remains to show that p_n belongs to $A_n(T)$. To do this we need to show that there exists a point of the form $\lambda_n' r_n' + (1-\lambda_n')p_n$ against which the strategies in T_m^y yield $W_m^{y,*}$ and are at least as good replies as those in $\bar{T}_m^y$. In fact it is sufficient to show a weaker statement, one obtained by relaxing the requirement that the common payoff to the strategies in T_m^y is $W_m^{y,*}$. Indeed, suppose the common payoff is some $w < W_m^{y,*}$ (the argument for the other case being analogous) then we can find a point r_n'' in $P_n^y(\tilde{T}_n^y)$ where the strategies in $T_m^y \cup \bar{T}_m^y$ all yield the same payoff and this payoff is strictly greater than $W_m^{y,*}$; an average of the original point and r_n'' now shows that $p_n \in A_n(T)$. Thus, we will show that there exists a point of the form $\lambda_n' r_n' + (1-\lambda_n')p_n$ against which the strategies in T_m^y are equally good replies and are at least as good a replies as those in $\bar{T}_m^y$. Suppose, to the contrary, that this statement is not true. Then, letting K_n be the convex hull of p_n and $P_n^y(T_n^y)$, we see that some strategy r_m^* with support contained in T_m^y is weakly dominated by another strategy $\hat{r}_m$ whose support is contained in $T_m^y \cup \bar{T}_m^y$ when we restrict n to the set K_n of strategies. Since all the strategies in $T_m^y \cup \bar{T}_m^y$ yield the same payoff against which q_n, which belongs to the interior of, say, $P_n^y(\bar{T}_n^y)$, the strategies r_m^* and $\hat{r}_m$ yield the same payoff against every strategy in $P_n^y(\bar{T}_n^y)$, since otherwise, r_m^* would not be dominated by $\hat{r}_m$. Therefore, by the genericity of the game, the strategies r_m^* and $\hat{r}_m$ induce the same outcome against every strategy in $P_n^y(\bar{T}_n^y)$. Consequently, any node $x \in X_m$ that is not excluded by q_n nor by either r_m^* or $\hat{r}_m$, is enabled by the other as well and the actions prescribed by behavioral strategies equivalent to these two agree at such a node. If $x \in X_m$ is node that is excluded by q_n, then it is enabled by $(1-\lambda_n)p_n + \lambda_n r_n$, since p_n is completely mixed; therefore, by weak sequential rationality of r_m, if x is not excluded by r_m^*, then r_m^* prescribes choices at x that are optimal against $(1-\lambda_n)p_n + \lambda_n r_n$. This implies that r_m^* is at least as good a reply as $\hat{r}_m$ against $q_n(\varepsilon)$ for all ε, which implies that it is not dominated by $\hat{r}_m$ as claimed. Thus $p_n \in A_n(T)$. □

Lemma A.5. *A^y is a pseudo-manifold of dimension $l_m^y + l_n^y$.*

Proof of Lemma. For each T, the dimension of $A_m(R)$ is $l_m + l_m(T_m^y) - l_n(T_n^y) + 1$; that of $A_n(T)$ is $l_n + l_n(T_n^y) - l_m(T_m^y) - 1$, with the convention that $l_m(\emptyset) = -1$. Hence the dimension of A^y is $l_m^y + l_n^y$. We now prove the pseudo-manifold property. To prove this, it is sufficient to show that for each T, each maximal face $A'(T)$ of $A_m(T) \times A_n(T)$ belongs either to ∂A^y or admits a decomposition into finitely many polyhedra $A^1(T), \ldots, A^k(T)$ of the same dimension as $A'(T)$ such that each $A^i(T)$ is a subset of a maximal proper face of

exactly one other $A(T')$. In fact the only sets $A'(T)$ that we need to decompose are of the form $A_m(T) \times (B_n(T) \times C')$ where C' is a maximal proper face of $C(T)$. By Lemma A.2, this set can be written as the union $A_m(T) \times (\cup_{T'} B_n(T') \times C')$; each $A_m(T) \times (B_n(T') \times C')$ is a maximal proper face of $A_m(T') \times A_n(T')$.

Now if $A'(T)$ is a maximal face of $A(T)$, then either there exists a maximal proper face $A'_m(T)$ of $A_m(T)$ such that $A'(T) = A'_m(T) \times A_n(T)$ or there exists a maximal proper face $A'_n(T)$ of $A_n(T)$ such that $A'(T) = A_m(T) \times A'_n(T)$. Consider the former case. Here, by Lemma A.3, for all $(p_m, q_m) \in A_m(T)$ exactly one of the following inequalities hold: (i) $p_m \in \partial P^y_m$; (ii) there exists $s_n \notin T^y_n$ that is a best reply against q_m; (iii) there exists a maximal proper face of $P^y_m(T^y_m)$ such that q_m is a convex combination of p_m and a point on this face. In case (i), $A'(T)$ belongs to ∂A^y and is not a face of any other $A(T)$. In case (ii), letting T'_n be the set of pure strategies that belong to the minimal face of P^y_n that is spanned by T^y_n and s_n, and $T' = (T_m, T'_n)$, we have two possibilities: (a) $C_n(T) = C_n(T')$; (b) $C_n(T) \subsetneq C_n(T')$. In the former case, $B_n(T)$ is a maximal proper face of $B_n(T')$ by property (iii) of Lemma A.1 and hence $A'(T)$ is a maximal proper face of $A_m(T') \times (B_n(T') \times C_n(T))$. In the latter case, as we saw at the end of the last paragraph, $A'(T)$ is a face of $A_m(T') \times (B_n(T') \times C_n(T'))$. In case (iii), letting $T' = (T'_m, T^y_n)$, where $P^y_m(T'_m)$ is the maximal face we again have two possibilities: (a) $C_n(T) = C_n(T')$; (b) $C_n(T) \subsetneq C_n(T')$. Under (a), the strategies in $T_m \setminus T'_m$ belong to $\bar{T}'_m$ (the strategies other than those in T'_m that are best replies against all strategies in $C_n(T')$). Hence, $B_n(T)$ is a maximal proper face of $B_n(T')$ by property (ii) of Lemma A.1, and $A'(T)$ is a maximal proper face of $A_m(T') \times A_n(T')$. Under case (b), the argument is as under case (ii)(b).

In case $A'(T) = A_m(T) \times A'_n(T)$ there are four possibilities: (i) p_n belongs to the boundary of P^y_n; (ii) there exists a strategy $s^y_m \in \bar{T}^y_m$ such that s^y_m is as good a reply against every q_n for which $(1-\lambda)r_n + \lambda q_n$ belongs to $B'_n(T)$; moreover, in this case, $P^y_m(T^y_m)$ is a maximal proper face of the smallest face of P^y_m that includes the strategies T^y_m and s_m; (iii) for all points $p_n \equiv \lambda q_n + (1-\lambda)r_n$, r_n belongs to a maximal proper face $P_n(T'_n)$ and $C_n(T') = C_n(T)$; (iv) for all (p_n, q_n) in A'_n, q_n belongs to a maximal proper face of $C_n(T)$;

In case (i), $A'(T)$ belongs to the boundary of A^y since its projection is to the boundary of P^y. In case (ii), let T'_m be the set of pure strategies that are vertices of the smallest face of P^y_m that is spanned by T^y_m and s_m. Then $A'(T)$ is a face of $A_m(T') \times A_n(T')$, where $T' = (T'_m, T_n)$. In case (iii), $A'(T)$ is a face of $A_m(T') \times A_n(T')$ where $T' = (T^y_m, T'_n)$. In case (iv), the decomposition mentioned above applies: $A'(T)$ is the union of subsets $A'(T') \equiv (A_m(T) \times B_n(T') \times C'_n)$ for $T' \in \mathcal{T}$, with each $A'(T')$ being a subset of $A_m(T') \times A_n(T')$. □

Lemma A.6. *The projection π^y from A^y to P^y has degree one.*

Proof of Lemma. We prove this lemma by showing that there exists an open subset U of P^y such that $(\pi^y)^{-1}(p)$ is a singleton for each $p \in U$. Let b^* be the backward induction equilibrium of Γ^y represented in behavioral strategies. Observe that the payoff to m from the equilibrium b^* in Γ^y is strictly smaller than $W_m^{y,*}$, since in the subgame-perfect equilibrium of Γ, player m avoids the subgame Γ^y. By genericity of payoffs, there exists a small neighborhood $\bar{V}$ of b^* such that for each b in the subset V of $\bar{V}$ consisting of completely mixed behavioral strategies, b^* is the unique best reply and b_m^* yields strictly less than $W_m^{y,*}$ against every point in V. Let U be the subset of $P^y \setminus \partial P^y$ consisting of mixed strategies that are equivalent to some behavioral strategy in V. Then U is an open subset of the interior of P^y. We claim now that for each $p \in U$, $(\pi^y)^{-1}(p) = (p, q)$, where $q_m = p_m$ and q_n^* is the mixed strategy that is equivalent to b_n^*. To prove this, fix $p \in U$ and let q be a point such that $(p, q) \in A^y$. If $q_m = p_m$ then obviously $q_n = q_n^*$ and we are done. Suppose $q_m \neq p_m$; then there exists $\lambda \in (0, 1)$ and r_m such that $q_m = (1 - \lambda)p_m + \lambda r_m$ and r_m is a best reply to q_n and yields payoff $W_m^{y,*}$. We show in this case that $r_m = q_m^*$ and $q_n = q_n^*$, which would imply a contradiction because against b_n^*, the strategy b_m^* does not yield $W_m^{y,*}$. To prove this last point, we show that if r_m (resp. q_n) does not exclude a node of player m (resp. n) then the action prescribed by r_m (resp. q_n) there coincides with the backward induction solution. The proof of this point is by backward induction on the tree. This is obviously true at all end-game nodes. So let y' be a node of one of the players that is not excluded by that player and such that for all nodes following y' our induction hypothesis holds. Suppose y' is a node of m (the other case is analogous). If y' is excluded by q_n then it is enabled by p_n, which is in the interior of P_n^y, and by construction of U, the backward induction choice at y' is the response chosen by r_m. If y' is not excluded by q_n then, by induction, q_n prescribes the same continuation as q_n^*. Thus, the beliefs of player m at y' are that n's play is dictated by some average of q_n^* and p_n. By construction, the backward induction choice is the best reply against either of those two strategies for player n. So, the induction hypothesis holds at y'. □

This concludes the proof of the Proposition. □

References

[1] Blume, L., A. Brandenburger, and E. Dekel (1991): Lexicographic Probabilities and Equilibrium Refinements, Econometrica, 59, 81-98.

[2] Govindan, S., and T. Klumpp (2002): Perfect Equilibrium and Lexicographic Beliefs, International Journal of Game Theory, 31, 229-243.

[3] Govindan, S., and R. Wilson (2001): Direct Proofs of Generic Finiteness of Nash Equilibrium Outcomes, Econometrica, 69, 765-769.

[4] Govindan, S., and R. Wilson (2002): Structure Theorems for Game Trees, Proceedings of the National Academy of Sciences USA, 99, 9077-9080. URL: www.pnas.org/cgi/reprint/99/13/9077.pdf

[5] Govindan, S., and R. Wilson (2002): Maximal Stable Sets of Two-Player Games, International Journal of Game Theory, 30, 557-566.

[6] Govindan, S., and R. Wilson (2005): Essential Equilibria, Proceedings of the National Academy of Sciences USA, 102, 15706-15711. URL: www.pnas.org/cgi/reprint/102/43/15706.pdf.

[7] Govindan, S., and R. Wilson (2006): Sufficient Conditions for Stable Equilibria, Theoretical Economics, 1, 167-206.

[8] Govindan, S., and R. Wilson (2008): Axiomatic Theory of Equilibrium Selection in Signaling Games with Generic Payoffs, Research Report 2000, Stanford Business School, Stanford CA.

[9] Govindan, S., and R. Wilson (2008): Metastable Equilibria, Mathematics of Operations Research, 33, 787-820.

[10] Hillas, J., and E. Kohlberg (2002): Conceptual Foundations of Strategic Equilibrium, in R. Aumann and S. Hart (eds.), Handbook of Game Theory, III, 1597-1663. New York: Elsevier.

[11] Kohlberg, E. (1990): Refinement of Nash Equilibrium: The Main Ideas, in T. Ichiishi, A. Neyman, and Y. Tauman (eds.), Game Theory and Applications. San Diego: Academic Press.

[12] Kohlberg, E., and J.-F. Mertens (1986): On the Strategic Stability of Equilibria, Econometrica, 54, 1003-1037.

[13] Kohlberg, E., and P. Reny (1997): Independence on Relative Probability Spaces and Consistent Assessments in Game Trees, Journal of Economic Theory, 75, 280-313.

[14] Koller, D., and N. Megiddo (1992): The Complexity of Two-Person Zero-Sum Games in Extensive Form, Games and Economic Behavior, 4, 528-552.

[15] Koller, D., N. Megiddo, and B. von Stengel (1996): Efficient Computation of Equilibria for Extensive Two-Person Games, Games and Economic Behavior, 14, 247-259.

[16] Kreps, D., and R. Wilson (1982): Sequential Equilibria, Econometrica, 50, 863-894.

[17] Kuhn, H. (1953): Extensive Games and the Problem of Information, in H. Kuhn and A. Tucker (eds.), Contributions to the Theory of Games, II, 193-216. Princeton: Princeton University Press. Reprinted in H. Kuhn (ed.), Classics in Game Theory, Princeton University Press, Princeton, New Jersey, 1997.

[18] Mertens, J.-F. (1989): Stable Equilibria–A Reformulation, Part I: Definition and Basic Properties, Mathematics of Operations Research, 14, 575-625.

[19] Mertens, J.-F. (1991): Stable Equilibria–A Reformulation, Part II: Discussion of the Definition and Further Results, Mathematics of Operations Research, 16, 694-753.

[20] Mertens, J.-F. (1992): The Small Worlds Axiom for Stable Equilibria, Games and Economic Behavior, 4, 553-564.

[21] Nash, J. (1950): Equilibrium Points in N-Person Games, Proceedings of the National Academy of Sciences USA, 36, 48-49.

[22] Nash, J. (1951): Non-Cooperative Games, Annals of Mathematics, 54, 286-295.

[23] Reny, P. (1992): Backward Induction, Normal Form Perfection and Explicable Equilibria, Econometrica, 60, 627-649.

[24] Savage, L.J. (1954): Foundations of Statistics, New York: John Wiley & Sons.

[25] van Damme, E. (2002): Strategic Equilibrium, in R. Aumann and S. Hart (eds.), Handbook of Game Theory, III, 1523-1596. New York: Elsevier.

Department of Economics, University of Iowa, Iowa City IA 52242, USA.
E-mail address: srihari-govindan@uiowa.edu

Stanford Business School, Stanford, CA 94305-5015, USA.
E-mail address: rwilson@stanford.edu