Econometrica, Submission #6956, revised

ON FORWARD INDUCTION

SRIHARI GOVINDAN AND ROBERT WILSON

ABSTRACT. A player's pure strategy is called relevant for an outcome of a game in extensive form with perfect recall if there exists a weakly sequential equilibrium with that outcome for which the strategy is an optimal reply at every information set it does not exclude. The outcome satisfies forward induction if it results from a weakly sequential equilibrium in which players' beliefs assign positive probability only to relevant strategies at each information set reached by a profile of relevant strategies. We prove that if there are two players and payoffs are generic then an outcome satisfies forward induction if every game with the same reduced normal form after eliminating redundant pure strategies has a sequential equilibrium with an equivalent outcome. Thus in this case forward induction is implied by decision-theoretic criteria.

This paper has two purposes. One is to provide a general definition of forward induction for games in extensive form with perfect recall. As a refinement of weakly sequential equilibrium, forward induction restricts the support of a player's belief at an information set to others' strategies that are optimal replies to some weakly sequential equilibrium with the same outcome, if there are any that reach that information set.

The second purpose is to resolve a conjecture by Hillas and Kohlberg [27, §13.6], of which the gist is that 'invariant backward induction outcomes satisfy forward induction.' An outcome is invariant if every extensive form representing the same strategic situation (i.e. they have the same reduced normal form) has a sequential equilibrium with an equivalent outcome. For a game with two players and generic payoffs we prove that an invariant outcome satisfies forward induction.

The definitions and theorem are entirely decision-theoretic. None of the technical devices invoked in game theory, such as perturbations of players' strategies or payoffs, are needed.[1]

Date: December 2004. Submitted January 2007. Revised January 29, 2008.

Key words and phrases. game theory, equilibrium refinement, forward induction, backward induction.

JEL subject classification: C72.

This work was funded in part by a grant from the National Science Foundation of the United States. We are grateful for superb insights and comments from referee #3.

[1]However, we retain Kreps and Wilson's [33] definition of sequential equilibrium that 'fully consistent' beliefs are limits of beliefs induced by sequences of completely mixed strategies converging to equilibrium strategies.

Sections 1 and 2 review the motivations for backward induction and forward induction. Sections 3 and 4 provide general definitions of forward induction and invariance. The formulation and proof of the theorem are in Sections 5 and 6. Section 7 examines an alternative version of forward induction and Section 8 mentions economic applications.

1. Introduction

We consider a finite game in extensive form specified by a game tree and an assignment of players' payoffs to its terminal nodes. We assume perfect recall, so the game tree induces a decision tree for each player. A pure strategy for a player specifies an action at each of his information sets, and a mixed strategy is a distribution over pure strategies. A mixed strategy induces a behavioral strategy that mixes anew according to the conditional distribution among actions at each information set. Kuhn [34, Thm. 4] established for a game with perfect recall that each behavioral strategy is induced by a mixed strategy, and vice versa, inducing the same distribution on histories of play.

1.1. **Backward Induction.** Economic models formulated as games typically have multiple Nash equilibria. Decision-theoretic criteria are invoked to select among Nash equilibria. For a game in extensive form with perfect recall, the primary criterion is backward induction. Backward induction is invoked to eliminate Nash equilibria that depend on implausible behaviors at information sets excluded by other players' equilibrium strategies. Thus backward induction requires that a player's strategy remains optimal after every contingency, even those that do not occur if all players use equilibrium strategies.

We assume here that backward induction is implemented by sequential equilibrium, essentially as defined by Kreps and Wilson [33, p. 872,882]. A sequential equilibrium is a pair of profiles of players' behavioral strategies and beliefs. Here we define a player's belief to be a conditional probability system (i.e. satisfying Bayes' Rule where well defined) that at each of his information sets specifies a distribution over pure strategies that do not exclude the information set from being reached. A player's belief is required to be 'fully consistent' in that it is a limit of the conditional distributions induced by profiles of completely mixed or equivalent behavioral strategies converging to the profile of equilibrium strategies.

Kreps and Wilson's exposition differs in that for a player's belief at an information set they use only the induced distribution over nodes in that information set. This restriction cannot be invoked here since the purpose of forward induction is to ensure that the support of a player's belief at an information set is confined to others' optimal strategies wherever possible, both before this information set as in their formulation in terms of nodes, and also

subsequently in the continuation of the game. Thus we use throughout the more general specification that a player's belief is over strategies.

The defining feature of a sequential equilibrium is the requirement that in the event an information set is reached the player acting there behaves according to a strategy that in the continuation is optimal given his belief about nature's and other players' strategies. A weakly sequential equilibrium as defined by Reny [41, p. 631] is the same as a sequential equilibrium except that if a player's strategy excludes an information set from being reached then his continuation strategy there need not be optimal. Section 3 provides a formal definition of weakly sequential equilibrium, which is then used in our definition of forward induction.

1.2. **Forward Induction.** Kohlberg and Mertens [29, §2.3] emphasize that refining Nash equilibrium to sequential equilibrium is not sufficient to ensure that behaviors are justified by plausible beliefs. McLennan [35, p. 901] and Reny [41, p. 639] propose different algorithms for iterative elimination of beliefs that are implausible according to some criterion. McLennan defines the set of justifiable equilibria iteratively by excluding a sequential equilibrium that includes a belief for one player that assigns positive probability at an information set to an action of another player that is not optimal in any sequential equilibrium in the restricted set obtained in the previous iteration. Reny defines the set of explicable equilibria iteratively by excluding a belief that assigns positive probability to a pure strategy that is not a best response to some belief in the restricted set obtained in the previous iteration. Essentially, these procedures apply variants of Pearce's [39] iterative procedure for identifying rationalizable strategies to the more restrictive context of sequential equilibria. Somewhat similar is Kohlberg and Mertens' [29, Prop. 6] proof that a stable set of Nash equilibria contains a stable set of the game obtained by deleting a strategy that is not an optimal reply to any equilibrium in the set.

Kohlberg and Mertens label this result Forward Induction but they and other authors do not define the criterion explicitly. The main idea is the one expressed by Hillas and Kohlberg [27, §42.13.6] in their recent survey: "Forward induction involves an assumption that players assume, even if they see something unexpected, that the other players chose rationally in the past," to which one can add that 'and other players will choose rationally in the future.' This is implicit since rationality presumes that prior actions are part of an optimal strategy. See also Kohlberg [28] and van Damme [16].

Studies of particular classes of two-player games with generic payoffs reveal aspects of what this idea entails. (A broader range of economic applications is described in Section 8.) Outside-option games (as in Example 2.1 below) are addressed by van Damme [15, p. 485]. He proposes as a minimal requirement that forward induction should select a sequential

equilibrium in which a player rejects the outside option if in the ensuing subgame there is only one equilibrium whose outcome he prefers to the outside option. Signaling games (as in Example 2.2 below) are addressed by Cho and Kreps [10, p. 202]. They propose an Intuitive Criterion that is refined further by Banks and Sobel [4, §3] to obtain criteria called Divinity and Universal Divinity. These are obtained from iterative application of criteria called D1 and D2 by Cho and Kreps [10, p. 205] using algorithms akin to those of McLennan and Reny.

Briefly, a sequential equilibrium satisfies the Intuitive Criterion if no type of the sender could obtain a payoff higher than his equilibrium payoff were he to choose a non-equilibrium message and the receiver responds with an action that is an optimal reply to a belief that imputes zero probability to nature's choice of those types that cannot gain from such a deviation regardless of the receiver's reply. The D1 criterion requires that after an unexpected message the receiver's belief imputes zero probability to a type of the sender for which there is another type who prefers this deviation for a larger set of those responses of the receiver that are justified by beliefs concentrated on types who could gain from the deviation and response. See also Cho and Sobel [11] and surveys by van Damme [16], Fudenberg and Tirole [18, §11], Hillas and Kohlberg [27], and Kreps and Sobel [32].

Battigalli and Siniscalchi [6, §5] derive the Intuitive Criterion from an epistemic model. They say that a player strongly believes that an event is true if he remains certain of this event after any history that does not contradict this event. They consider a signaling game and a belief-complete space of players' types; e.g. one containing all possible hierarchies of conditional probability systems (beliefs about beliefs) that satisfy a coherency condition. Say that a player expects an outcome if his first-order beliefs are consistent with this outcome, interpreted as a probability distribution on terminal nodes of the game tree. They show that an outcome of a sequential equilibrium satisfies the Intuitive Criterion under the following assumption about the epistemic model:

> The sender (1) is rational and (2) expects the outcome and believes that (2a) the receiver is rational and (2b) the receiver expects the outcome and *strongly believes* that (2b.i) the sender is rational and (2b.ii) the sender expects the outcome and believes the receiver is rational. [6, Prop. 11]

The key aspect of this condition is the receiver's strong belief in the sender's rationality. This implies that the receiver sustains his belief in the sender's rationality after any message for which there exists some rational explanation for sending that message.

These contributions agree that forward induction should ensure that a player's belief assigns positive probability only to a restricted set of strategies of other players. In each case the restricted set comprises strategies that satisfy minimal criteria for rational play.

1.3. **Synopsis.** In Section 2 we illustrate further the motivation for forward induction via two standard examples from the literature. Our analyses of these examples anticipate the theorem in Section 6 by showing that the result usually obtained by 'forward induction reasoning' is implied by the decision-theoretic criterion called invariance. Invariance requires that the outcome should be unaffected by whether a mixed strategy is treated as a pure strategy.

In Section 3 we propose a general definition of forward induction. Its key component specifies *relevant* pure strategies, i.e. those that satisfy minimal criteria for rational play resulting in any given outcome—the induced probability distribution on terminal nodes of the game tree and thus on possible paths of equilibrium play. Our definition says that a pure strategy is relevant if there is some weakly sequential equilibrium with that outcome for which the strategy prescribes an optimal continuation at every information set the strategy does not exclude.[2] We then say that an outcome satisfies forward induction if it results from a weakly sequential equilibrium in which each player's belief at an information set reached by relevant strategies assigns positive probability only to relevant strategies.

In Section 6 we prove for general two-player games with generic payoffs that backward induction and invariance imply forward induction. Thus for such games forward induction is implied by standard decision-theoretic criteria.

2. Examples

In this section we illustrate the main ideas with two standard examples. These examples illustrate how a test for forward induction can reject some sequential equilibria in favor of others. Each example is first addressed informally using the 'forward induction reasoning' invoked by prior authors. The literature provides no formal definition of forward induction and we defer statement of our definition to Section 3, but the main idea is evident from the context. Each example is then analyzed using the decision-theoretic criterion called invariance to obtain the same result. Invariance is defined formally in Section 4 and invoked in Theorem 6.1, but in these examples and the theorem it is sufficient to interpret invariance as requiring only that the outcome resulting from a sequential equilibrium is not affected by adding a redundant pure strategy, i.e. a pure strategy whose payoffs for all players are replicated by a mixture of other pure strategies.

[2]See [21] for a version of this paper that obtains results for the weaker concept of relevant actions, rather than strategies. We are indebted to a referee who provided an example of an irrelevant strategy that uses only relevant actions.

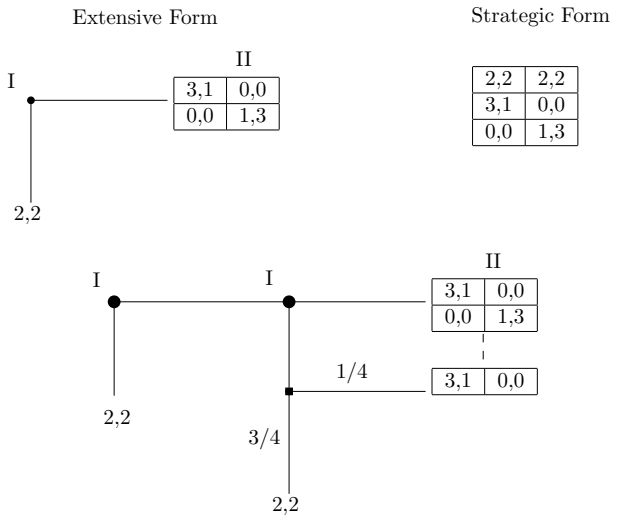

FIGURE 1. Two versions of a game with an outside option

2.1. **An Outside-Option Game.** The top panel of Figure 1 displays the extensive and normal forms of a two-player game consisting of a subgame with simultaneous moves that is preceded by an outside option initially available to player I. The component of Nash equilibria in which player I chooses his outside option includes an equilibrium in which player II's strategy has probability 2/3 of his left column and therefore player I is indifferent about deviating to his top row in the subgame, whereas there is no such equilibrium justifying deviating to the bottom row. Alternatively, player I might anticipate that II will recognize rejection of the outside option as a signal that I intends to choose the top row and therefore II should respond with the left column.

To apply forward induction one excludes from the support of II's belief the dominated strategy in which I rejects the outside option and chooses the bottom row in the subgame. The test fails because then II is sure that if I rejects the outside option then he will play the top row, and therefore II's optimal strategy is the left column, and anticipating this, player I rejects the outside option.

As in Hillas [26, Figure 2], one can invoke invariance to obtain this conclusion. The bottom panel of Figure 1 shows the expanded extensive form after adjoining the redundant strategy in which, after tentatively rejecting the outside option, player I randomizes between the outside option and the top row of the subgame with probabilities 3/4 and 1/4. Player II does not observe which strategy of player I led to rejection of the outside option. In the unique sequential equilibrium of this expanded game player I rejects both the outside option and the redundant strategy and then chooses the top row of the final subgame.

A lesson from this example is that an expanded game has imperfect information in the sense that II has imperfect observability about whether I chose the redundant strategy. This is significant for II because I retains the option to choose the bottom row in the subgame iff he rejected the redundant strategy. Therefore, even though subgame perfection could suffice in the original game, in general one needs sequential equilibrium to analyze the expanded game. Note too that addition of the redundant strategy alters equilibrium strategies but if invariance is satisfied then the induced probabilities of actions along equilibrium paths are preserved and thus so too is the predicted outcome. One could explicitly map equilibria of an expanded game into induced behavioral probabilities of actions in the original game, but we omit this complication.

2.2. **A Signaling Game.** The top panel of Figure 2 displays the two-player two-stage signaling game Beer-Quiche studied by Cho and Kreps [10, §II] and discussed further by Kohlberg and Mertens [29, §3.6.B] and Fudenberg and Tirole [18, §11.2].

Consider sequential equilibria with the outcome QQ-R; that is, both types W and S of player I (the sender) choose Q and player II (the receiver) responds to Q with R and to B with a probability of F that is $\geqslant 1/2$. The equilibria in this component are sustained by player II's belief after observing B that imputes to I's type W a greater likelihood of having deviated than to type S. In all these equilibria, B is not an optimal action for type W. But in the equilibrium for which player II assigns equal probabilities to W and S after observing B and mixes equally between F and R, type S is indifferent between Q and B. If II recognizes this as the source of I's deviation then he infers after observing B that I's type is S and therefore chooses R. Alternatively, if player I's type is S then he might deviate to B in hopes that this action will credibly signal his type, since his equilibrium payoff is 2 from Q but he obtains 3 from player II's optimal reply R if the signal is recognized, but type W has no comparable incentive to deviate—this is the 'speech' suggested by Cho and Kreps [10, p. 180,181] to justify their Intuitive Criterion.

One applies forward induction by excluding from the support of II's belief after observing B those strategies that take action B when I's type is W. In fact, the sequential equilibria in which both types of I choose Q do not survive this restriction on II's belief because II's optimal response to B is then R, which makes it advantageous for player I's type S to deviate by choosing B. Thus sequential equilibria with the outcome QQ-R do not satisfy forward induction. This leaves only sequential equilibria with the outcome BB-R in which both types of player I choose B and II chooses R after observing B.

As in Example 2.1 one can obtain this same conclusion by invoking invariance. The bottom panel of Figure 2 shows the extensive form after adjoining a mixed action X for type S of

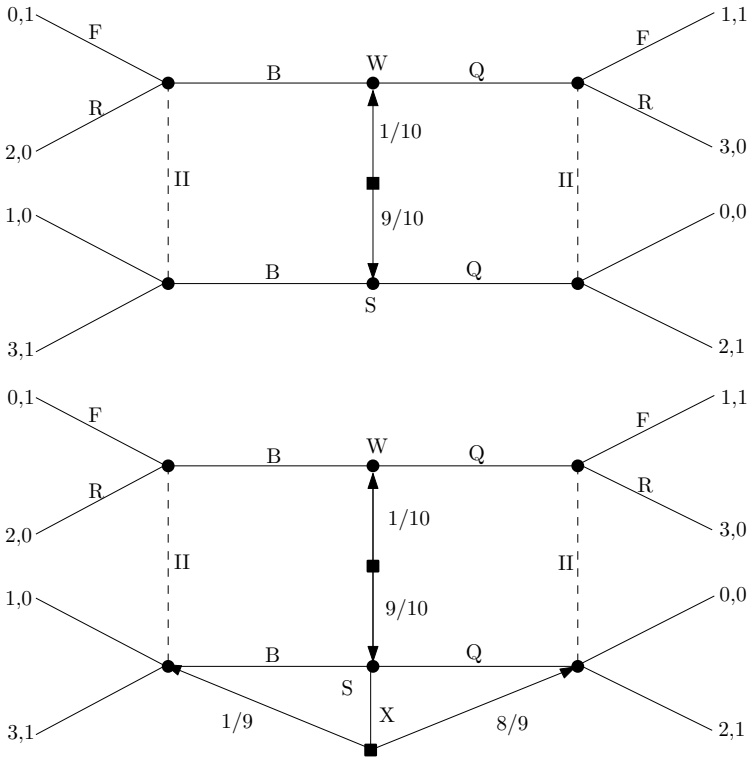

FIGURE 2. Two versions of the Beer-Quiche game

player I that produces a randomization between B and Q with probabilities 1/9 and 8/9. Denote by BQ player I's pure strategy that chooses B if his type is W and chooses Q if his type is S, and similarly for his other pure strategies. The normal form of this expanded game is shown in Table 1 with all payoffs multiplied by 10 (we intentionally omit the pure strategy BX to keep the analysis simple). Now consider the following extensive form that has the same reduced normal form. Player I initially chooses whether or not to use his pure strategy QQ, and if not then subsequently he chooses among his other pure strategies BB, BQ, QB, and QX. After each of these five pure strategies, the extensive form in the bottom panel ensues, but with I's action dictated by his prior choice of a pure strategy. That is, nature chooses I's type to be W or S, the selected pure strategy dictates the subsequent choice of B or Q, and then player II (still having observed only which one of B or Q was chosen) chooses F or R. At player I's information set where, after rejecting QQ, he chooses among his other pure strategies, a sequential equilibrium requires that he assigns zero probability to BQ because it is strictly dominated by QX in the continuation. At player II's information set after observing B a sequential equilibrium requires that his behavioral strategy is an optimal reply to some fully consistent belief about those strategies of player I that reach

		B:	F	F	R	R
W	S	Q:	F	R	F	R
B	B		9,1	9,1	29,9	29,9
B	Q		0,1	18,10	2,0	20,9
Q	B		10,1	12,0	28,10	30,9
Q	Q		1,1	21,9	1,1	21,9
Q	X		2,1	20,8	4,2	22,9

TABLE 1. Strategic form of the Beer-Quiche game with the redundant strategy QX

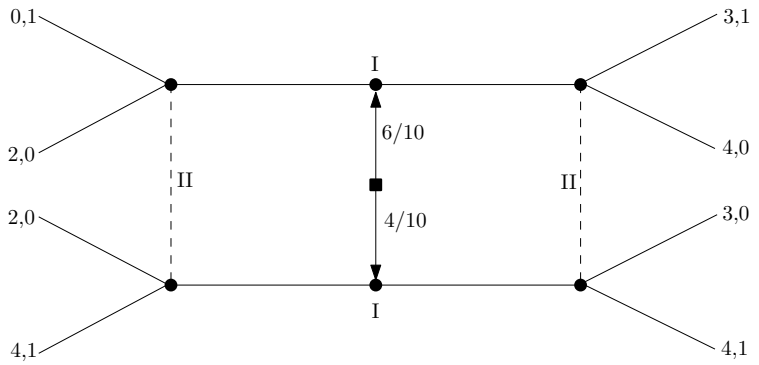

FIGURE 3. A signaling game with pooling and separating equilibria

this information set. But every mixture of I's pure strategies BB, QB, and QX implies that, given his choice of B, the induced conditional probability that his type is S exceeds 9/10. Therefore, player II's reply to B must be R in every sequential equilibrium of this extensive form. Hence the sequential equilibria with outcome QQ-R are inconsistent with invariance, in agreement with failure to satisfy forward induction.

A similar analysis applies to the game in Figure 3, which resembles games considered in studies of signaling via costly educational credentials in labor markets as in Spence [42] and Kreps [30, §17.2]. In this game forward induction rejects the outcome of the pooling equilibrium in which both types of player I move right and II responds to right with up (and to left with probability of up $\geqslant 1/2$), and accepts the outcome of the separating equilibrium in which only I's top type moves right. As mentioned in Section 8, the most common use of forward induction in economic models is to reject pooling equilibria in favor of a separating equilibrium, although often what is actually assumed is a weaker implication of forward induction.

The way in which invariance is invoked in Example 2.2 is indicative of the proof of Theorem 6.1 in Section 6.

3. Definition of Forward Induction

In this section we propose a general definition of forward induction for games in extensive form with perfect recall. Appendix B describes a slightly stronger definition for games in normal form.

Our definition of forward induction relies on the solution concept called weakly sequential equilibrium by Reny [41, p. 631]. Recall from Section 1.1 that a weakly sequential equilibrium is the same as a sequential equilibrium except that a player's strategy need not be optimal at information sets it excludes. See Reny [41, §3,4] for an expanded justification of weakly sequential equilibrium as the right concept for analysis of forward induction.

Our definition differs from Reny's in that we interpret players' beliefs as specifying distributions over others' strategies. Beliefs over strategies typically encode more information than necessary to implement sequential rationality, i.e. as in Kreps and Wilson [33], the conditional distribution over nodes in an information set suffices to verify optimality. However, it is only from a belief specified as a conditional distribution over strategies that one can verify whether a player's belief recognizes the rationality of others' strategies. As Examples 2.1 and 2.2 illustrate, the purpose of forward induction as a refinement is to reject outcomes that deter player I's deviation by the threat of II's response that is optimal for II only because his belief does not recognize I's deviation as part of an optimal strategy for some equilibrium with the same outcome. To reject such outcomes, it is sufficient that the support of II's belief is confined to I's pure strategies that are optimal replies at information sets they do not exclude.

The following definition is the analog of the definitions in Kreps and Wilson [33] and Reny [41].

Definition 3.1 (Weakly Sequential Equilibrium)**.** A weakly sequential equilibrium is a pair (b, μ) of profiles of players' behavioral strategies and beliefs. At each information set h_n of player n his behavioral strategy specifies a distribution $b_n(\cdot|h_n)$ over his feasible actions, and his belief specifies a distribution $\mu_n(\cdot|h_n)$ over profiles of nature's and other players' pure strategies that enable h_n to be reached. These profiles are required to satisfy:

(1) Full Consistency: There exists a sequence $\{b^k\}$ of profiles of completely mixed behavioral strategies converging to b and a sequence $\{\sigma^k\}$ of completely mixed equivalent normal-form strategies such that for each information set of each player the conditional distribution specified by μ is the limit of the conditional distributions obtained from $\{\sigma^k\}$.[3]

[3]The belief $\mu_n(\cdot|h_n)$ might entail correlation; cf. Kreps and Ramey [31].

(2) Sequential Rationality: For each player n and each information set h_n that b_n does not exclude, each action in the support of $b_n(\cdot|h_n)$ is part of a pure strategy that is an optimal reply to $\mu_n(\cdot|h_n)$ in the continuation from h_n.

A sequential equilibrium is defined exactly the same except that each player's actions must be optimal at all his information sets, including those excluded by his equilibrium strategy.

We interpret forward induction as a property of an outcome of the game, defined as follows.

Definition 3.2 (Outcome of an Equilibrium)**.** The outcome of an equilibrium of a game in extensive form is the induced probability distribution over the terminal nodes of the game tree.

A key feature in the definition of forward induction is the concept of a relevant strategy.

Definition 3.3 (Relevant Strategy)**.** A pure strategy of a player is relevant for a given outcome if there is a weakly sequential equilibrium with that outcome for which the strategy at every information set it does not exclude prescribes an optimal continuation given the player's belief there.

Thus a relevant strategy is optimal for some expectation about others' equilibrium play with that outcome, and his beliefs at events after their deviations. For instance, in Example 2.2 of a signaling game, the strategy QB of the sender I in which type W chooses Q and type S chooses B is relevant for the outcome QQ-R because it is an optimal reply to the sequential equilibrium with that outcome in which the receiver II responds to B by using F and R with equal probabilities. But the strategies BB and BQ are irrelevant because B is not an optimal reply for I's type W to any sequential equilibrium with outcome QQ-R.

For the standard examples in Section 2 it is sufficient to interpret forward induction as requiring merely that player II's belief at the information set excluded by I's equilibrium strategy imputes positive probability only to the node reached by I's non-equilibrium relevant strategy. For general games, however, a stronger requirement is desirable.

We propose a general definition of forward induction that identifies those outcomes resulting from the conjunction of rational play and belief that others' play is rational, and thus minimally consistent with Battigalli and Siniscalchi's [6] epistemic model of strong belief in rationality. Because relevant strategies are optimal, hence rational, in some weakly sequential equilibrium with the same outcome, they are the minimal set for which one can require the support of one player's belief to recognize the rationality of other players' strategies—indeed that is the lesson from the standard examples in Section 2. Our proposed definition of a forward induction outcome therefore requires that the outcome results from a weakly

sequential equilibrium in which every player maintains the hypothesis that other players are using relevant strategies throughout the game, so long as that hypothesis is tenable.

To simplify terminology, say that an information set is relevant for an outcome if it is not excluded by every profile of relevant strategies. (This differs from Kuhn's [34, Def. 6] and Reny's [41, p. 631] definition of an information set that is relevant for a pure strategy because the information set is not excluded by that strategy.) Then we define a forward induction outcome as follows.

Definition 3.4 (Forward Induction)**.** An outcome satisfies forward induction if it results from a weakly sequential equilibrium in which at every relevant information set the support of the belief of the player acting there is confined to profiles of nature's strategies and other players' relevant strategies.

Section 7 compares this definition with Reny's alternative interpretation. Applied to the standard examples in Section 2, our definition yields the conclusions obtained from 'forward induction reasoning' in the literature. For instance, in Example 2.2 of a signaling game, the outcome QQ-R does not satisfy forward induction because the definition requires that after observing B player II assigns zero probability to I's irrelevant strategies BB and BQ, and thus assigns positive probability only to I's relevant strategy QB that enables II's information set B to be reached.

More generally, forward induction implies the Intuitive Criterion, D1, D2 (whose iterative version defines Universal Divinity) and Cho and Kreps' [10, § IV.5] strongest criterion, called Never Weak Best Response (NWBR) as defined for signaling games. These implications are verified by showing that a strategy s of the sender is irrelevant if s prescribes that his type t sends a message m that is not sent by any type in some weakly sequential equilibrium with the given outcome, and the pair (t, m) satisfies any of these criteria. For instance, the criterion NWBR excludes the strategy s from the receiver's belief if the continuation strategy m at information set t yields exactly the sender's type-contingent payoff from the given outcome for some beliefs and optimal responses of the receiver only when some other type t' that could send m would get a type-contingent payoff that is higher than from the designated outcome for the same or a larger set of the receiver's optimal responses. But this condition implies that there is no weakly sequential equilibrium with the same outcome for which m is an optimal action for type t. Were there such an equilibrium then the receiver could use any such response at the off-the-equilibrium-path information set m, but then type t' could obtain a superior payoff by sending m. Thus, m cannot be an optimal continuation

by type t in any weakly sequential equilibrium with the given outcome, and therefore s is an irrelevant strategy.

4. Definition of Invariance

In this section we define invariance as a property of a solution concept. First we define relations of equivalence between games and between outcomes of equivalent games.

Recall that a player's pure strategy is redundant if its payoffs for all players are replicated by a mixture of his other pure strategies. From the normal form of a game one obtains its reduced normal form by deleting redundant strategies. Thus the reduced normal form is the minimal representation of the essential features of the strategic situation.

Definition 4.1 (Equivalent Games)**.** Two games are equivalent if their reduced normal forms are the same up to relabeling of strategies.

As specified in Definition 3.2, the outcome of an equilibrium of a game in extensive form is the induced probability distribution on terminal nodes, and thus on the paths through the tree. Associated with each outcome is a set of profiles of nature's and players' mixed strategies that result in the outcome, and in turn each such profile can be replicated by a profile of mixed strategies in the reduced normal form. Hence we define equivalent outcomes as follows.

Definition 4.2 (Equivalent Outcomes of Equivalent Games)**.** Outcomes of two equivalent games are equivalent if they result from the same profile of mixed strategies of their reduced normal form.

Trivially, the outcome of any Nash equilibrium is equivalent to the outcome of a Nash equilibrium of any equivalent game. For any solution concept that is a refinement of Nash equilibrium, we define invariance as follows.

Definition 4.3 (Invariant Outcome)**.** An outcome is invariant for a solution concept if every equivalent game has an equivalent outcome of an equilibrium selected by the solution concept.

This definition is used in Section 6 where the solution concept is sequential equilibrium. Existence of invariant outcomes of sequential equilibria is implied by Mertens' [36] demonstration that one is included among outcomes of a stable set of Nash equilibria.[4]

[4]In [20] we prove that if a solution concept satisfies invariance and a condition called strong backward induction then it selects sets of equilibria that are stable in the weaker sense defined by Kohlberg and Mertens [29, § 3.5].

5. Formulation

In this section we introduce notation used in the proof of the theorem in Section 6.

Let Γ be a game in extensive form with perfect recall. For each player n let H_n be the collection of his information sets, and let S_n, B_n and Σ_n be his sets of pure, behavioral and mixed strategies. A pure strategy chooses an action at each information set in H_n, a behavioral strategy chooses a distribution over actions at each information set, and a mixed strategy chooses a distribution over pure strategies. We say that a pure strategy *enables* an information set if the strategy's prior actions do not exclude the information set from being reached, and similarly for behavioral and mixed strategies.

Let $\Sigma(P)$ and $B(P)$ be the sets of Nash equilibria of Γ represented as profiles of mixed and behavioral strategies, respectively, that result in the outcome P. Let $BM(P)$ be the set of weakly sequential equilibria whose outcome is P, where each $(b, \mu) \in BM(P)$ consists of a profile $b \in B(P)$ of players' behaviorial strategies and a profile μ of players' fully consistent beliefs. As in Definition 3.1, in each weakly sequential equilibrium (b, μ) the belief of a player n at his information set $h_n \in H_n$ is a probability distribution $\mu_n(\cdot|h_n)$ over nature's and other players' pure strategies that reach h_n.

Given an outcome P, say that a P-path through the game tree is one that terminates at a node in the support of P. Actions on P-paths are called equilibrium actions. Let $H_n(P)$ be the collection of player n's information sets that intersect P-paths. Obviously, every equilibrium in $B(P)$ prescribes the same mixture at each information set in $H_n(P)$. Let $S_n(P) \subset S_n$ comprise those pure strategies s_n of player n such that s_n chooses an equilibrium action at every information set in $H_n(P)$ that s_n enables. Note that if $\sigma \in \Sigma(P)$ then the support of σ_n is contained in $S_n(P)$. Moreover, every strategy in $S_n(P)$ is optimal against every equilibrium in $\Sigma(P)$. Partition the complement $T_n(P) \equiv S_n \setminus S_n(P)$ into subsets R_n and Q_n of n's pure strategies that are relevant and irrelevant, respectively, for P. Note that $S_n(P)$ may contain irrelevant strategies.

Define an equivalence relation among player n's pure strategies as follows. Two strategies are equivalent if they prescribe the same action at each information set in $H_n(P)$. Let $\mathcal{E}_n(P)$ be the set of equivalence classes. Denote a typical element of $\mathcal{E}_n(P)$ by E_n and let $E_n(s_n)$ be the equivalence class that contains s_n. Let $\mathcal{E}_n^\circ(P)$ be the subcollection of equivalence classes that contain strategies in $S_n(P)$. Thus, any strategy that is used in some equilibrium in $\Sigma(P)$ belongs to some equivalence class in $\mathcal{E}_n^\circ(P)$, while any strategy that is in $T_n(P)$ does not.

If R_n is not empty then for each probability $\delta \in (0, 1)$ let t_n^δ be a mixed strategy of n of the form $[1-\delta]s_n^* + \delta\rho_n$, where s_n^* is a strategy in $S_n(P)$ and ρ_n is a mixed strategy whose support

is R_n. Since s_n^* is a best reply against every equilibrium in $\Sigma(P)$, t_n^δ is an approximate best reply against equilibria in $\Sigma(P)$ when δ is small, a fact we need in the next section. Define a game G^δ in normal form by adding to the normal form G of Γ the redundant pure strategy t_n^δ for each player n for whom R_n is not empty. In particular, t_n^δ is added iff there is some information set in $H_n(P)$ where some non-equilibrium action is part of a relevant strategy. (In Example 2.1 the redundant strategy t_n^δ for player I is the one shown in Figure 1 with parameter $\delta = 1/4$; and in Example 2.2 it is the strategy QX with parameter $\delta = 1/9$.)

Next we define a game Γ^δ in extensive form whose normal form is equivalent to G^δ, and thus also Γ^δ is equivalent to Γ. A path of play in Γ^δ consists of choices by players in an initial stage, followed by a path of play in a copy of Γ. In the subsequent play of Γ, no player is informed about choices made by other players in the initial stage of Γ^δ. The rules of Γ^δ are the following. If R_n is empty then in Γ^δ player n chooses among all his equivalence classes in $\mathcal{E}_n(P)$ in the initial stage. If R_n is not empty then in the initial stage he first chooses whether to play an equivalence class in $\mathcal{E}_n^\circ(P)$ or not. If he decides to play something in $\mathcal{E}_n^\circ(P)$ then he chooses one of these equivalence classes; or if he chooses not to, then he proceeds to a second information set where he chooses to play either the redundant pure strategy t_n^δ or an equivalence class among those not in $\mathcal{E}_n^\circ(P)$. After these initial stages for all players, Γ^δ evolves the same as Γ does, i.e. a copy of Γ follows each sequence of choices in the initial stage. In Γ^δ the information sets in Γ are expanded to encompass appropriate copies of Γ to represent that no player ever observes what others chose in the initial stage; thus, the information revealed in Γ^δ is exactly the same as in Γ. The information set $h_n \in H_n$ in Γ has in Γ^δ for each $E_n \in \mathcal{E}_n(P)$ an expanded copy $h_n^\delta(E_n)$, and a copy $h_n^\delta(t_n^\delta)$. Nature makes the choice at the expansions of those information sets in $H_n(P)$ (but not at expansions of those in $H_n \setminus H_n(P)$) according to the equivalence class chosen in the initial stage, or at all expansions of information sets in H_n if t_n^δ was chosen. That is, if n chooses t_n^δ at the second information set then nature automatically implements the entire strategy; but if he chooses some equivalence class E_n in $\mathcal{E}_n(P)$ then nature implements actions prescribed by E_n at each $h_n^\delta(E_n)$ when $h_n \in H_n(P)$ and leaves to him to choose at those that are not, if and when they occur.

A pure strategy $s_n \in S_n$ can be implemented in Γ^δ by first choosing $E_n(s_n)$ in the initial stage and then making the choices prescribed by s_n at all $h_n \in H_n \setminus H_n(P)$. And, any strategy in Γ^δ that begins by choosing some equivalence class E_n in the first stage ends up implementing some $s_n \in E_n$. Observe too that the redundant pure strategy t_n^δ, when available, ends up implementing the mixture given by t_n^δ. Thus, it is obvious that G^δ is obtained from the normal form of Γ^δ by deleting some redundant pure strategies in the

latter that are duplicates of other pure strategies. Hence, by a slight abuse of notation, we view G^δ as the normal form of Γ^δ. The game Γ^δ is now easily seen to be equivalent to Γ.

Suppose $h_n \in H_n \backslash H_n(P)$. If an equivalence class E_n contains a pure strategy s_n that enables h_n in Γ, then in Γ^δ the corresponding strategy s_n—i.e. choosing E_n in the initial stage and then making s_n's choices at all $h'_n \notin H_n(P)$—enables $h_n^\delta(E_n)$. Conversely, if E_n does not contain such an s_n then there is an information set $v_n \in H_n(P)$ that precedes h_n, is enabled by E_n, and where E_n makes a choice different from the one that leads to h_n. Thus, in Γ^δ nature's choice at $v_n^\delta(E_n)$ prevents $h_n^\delta(E_n)$ from being reached. Therefore, to analyze the game Γ^δ we need to consider only information sets $h_n^\delta(E_n)$ where E_n contains a strategy that enables h_n in Γ. For simplicity in this section and the next, by an information set $h_n^\delta(E_n)$ in Γ^δ of player n, we mean an $h_n \in H_n \backslash H_n(P)$ and an E_n that contains a strategy that enables h_n in Γ.

Now assume the game has two players. We use m to denote the opponent of player n. Suppose that E_n and E'_n are two equivalence classes that contain strategies that enable some $h_n \notin H_n(P)$. The information that n has at $h_n^\delta(E_n)$ and $h_n^\delta(E'_n)$ about m's choices are the same at both information sets. Therefore, a pure strategy of m in G^δ enables one iff it enables the other. In particular, in a sequential equilibrium $(\tilde{b}^\delta, \tilde{\mu}^\delta)$ of Γ^δ, player n's belief at $h_n^\delta(E_n)$ is independent of E_n and can thus be denoted $\tilde{\mu}_n^\delta(\cdot|h_n)$.

Likewise, suppose $\tilde{\sigma}_m^\delta$ is a mixed strategy of m in G^δ that enables some information set $h_n^\delta(E_n)$ of n. Then $\tilde{\sigma}_m^\delta$ induces a conditional distribution $\tilde{\tau}_m^\delta$ over the pure strategies of m in G^δ that enable $h_n^\delta(E_n)$. Let σ_m and τ_m be the equivalent strategies in G. It is easily checked that τ_m is the conditional distribution induced by σ_m over the pure strategies that enable h_n. And, an action a_n at $h_n^\delta(E_n)$ in Γ^δ is optimal against $\tilde{\tau}_m^\delta$ iff it is optimal against τ_m in Γ.

6. Statement and Proof the Theorem

In this section we show for two-player games with generic payoffs that an invariant backward induction outcome satisfies forward induction.

The notion of genericity we invoke is the following. Let $\mathcal{G}$ be the space of all games generated by assigning payoffs to the terminal nodes of a fixed two-player game tree. In [19] we show that there exists a closed lower-dimensional subset $\mathcal{G}_0$ such that for each game not in $\mathcal{G}_0$ there are finitely many outcomes of Nash equilibria. For technical reasons, in Appendix A we construct another closed lower-dimensional subset denoted $\mathcal{G}_1$. Now a game is generic if it is in the complement of both $\mathcal{G}_0$ and $\mathcal{G}_1$. With this, the formal statement of our theorem is:

Theorem 6.1. *An outcome of a two-player game with perfect recall and generic payoffs satisfies forward induction if it is invariant for the solution concept sequential equilibrium.*

Proof. Assume that Γ is a two-player game in extensive form with perfect recall and generic payoffs. Assume also that P is an invariant sequential equilibrium outcome of Γ, i.e. each game equivalent to Γ has a sequential equilibrium whose outcome is equivalent to P. Because Γ is equivalent to the game Γ^δ defined in Section 5, Γ^δ has a sequential equilibrium $(\tilde{b}^\delta, \tilde{\mu}^\delta)$ whose outcome is equivalent to P. Because $\tilde{\mu}^\delta$ is fully consistent, there exists a sequence $\{\tilde{b}^\delta_\varepsilon\}$ of profiles of completely mixed behavioral strategies that converges as $\varepsilon \downarrow 0$ to $\tilde{b}^\delta$ and a corresponding equivalent sequence $\{\tilde{\sigma}^\delta_\varepsilon\}$ of profiles of completely mixed strategies in the normal form G^δ that converges to some profile $\tilde{\sigma}^\delta$ and such that the belief profile $\tilde{\mu}^\delta$ is the limit of the beliefs derived from the sequence $\{\tilde{\sigma}^\delta_\varepsilon\}$.

Since $\tilde{b}^\delta$ induces an outcome that is equivalent to P, the strategy $\tilde{\sigma}^\delta$, which is equivalent to $\tilde{b}^\delta$, has its support in $S_n(P)$ for each n: indeed, a strategy in $T_n(P)$, or the strategy t^δ_n when available, chooses a non-equilibrium action at some $h_n \in H_n(P)$ in Γ that it enables. Therefore, under $\tilde{b}^\delta$ each player n in the initial stage assigns positive probability only to choices of equivalence classes in $\mathcal{E}^\circ_n(P)$.

Corresponding to the sequence $\{\tilde{\sigma}^\delta_\varepsilon\}$ there is an equivalent sequence $\{\sigma^\delta_\varepsilon\}$ of profiles of mixed strategies in the normal form of Γ for which there is an equivalent sequence $\{b^\delta_\varepsilon\}$ of profiles of behavioral strategies in the extensive form of Γ. Let μ^δ_ε be the profile of beliefs induced by $\sigma^\delta_\varepsilon$. Denote selected limit points of these sequences by σ^δ, b^δ, and μ^δ. By construction, $\sigma^\delta \in \Sigma(P)$, $b^\delta \in B(P)$ and μ^δ is fully consistent. It follows from our remarks at the end of the previous section that for each n and $h_n \notin H_n(P)$: $\mu^\delta_n(\cdot|h_n)$ is equivalent to $\tilde{\mu}^\delta_n(\cdot|h_n)$, and an action at h_n is optimal in Γ against $\mu^\delta_n(\cdot|h_n)$ iff it is optimal in Γ^δ at $h^\delta_n(E_n)$ against $\tilde{\mu}^\delta_n(\cdot|h_n)$ for the corresponding copies in Γ^δ.

Next we argue that (b^δ, μ^δ) is a weakly sequential equilibrium of Γ. Let h_n be an information set of player n that b^δ_n enables. We need to show that the choice made by b^δ_n at h_n is optimal against $\mu^\delta_n(\cdot|h_n)$. If h_n belongs to $H_n(P)$ then $\mu_n(\cdot|h_n)$ is derived from σ^δ_m and obviously b^δ_n chooses optimally at h_n. Suppose that $h_n \notin H_n(P)$. Let a_n be an arbitrary action at h_n that is chosen with positive probability by b^δ_n. Since h_n is enabled by b^δ_n there exists a pure strategy s_n in the support of σ^δ that enables h_n and chooses a_n there. Since σ^δ is equivalent to $\tilde{\sigma}^\delta$, in $\tilde{b}^\delta$ player n with positive probability chooses $E_n(s_n)$ and then makes the choices prescribed by s_n. Sequential rationality of a_n at $h^\delta_n(E_n)$ implies its optimality against $\tilde{\mu}^\delta_n(\cdot|h_n)$. Hence, from the previous paragraph a_n is optimal against $\mu^\delta_n(\cdot|h_n)$ in Γ. Since a_n was arbitrary, this shows that (b^δ, μ^δ) is a weakly sequential equilibrium of Γ.

For some sequence $\delta \downarrow 0$, $(\sigma^\delta, b^\delta, \mu^\delta)$ converges to some limit point (σ, b, μ). Clearly, $\sigma \in \Sigma(P)$, $b \in B(P)$, μ is fully consistent, and $(b, \mu) \in BM(P)$ is a weakly sequential equilibrium of Γ because $BM(P)$ is a closed set.

It remains to prove the forward induction property for the belief profile μ. For each n for whom R_n is not empty, and each δ, let $\{\tilde{\tau}^\delta_{n,\varepsilon}\}$ be the sequence of conditional distributions over $T^\delta_n \equiv T_n(P) \cup \{\, t^\delta_n \,\}$ induced by the sequence $\{\tilde{\sigma}^\delta_\varepsilon\}$ and let $\tilde{\tau}^\delta_n$ be a limit point. The sequence $\{\tilde{\tau}^\delta_{n,\varepsilon}\}$ and therefore its limit are determined by choices made after n chooses in the initial stage to avoid equivalence classes in $\mathcal{E}^\circ_n(P)$. Therefore, the probability of t^δ_n is nonzero under $\tilde{\tau}^\delta_n$ iff t^δ_n is chosen with positive probability at n's second information set in the initial stage; and, the probability of $s_n \in T_n(P)$ is nonzero under $\tilde{\tau}^\delta_n$ iff n chooses $E(s_n)$ with positive probability at this stage and then implements the choices of s_n with positive probability after this choice. Express $\tilde{\tau}^\delta_n$ as a convex combination $\tilde{\alpha}^\delta_n t^\delta_n + [1 - \tilde{\alpha}^\delta_n]\hat{\tau}^\delta_n$ where the support of $\hat{\tau}^\delta_n$ is contained in $T_n(P)$. Then sequential rationality at the initial stage after rejecting equivalence classes in $\mathcal{E}^\circ_n(P)$ and at subsequent information sets have the following two implications. First, $\tilde{\alpha}^\delta_n$ is nonzero only if t^δ_n is at least as good a reply as each $s_n \in T_n(P)$ against b^δ_m. Second, if $\tilde{\alpha}^\delta_n < 1$ then a strategy $s_n \in T_n(P)$ belongs to the support of $\hat{\tau}^\delta_n$ only if it is at least as good a reply against b^δ_m as the other strategies in T^δ_n; and for each $h_n \notin H_n(P)$ enabled by s_n in Γ, the choice prescribed by s_n at h_n is optimal at $h^\delta_n(E_n(s_n))$ given the belief $\tilde{\mu}^\delta_n(\cdot|h_n)$. If an information set $h^\delta_m(E_m)$ of player m is enabled by $\tilde{\tau}^\delta_n$ but not by $\tilde{\sigma}^\delta_n$ then the beliefs $\tilde{\mu}^\delta_m(\cdot|h_m)$ are derived from $\tilde{\tau}^\delta_n$.

The sequence $\{\tilde{\tau}^\delta_{n,\varepsilon}\}$ induces a corresponding sequence of equivalent strategies in G that induces a sequence of conditional distributions over $T_n(P)$. Because $t^\delta_n = [1 - \delta]s^*_n + \delta\rho_n$, the limit of the corresponding sequence of equivalent strategies in G is $[1 - \delta]\tilde{\alpha}^\delta_n s^*_n + \delta\tilde{\alpha}^\delta_n\rho + [1 - \tilde{\alpha}^\delta_n]\hat{\tau}^\delta_n$. Therefore, the limit of the sequence of conditional distributions over $T_n(P)$ is $\tau^\delta_n = \alpha^\delta_n\rho + [1 - \alpha^\delta_n]\hat{\tau}^\delta_n$, where $\alpha^\delta_n = \tilde{\alpha}^\delta_n\delta/[\tilde{\alpha}^\delta_n\delta + (1 - \tilde{\alpha}^\delta_n)]$. Obviously if an information set h_m of player m is enabled by τ^δ_n but not by σ^δ_n then the beliefs $\mu^\delta_m(\cdot|h_m)$ are those derived from τ^δ_n.

Passing to a subsequence if necessary, the limit τ_n of the sequence τ^δ_n can be expressed as a convex combination $\tau_n = \alpha_n\rho + [1 - \alpha_n]\hat{\tau}_n$ where α_n and $\hat{\tau}_n$ are the limits of α^δ_n and $\hat{\tau}^\delta_n$, respectively. As in the previous paragraph, if an information set h_m of player m is enabled by τ_n but not by σ_n then the beliefs $\mu_m(\cdot|h_m)$ are those derived from τ_n.

Claim 6.2. *(1) $\alpha_n > 0$. (2) If $\alpha_n < 1$ then the support of $\hat{\tau}_n$ consists of strategies in R_n. In particular, for each s_n in its support and each information set h_n that s_n enables, the choice of s_n at h_n is optimal given (b, μ).*

Proof of Claim. We prove (2) first. Suppose $\alpha_n < 1$. Let s_n be a strategy in $T_n(P)$ that is not optimal in reply to (b, μ). We show that s_n is not in the support of $\hat{\tau}_n^\delta$ for all sufficiently small δ, which proves the second statement. Let h_n be an information set that s_n enables where its action is not optimal. If $h_n \in H_n(P)$ then every strategy in $E_n(s_n)$ is suboptimal against b_m. Because s_n^*, the strategy that belongs to $S_n(P)$ and to the support of t_n^δ for all δ, is optimal against b_m^δ for all δ, and b is the limit of b^δ, for sufficiently small δ the strategy t_n^δ does better against b^δ than every strategy in the equivalence class $E_n(s_n)$. At the second information set in the initial stage of Γ^δ where n decides among the redundant strategy t_n^δ and equivalence classes not in $\mathcal{E}_n^\circ(P)$, sequential rationality implies that he chooses the equivalence class $E_n(s_n)$ with zero probability for all small δ. As we remarked above, this implies that for such δ, the probability of s_n is zero in $\hat{\tau}_n^\delta$.

If $h_n \notin H_n(P)$ then there exists another strategy s_n' in the equivalence class $E_n(s_n)$ that agrees with s_n elsewhere but prescribes an optimal continuation at h_n. Obviously, for all small δ, s_n' is a better reply than s_n in reply to (b^δ, μ^δ). But then sequential rationality at the copy $h_n^\delta(E_n(s_n))$ of h_n in Γ^δ for such small δ implies that he would choose according to s_n' there and not s_n. Again, the probability of s_n under $\hat{\tau}_n^\delta$ is zero for small δ. Thus every strategy in the support of $\hat{\tau}_n$ is optimal in reply to (b, μ) and therefore is a relevant strategy.

It remains to show that $\alpha_n \neq 0$. Suppose to the contrary that $\alpha_n = 0$. Let $\hat{S}_n$ be the set of strategies in the support of either σ_n or $\hat{\tau}_n$. Let $\hat{H}_m$ be the collection of information sets in H_m enabled by strategies in $\hat{S}_n$. Because (b, μ) is a weakly sequential equilibrium we obtain the following properties for each information set h_m in $\hat{H}_m$ of player m: if h_m is enabled by σ_n then the action prescribed by b_m at h_m is optimal against σ_n; if h_m is enabled only by $\hat{\tau}_n$ then the action prescribed by b_m is optimal against $\hat{\tau}_n$. Therefore, for each small $\eta > 0$ there exists a perturbation $\Gamma(\eta)$ of Γ, where only m's payoffs are perturbed, such that: σ_m is optimal against $\sigma_n(\eta) \equiv [1-\eta]\sigma_n + \eta\hat{\tau}_n$ in $\Gamma(\eta)$; and $\Gamma(\eta)$ converges to Γ as η goes to zero. As we argued above, $\hat{\tau}_n$ is optimal against σ_m in Γ. Therefore, for all small η, $(\sigma_m, \sigma_n(\eta))$ is an equilibrium of $\Gamma(\eta)$. Since Γ is generic, it belongs to some component C of the open set $\mathcal{G} \setminus \mathcal{G}_0$, where $\mathcal{G}_0$ is the set constructed in Appendix A. Therefore, the sequence $\Gamma(\eta)$ is in C for all small η. By Lemma A.1 in Appendix A, there exists a strategy τ_n' such that: (i) the support of τ_n' equals $\hat{S}_n$; and (ii) σ_m is a best reply against τ_n' in Γ. Therefore, for all $0 \leqslant \varepsilon \leqslant 1$, $(\sigma_m, (1-\varepsilon)\sigma_n + \varepsilon\tau_n')$ is an equilibrium of Γ. But because the strategies in the support of $\hat{\tau}_n$ choose a non-equilibrium action at some $h_n \in H_n(P)$ that they enable, all these equilibria result in different outcomes. This is impossible because the payoffs in Γ are generic and therefore Γ has only finitely many equilibrium outcomes [19]. Thus $\alpha_n \neq 0$. □

Now we prove that P satisfies forward induction by showing that (b, μ) induces beliefs that assign positive probability only to relevant strategies. Let h_n be an information set of n that is enabled by b_n. If $h_n \in H_n(P)$ then obviously n's belief over the continuation strategies of m is the one derived from σ_m, and strategies in the support of σ_m are obviously relevant. If $h_n \notin H_n(P)$ then the only strategies of m that enable h_n are those in $T_m(P)$. If there is no strategy in R_m that enables h_n then there is nothing to prove. Otherwise, the subset of strategies in R_m that enable h_n is not empty and then the strategy τ_m, which by the above Claim has R_m as its support, enables h_n. Therefore, $\mu_n(\cdot|h_n)$ is derived from τ_m, and in this case too, the restriction on beliefs imposed by forward induction holds. Thus P satisfies forward induction. □

Theorem 6.1 resolves a conjecture by Hillas and Kohlberg [27, §13.6]. Its remarkable aspect is that backward induction and invariance suffice for forward induction—if there are two players and payoffs are generic. No further assumption about rationality of behavior or plausibility of beliefs is invoked, nor are perturbations of strategies invoked as in studies of perfect equilibria and stable sets of equilibria. And, for signaling games there is no reliance on Cho and Kreps' [10, p. 181] auxiliary scenario in which the sender makes a 'speech' that the other's intransigent belief ignores the fact that a deviation would be rational provided merely that the receiver recognizes and acts on its implications by excluding irrelevant strategies from the support of his belief.

Invariance excludes one particular presentation effect by requiring that the outcome should not depend on whether a mixed strategy is treated as an additional pure strategy. One interpretation of forward induction is that it excludes another presentation effect by requiring that the outcome does not depend on irrelevant strategies. Indeed, van Damme [16, p. 1555] interprets forward induction as akin to the axiom called 'independence of irrelevant alternatives' in social choice theory. In the case of a game the analog of social choice is the outcome (the probability distribution on terminal nodes) and the irrelevant alternatives are players' irrelevant strategies.

7. Reny's Interpretation of Forward Induction

An implication of Theorem 6.1 is that there is no conflict between backward and forward induction if one adopts the decision-theoretic principle of invariance. This conclusion depends on our definitions of relevant strategies and forward induction outcomes; e.g. we interpret forward induction as a refinement of weakly sequential equilibrium that ensures the outcome does not depend on one player believing the other is using an irrelevant strategy at a relevant information set.

In this section we compare our definitions with the principle alternative, represented by the discussion in Reny [41, §4]. He invokes 'best response motivated inferences' as an instance of 'forward induction logic' and concludes from an example that it can conflict with backward induction.

Although he does not propose an explicit definition, the main ingredients differ from our formulation as follows. Our definitions are narrow—we interpret forward induction as a property of an outcome of a weakly sequential equilibrium, and ask only that the outcome results from one in which the support of a player's belief at a relevant information set is confined to relevant strategies, which we limit to those strategies that are optimal replies to some weakly sequential equilibrium. Reny's view applies forward induction reasoning directly to players' strategies rather than outcomes, and applies it to more information sets and more strategies. At every information set not excluded by a player's own strategy, he asks only that the support of the player's belief is confined to those strategies reaching that information set for which there are some beliefs of the other player that would justify using them.

The implications of Reny's expanded view of forward induction reasoning are illustrated by his motivating example [41, Figure 3]. The top panel of Figure 4 shows a game in which players I and II alternately choose whether to end the game. Reny argues that this example shows a tension between forward and backward induction. He observes that I's choice of the pure strategy D strictly dominates Ad. He infers from this that forward induction should require that if I rejects D then II must believe that I's strategy is surely Aa, and hence II's only optimal reply is Ad. But backward induction requires each player to choose d, and before that D, which contradicts the seeming implication of forward induction that II's strategy should be Ad. From this he concludes that II's backward induction strategy is "rendered 'irrational' " and thus "the inappropriateness, indeed the *inapplicability* of the usual backward programming argument in the presence of best response motivated inferences." [41, p. 637, italics in original].

Our analysis of this example differs in two respects. First, I's only relevant strategy is to choose D initially, so forward induction according to our definition has no implications for II's beliefs. This is so because our definitions identify outcomes resulting from the conjunction of rational play and beliefs that other players are playing rationally; hence we apply them only to information sets reached by rational play as represented by relevant strategies. In contrast, Reny applies forward induction to the belief of a player when the other player's strategy is an optimal reply to an arbitrary belief. In the top panel of Figure 4, for player II to choose A requires that either II believes I is irrational, or II believes that I believes II

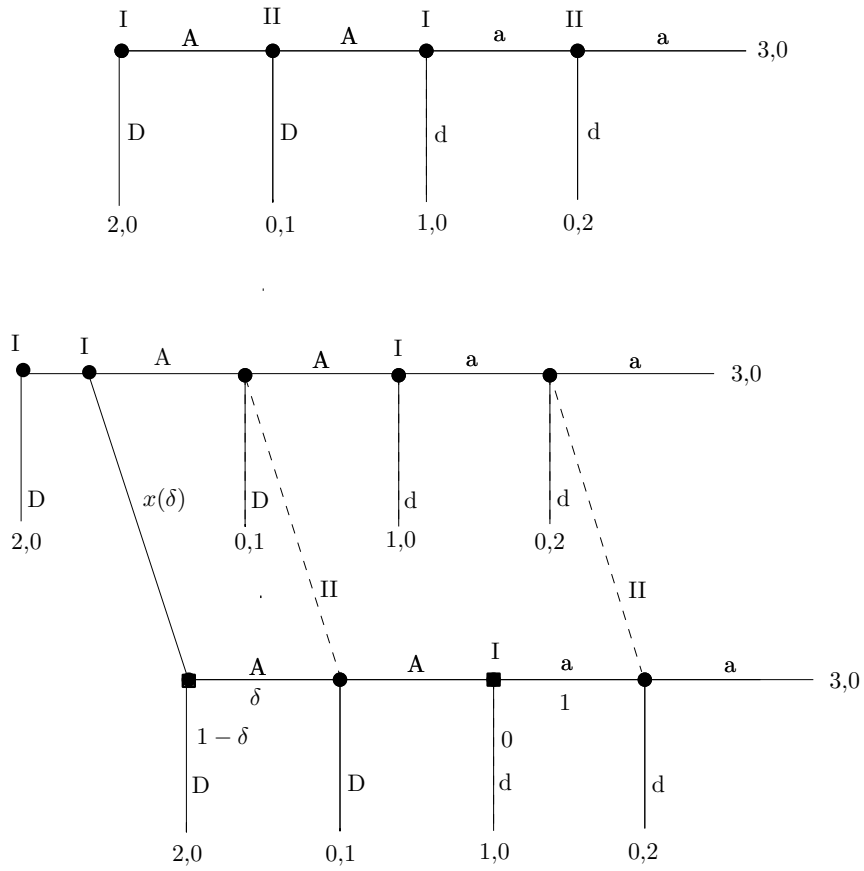

FIGURE 4. Top panel: Reny's example of a game between players I and II. Bottom panel: The game modified so that player I can choose the redundant strategy $x(\delta)$ after rejecting D.

is irrational. Specifically, for II at her first decision node to believe that I's strategy is Ad amounts to believing that I is irrational (because D dominates Ad as noted above); and for II to believe that I's strategy is Aa and that I is rational requires II to ascribe to I a belief that II's strategy is Aa with high probability, which is an irrational strategy for II (because at II's second decision node the continuation d dominates the continuation a).

Our view is that a coherent theory of rational play and beliefs that others are playing rationally (i.e. a theory consistent with strong belief in rationality) is possible only with the more circumscribed definition of relevant strategies that we propose. Note, however, that we admit fewer strategies as relevant but restrict beliefs only at relevant information sets, so our definition of forward induction is neither stronger nor weaker than Reny's interpretation.

The other respect in which our analysis differs is that we invoke invariance. The bottom panel of Figure 4 shows an expanded extensive form in which player I can reject D and then choose between action A or the new pure strategy $x(\delta)$ for some probability $\delta \in (1/2, 1)$.

(See [20, §2.3] for a similar example.) The two information sets indicate that player II cannot know whether I chose A or $x(\delta)$ after rejecting D. The branch points indicated by black boxes refer to moves by nature, i.e. nature takes over and implements the strategy $x(\delta)$ using the indicated probabilities $(1-\delta, \delta)$ and $(0, 1)$ at I's first and second information sets after I chooses $x(\delta)$. Note that $x(\delta)$ is redundant because it is replicated by the mixed strategy that chooses between D and Aa with probabilities $1-\delta$ and δ.

In the expanded game it is somewhat arbitrary whether one supposes that I can choose $x(\delta)$ before or after D. We use the latter because then it is easy to construct the unique Nash equilibrium of the subgame that begins after I rejects D. This is a sequential equilibrium in which I chooses $x(\delta)$ with probability $1/[1+\delta]$ and otherwise chooses A and then d; and, if A occurs then II chooses D with probability $2\delta - 1$ and otherwise A and then d. Consistent with Bayes' Rule, II's strategy is supported by beliefs at her first and second information sets that the conditional probabilities are respectively $1/2$ and 1 that I chose $x(\delta)$. As δ approaches $1/2$, II's strategy in this equilibrium converges to Ad, and as δ approaches 1, to D, which correspond to the two strategies by II that Reny considers. (When $\delta = 1/2$ the Nash equilibria of the subgame require only that I's probability of $x(1/2)$ is at least $2/3$, and II's belief changes accordingly; and when $\delta = 1$ the game is essentially the same as the original game since $x(1)$ is a duplicate of Aa.)

Thus we find that II might be indifferent between her backward induction strategy D and her strategy Ad that Reny concludes is implied by best response motivated inferences. Therefore, Reny's conclusion that II's backward induction strategy is rendered irrational depends on rejecting invariance as a decision-theoretic principle.

Because Theorem 6.1 is restricted to games with two players and generic payoffs, it does not establish that our definitions of relevant strategies and forward induction outcomes are surely the right ones for general games. But it suggests that similar definitions can enable 'forward induction reasoning' to be justified by decision-theoretic criteria.

8. Economic Applications

We conclude by mentioning some prominent applications of forward induction, although in every case the authors use minimal assumptions or equilibrium selection criteria that are weaker than forward induction as we define it in Section 3.

One of the main applications of forward induction in economics is to game-theoretic models of entry deterrence and contestability in dynamic oligopolies with high fixed costs; e.g. Ponssard [40]. These models resemble Example 2.1 in that there is one equilibrium in which a low-cost firm is deterred from entering a market because it expects a high-cost incumbent

to meet entry with competition, but in another equilibrium (the one whose outcome satisfies forward induction) the high-cost firm exits after the low-cost firm enters. Ben-Porath and Dekel [5] study cases in which merely the option to engage in costly actions prior to a subgame selects an equilibrium of the subgame if one invokes iterative deletion of weakly dominated strategies. Hauk and Hurkens [25] study a version in which the forward induction outcome results from an evolutionary model.

Other applications resemble Example 2.2. These include explicit signaling games such as limit pricing as in Bagwell [2] and Milgrom and Roberts [37], and signaling of quality in labor markets by acquiring educational credentials as in Spence [42] and in product markets by advertising or warranties as in Bagwell and Ramey [3] and Milgrom and Roberts [38]. In many of these models, with sufficient regularity assumptions and a sufficiently rich set of signaling strategies for the sender, the outcomes of pooling and partially pooling equilibria do not satisfy forward induction, and thus forward induction selects the outcome of a separating equilibrium in which signals reveal players' types. This is to be expected since forward induction ensures maximum opportunity for credible signaling of private information by rejecting outcomes that depend on others' implausible or intransigent beliefs.

Among the important extensions of signaling games are dynamic models of bargaining by Admati and Perry [1], Compte and Jehiel [12], Cramton [13], Cramton and Tracy [14], Feinberg and Skrzypacz [17], and Grossman and Perry [22, 23] among others. In several of these the forward induction outcome results from a separating equilibrium in which the informed party credibly signals his type via sufficient delay in responding with a serious counteroffer. This outcome differs from outcomes of equilibria with partial pooling in which the uninformed party screens sets of types of the informed party via successively better offers. Gül and Sonnenschein [24] assume three conditions (most importantly that the informed party's strategy is stationary) that imply restriction to screening equilibria, and from this they obtain a proof of the Coase conjecture. But this result is not implied by equilibria of the model by Admati and Perry and the model by Cramton in which both parties have private information, and not necessarily by the model of Feinberg and Skrzypacz in which the uninformed party's probability distribution over the other's type is itself private information.

We caution that Theorem 6.1's restrictions to games with two players and generic payoffs have analogs in applications of forward induction. In Example 2.2 both outcomes QQ-R and BB-R satisfy forward induction if the two types of player I are treated as distinct players in a three-player game. Chen, Kartik, and Sobel [9, p. 118] emphasize that criteria like forward induction have limited power to select among outcomes of 'cheap talk' signaling games in which the sender's payoffs do not depend on his action and hence are nongeneric. However,

they show for a class of games with one-dimensional types that the 'most informative' equilibrium is a limit of equilibria of perturbed games in which the sender's payoffs depend on his action, provided that in these equilibria the sender's action is weakly increasing in his type and the receiver's action is weakly increasingly in the sender's action.

8.1. **Final Remarks.** We do not offer a new realm of applications. We do offer an explanation of why forward induction is a desirable refinement of sequential equilibrium in two-player games with generic payoffs. Theorem 6.1 says that if an outcome does not satisfy forward induction—that is, depends on one player believing the other is using an irrelevant strategy—then there is an equivalent game in which this outcome results only from Nash equilibria and not from any sequential equilibrium. Failure of an economic model to predict an outcome that satisfies forward induction could motivate reconsideration of whether the essential features of the strategic situation are well represented by the specific extensive form used in the model, or if one has confidence in the model then this prediction might be rejected because for the same model there necessarily exists another prediction that does satisfy forward induction.

References

[1] Admati, Anat, and Motty Perry (1987), "Strategic Delay in Bargaining," *Review of Economic Studies*, 54: 345–364.

[2] Bagwell, Kyle (1987), "Introductory Price as a Signal of Cost in a Model of Repeat Business," *Review of Economic Studies*, 54: 365–384.

[3] Bagwell, Kyle, and Garey Ramey (1988), "Advertising and Limit Pricing," *The RAND Journal of Economics*, 19: 59–71.

[4] Banks, Jeffrey, and Joel Sobel (1987), "Equilibrium Selection in Signaling Games," *Econometrica*, 55: 647–661.

[5] Ben-Porath, Elchanan, and Eddie Dekel (1992), "Signaling Future Actions and the Potential for Sacrifice," *Journal of Economic Theory*, 57: 36–51.

[6] Battigalli, Pierpaolo, and Marciano Siniscalchi (2002), "Strong Belief and Forward Induction Reasoning," *Journal of Economic Theory*, 106: 356-391.

[7] Blume, Lawrence, Adam Brandenburger, and Eddie Dekel (1991), "Lexicographic Probabilities and Equilibrium Refinements," *Econometrica*, 59: 81–98.

[8] Bochnak, J., M. Coste, and M-F. Roy (1998), *Real Algebraic Geometry*. Berlin: Springer-Verlag.

[9] Chen, Ying, Navin Kartik, and Joel Sobel (2008), "Selecting Cheap-Talk Equilibria," *Econometrica*, 76: 117-136.

[10] Cho, In-Koo, and David M. Kreps (1987), "Signalling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102: 179–221.

[11] Cho, In-Koo, and Joel Sobel (1990), "Strategic Stability and Uniqueness in Signaling Games," *Journal of Economic Theory*, 50: 381–413.

[12] Compte, Olivier, and Philippe Jehiel (2002), "On the Role of Outside Options in Bargaining with Obstinate Parties," *Econometrica*,70: 1477–1517.

[13] Cramton, Peter (1984), "Bargaining with Incomplete Information: An Infinite-Horizon Model with Two-Sided Uncertainty," *The Review of Economic Studies*, 51: 579–593.

[14] Cramton, Peter, and Joseph Tracy (1994), "Wage Bargaining with Time-Varying Threats," *Journal of Labor Economics*, 12: 594–617.

[15] van Damme, Eric (1989), "Stable Equilibria and Forward Induction," *Journal of Economic Theory*, 48: 476–496.

[16] van Damme, Eric (2002), "Strategic Equilibrium," in: R. Aumann and S. Hart (eds.), *Handbook of Game Theory*, Vol. 3, Chapter 41, 1523–1596. New York: Elsevier.

[17] Feinberg, Yossi, and Andrzej Skrzypacz (2005), "Uncertainty about Uncertainty and Delay in Bargaining," *Econometrica*, 73: 69-91.

[18] Fudenberg, Drew, and Jean Tirole (1993), *Game Theory*. Cambridge MA: MIT Press.

[19] Govindan, Srihari, and Robert Wilson (2001), "Direct Proofs of Generic Finiteness of Nash Equilibrium Outcomes," *Econometrica*, 69: 765–769.

[20] Govindan, Srihari, and Robert Wilson (2006), "Sufficient Conditions for Stable Equilibria," *Theoretical Economics*, 1: 167–206.

[21] Govindan, Srihari, and Robert Wilson (2007), "On Forward Induction," Research Report 1955, Stanford Business School, January.

[22] Grossman, Sanford, and Motty Perry (1986a), "Perfect Sequential Equilibrium," *Journal of Economic Theory*, 39: 97–119.

[23] Grossman, Sanford, and Motty Perry (1986b), "Sequential Bargaining under Asymmetric Information," *Journal of Economic Theory*, 39: 120–154.

[24] Gül, Faruk, and Hugo Sonnenschein, "On Delay in Bargaining with One-Sided Uncertainty," *Econometrica*, 56: 601–611 .

[25] Hauk, Esther, and Sjaak Hurkens (2002), "On Forward Induction and Evolutionary and Strategic Stability," *Journal of Economic Theory*, 106: 66–90.

[26] Hillas, John (1994), "How Much of Forward Induction is Implied by Backward Induction and Ordinality," University of Auckland, New Zealand.

[27] Hillas, John, and Elon Kohlberg (2002), "Conceptual Foundations of Strategic Equilibrium," in: R. Aumann and S. Hart (eds.), *Handbook of Game Theory*, Vol. 3, Chapter 42, 1597-1663. New York: Elsevier.

[28] Kohlberg, Elon (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.

[29] Kohlberg, Elon, and Jean-François Mertens (1986), "On the Strategic Stability of Equilibria," *Econometrica*, 54: 1003–1037.

[30] Kreps, David (1990), *A Course in Microeconomic Theory*. Princeton NJ: Princeton University Press.

[31] Kreps, David, and Garey Ramey (1987), "Structural Consistency, Consistency, and Sequential Rationality," *Econometrica*, 55: 1331–1348.

[32] Kreps, David M., and Joel Sobel (1994), "Signalling," in: R. Aumann and S. Hart, (eds.), *Handbook of Game Theory*, Vol. 2, Chapter 25, pp. 849–868. New York: Elsevier.

[33] Kreps, David, and Robert Wilson (1982), "Sequential Equilibria," *Econometrica*, 50: 863–894.

[34] Kuhn, Harold (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.

[35] McLennan, Andrew (1985), "Justifiable Beliefs in Sequential Equilibrium," *Econometrica*, 53: 889-904.
[36] Mertens, Jean-François (1989), "Stable Equilibria—A Reformulation, Part I: Definition and Basic Properties," *Mathematics of Operations Research*, 14: 575–625.
[37] Milgrom, Paul, and John Roberts (1982), "Limit Pricing and Entry under Incomplete Information: A General Equilibrium Analysis.," *Econometrica*, 50: 443–459.
[38] Milgrom, Paul, and John Roberts (1986), "Price and Advertising Signals of Product Quality," *Journal of Political Economy*, 94: 796–782.
[39] Pearce, David (1984), "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52: 1029-1050.
[40] Ponssard, Jean-Pierre (1991), "Forward Induction and Sunk Costs Give Average Cost Pricing," *Games Economic Behavior*, 3: 221–236.
[41] Reny, Philip (1992), "Backward Induction, Normal Form Perfection and Explicable Equilibria," *Econometrica*, 60: 627–649.
[42] Spence, A. Michael (1974), *Market Signaling*. Cambridge, MA: Harvard University Press.

Appendix A. Technical Lemma

Given an extensive form with two players, for each player n let S_n and Σ_n be n's sets of pure and mixed strategies, and let $S = S_1 \times S_2$ and $\Sigma = \Sigma_1 \times \Sigma_2$ be the product sets of profiles. Let $\mathcal{G}$ be the Euclidean space of extensive-form games generated by assigning payoffs to the players at the terminal nodes of the given extensive form.

Lemma A.1. *There exists a closed, lower-dimensional, semi-algebraic set $\mathcal{G}_1$ of $\mathcal{G}$ such that for each connected component C of $\mathcal{G} \setminus \mathcal{G}_1$, the following holds: if for some game $\Gamma \in C$ and profile $\sigma \in \Sigma$ the set of profiles of pure strategies that are the players' optimal replies to σ is $T = T_1 \times T_2 \subset S$, then for every game $\Gamma' \in C$ there exists a profile $\sigma' \in \Sigma$ with the same support as σ and such that in Γ' the set of pure optimal replies to σ' is T.*

Proof. Let $X = \mathcal{G} \times \Sigma$ and let $p : X \to \mathcal{G}$ be the natural projection. For each pair $R = R_1 \times R_2$ and $T = T_1 \times T_2$ of subsets of S, let $X(R,T)$ be the set of (Γ, σ) in X such that, for each n, R_n is the support of σ_n and T_n is the set of n's pure optimal replies in Γ to the mixed strategy σ_m of the other player. By the generic local triviality theorem [8] there exists a closed, lower-dimensional, semi-algebraic subset $\mathcal{G}_1$ of $\mathcal{G}$ such that for each connected component C of $\mathcal{G} \setminus \mathcal{G}_1$ there exist: (i) a semi-algebraic fibre F, (ii) for each pair (R,T) a subset $F(R,T)$ of F, and (iii) a homeomorphism $h : C \times F \to p^{-1}(C)$ with the properties that (a) $p \circ h(\Gamma, f) = \Gamma$ for all $\Gamma \in C$, and (b) h maps $C \times F(R,T)$ homeomorphically onto $p^{-1}(C) \cap X(R,T)$ for each (R,T).

Suppose T is the set of profiles of pure optimal replies to σ in a game $\Gamma \in C$. Let R be the support of σ. Then (Γ, σ) belongs to $X(R,T)$. Therefore there exists $f \in F(R,T)$ such that $h(\Gamma, f) = (\Gamma, \sigma)$. For each $\Gamma' \in C$, let $\sigma'(f)$ be the unique mixed strategy in Σ for which

$h(\Gamma', f) = (\Gamma', \sigma'(f))$. Then the support of $\sigma'(f)$ is R and the set of profiles of pure optimal replies in Γ' to $\sigma'(f)$ is T. □

Appendix B. Forward Induction in the Normal Form

The classical view in game theory is that the normal form of a game is sufficient to capture all strategically significant aspects. Hence the question arises as to whether we can state a comparable version of forward induction for a game in normal form. Here we provide one such definition.

The following three components of Definition 3.4 for a game in extensive form need to be rephrased in terms of the normal form: (1) weakly sequential equilibria, (2) relevant strategies, and (3) restriction of beliefs to those induced by relevant strategies whenever possible. As will be seen below, if the sequential rationality requirement in the definition of weakly sequential equilibria is strengthened slightly (and only for nongeneric games), then the corresponding definition of forward induction has a normal-form counterpart.

Given a game G in normal form, let σ be a profile of players' mixed strategies and let b be an equivalent profile in behavioral strategies for an extensive-form game Γ with that normal form. Reny [41, Prop. 1] shows that σ is a normal-form perfect equilibrium of G iff in Γ there exists a sequence b^ε of completely mixed profiles converging to b such that for each player n and each information set h_n that b_n does not exclude, the action prescribed by b_n at h_n is optimal against b^ε_{-n} for all small ε. Thus the difference between weakly sequential equilibrium and normal-form perfect equilibrium is analogous to that between sequential equilibrium and extensive-form perfect equilibrium: one requires optimality only in reply to the limit, while the other requires optimality in reply to the sequence as well. Reny also shows that weakly sequential equilibria coincide with normal-form perfect equilibria for generic extensive-form games. Therefore, a perfect equilibrium seems to be the right normal-form analog of a weakly sequential equilibrium.

Suppose Σ^* is a set of Nash equilibria of G. (To fix ideas, Σ^* could be the set $\Sigma(P)$ of equilibria inducing an outcome P in an extensive-form version of the game, but to allow applications to nongeneric games we want to allow multiple outcomes.) In the extensive-form case, we said that a strategy was relevant if it was optimal against a strategy-belief pair inducing the given outcome. But as noted above, if we insist on optimality along the sequence then the appropriate normal-form definition of a relevant strategy becomes: a strategy is relevant if it is optimal against a sequence of ε-perfect equilibria converging to an equilibrium in Σ^*.

Finally, we turn to belief restrictions. The idea in the extensive-form case is that if an information set h_n of player n is reached by a profile of relevant strategies of his opponents then he assigns zero probability to continuations that are enabled only by profiles that contain an irrelevant strategy for one of the other players. Let $R_{-n}(h_n)$ be the set of profiles of relevant strategies of n's opponents that reach such an h_n. If we use a sequence σ^ε normal-form profiles to generate players' beliefs and their continuation strategies, then the belief restriction says that n's belief at h_n and the continuation strategies of his opponents should be obtained from the limit of the sequence of conditional distributions over $R_{-n}(h_n)$ induced by the sequence σ^ε. That is, the beliefs at all information sets of all players that are reached by relevant strategies can be generated from the sequence of conditional distributions confined to relevant strategies.

Because we insist on optimality along the sequence, what we obtain is a perfect equilibrium with a restriction on the form of its representation as a lexicographic probability system, as in Blume, Brandenberger, and Dekel [7, Prop. 4,7]. The restriction is that any profile that includes an irrelevant strategy for some player should occur later in the lexicographic sequence than those that include only relevant strategies. This implements the basic requirement that each player believes the other is using a relevant strategy so long as that hypothesis is tenable. Thus, we are led to the following definition:

Definition B.1 (Normal-Form Forward Induction)**.** A set of Nash equilibria satisfies normal-form forward induction if it contains a perfect equilibrium whose lexicographic representation has all profiles of relevant strategies occurring before all profiles that include irrelevant strategies.

In general this is a stronger requirement than the one in the text. But for a generic two-player extensive-form game with perfect recall it can be shown that the set of weakly sequential equilibria inducing an outcome P satisfies the above definition iff P satisfies forward induction as defined in the text. The reason for this equivalence is similar to the reason that weakly sequential equilibria and normal-form perfect equilibria coincide for generic extensive-form games as established by Reny [41, Prop. 1]. An implication is that the analog of Theorem 6.1 is true with this definition of forward induction, i.e. the set of Nash equilibria resulting in an invariant sequential equilibrium outcome of a two-player game with perfect recall and generic payoffs satisfies normal-form forward induction.

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