FORWARD INDUCTION EQUILIBRIUM

PRISCILLA T. Y. MAN

Abstract. Forward induction is the notion that players in a game assume, even when confronted with an unexpected event, that their opponents chose rationally in the past and will choose rationally in the future. This paper modifies Govindan and Wilson's (2009, Econometrica 77(1), 1-28) definition of forward induction and constructs an admissible, invariant forward induction equilibrium concept for general games using normal form perfect equilibrium. Forward induction equilibrium according to this new definition exists for all finite, generic extensive form games with perfect recall. It does not satisfy backward induction. Yet for generic extensive form games the set of forward induction outcomes contains an invariant sequential equilibrium outcome. Forward induction is not equivalent to iterative elimination of strategies dominated at the equilibrium value. In signaling games, a forward induction equilibrium survives most existing equilibrium refinements.

1. Introduction

Forward induction is the idea that players in a game tend to make inferences from other players' behaviors even when confronted with the unexpected. While it has been a long-standing notion in game theory, only recently have Govindan and Wilson (2009b) provided a formal definition of forward induction for general extensive form games with perfect recall. This paper modifies their definition and explores the properties of the resulting equilibrium concept. There are two purposes for this exercise. First, to formalize forward induction for general games and understand its implications. We improve upon Govindan and Wilson

Date: November 11, 2009.
Key words and phrases. Forward induction, Equilibrium refinement, Backward induction, Iterative dominance.
I thank Phil Reny and Hugo Sonnenschein for their advice and encouragement. I have also received helpful comments from Andy McLennan and Roger Myerson. Funding from Henry Morgenthau Jr. Memorial Fund Dissertation Fellowship is gratefully acknowledged.

(2009b) definition to obtain an admissible, invariant forward equilibrium concept. This provides a formal expression of forward induction as an equilibrium concept for further analysis. The second purpose is to understand the consequences if we treat forward induction as a basic decision theoretic criterion for equilibrium concepts. Game theorists sometimes consider forward induction as a secondary requirement to be imposed only when its counterpart, backward induction, fails to have cutting power (for example, see van Damme (1989); Pearce (1984) could be viewed as an exception). By focusing solely on forward induction, we wish to explore the forces and limitations of forward induction as a basic requirement for solution concepts. This is not to say that forward induction should always prevail. Rather, we feel an understanding of forward induction as comprehensive as that of backward induction should provide good theoretical reasons for favoring one or another.

Central to forward induction is that players maintain the assumption that their opponents have maximized their expected utility in the past as long as this assumption is tenable — even if they observe the unexpected. In other words, if a player finds himself off the equilibrium path, he should not interpret it as a result of unintentional mistakes by his opponents as long as his opponents' deviations are "rationalizable". As a classic example, consider the Outside Option Game in Figure 1.1 in which Player 1 could choose between an outside option and playing a "battle of the sexes" game with Player 2. Player 1 prefers her favorite Nash equilibrium outcome (that is, (T, L)) in the battle of the sexes game to the outside option. Yet she prefers the outside option to the other two Nash equilibrium outcomes ((B, R) and the mixed strategy equilibrium) in the subgame. Typical forward induction argument rules out the outcome $(2, 2)$ in this game. To sustain $(2, 2)$ as an equilibrium outcome, Player 2 must put positive probability on R. However, Player 2 could reason: as long as Player 1 followed her equilibrium strategy, she would have got a payoff of 2 and Player 2 would not get to move. If Player 1 is not making a mistake letting Player 2 move, she must have chosen T, because B always gives her strictly less than 2. Yet if Player 2 believes with probability 1 that Player 1 had chosen T, his best response is L. This upsets the original equilibrium.

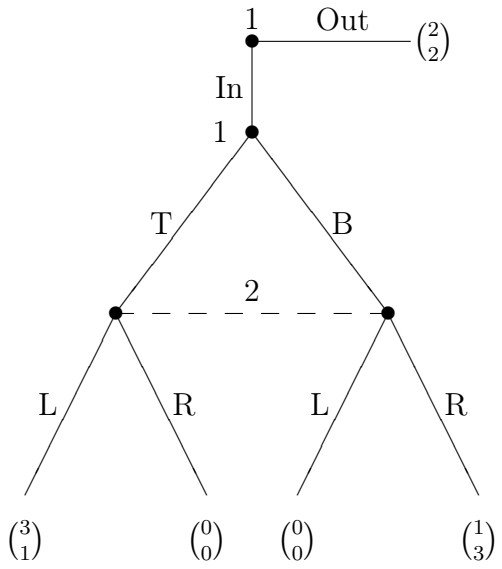


FIGURE 1.1. Outside Option Game

Simple as it is, this example reveals a crucial consequence of forward induction: that a subgame cannot be treated as a game on its own. In Figure 1.1, (B, R) is a strict Nash equilibrium in the subgame following "In". Most refinement concepts, including those motivated by forward induction, will not rule out a strict Nash equilibrium (neither will ours). Yet in the game containing this subgame, (B, R) is not part of a forward induction equilibrium. In other words, a forward induction solution of a subgame need not be part of the solution of the whole game. This violates one of the basic properties of backward induction.

Indeed, forward induction requires players to attach meanings to deviations. How a subgame is reached in a larger game conveys information about intended play in the subgame. This piece of information would be absent had the subgame been a game in its own right. Requiring forward inducting players to treat a subgame as a stand-alone game is requiring them to discard this piece of information when formulating the "meaning" of a deviation. We do not see any particular philosophical or decision theoretic reasons for requiring them to do so. However, if a subgame is not treated as a game in its own right, we cannot expect a forward induction solution to project into a subgame the same solution, nor could we require a forward induction solution of a subgame to be part of the solution of the whole game.

In short, it is not obvious whether forward induction reasoning is always compatible with backward induction.

An immediate consequence of this incompatibility is that a formal definition of forward induction cannot be built upon a backward induction solution concept such as sequential equilibrium. Govindan and Wilson (2009b) build their definition on weakly sequential equilibrium (Reny, 1992), which does not require backward induction. A weakly sequential equilibrium of an extensive form game is a consistent assessment such that each player's behavioral strategy constitutes an optimal continuation given his beliefs at those information sets not excluded by this strategy. An information set of a player is excluded by a behavioral strategy of his if this strategy reaches the information set with zero probability regardless of his opponents' strategies. We follow the interpretation by Reny (1992, p. 631-632) that a weakly sequential equilibrium strategy of a player consists of two parts. One part describes his rational plan at information sets that could potentially be reached given this strategy; another describes other players' predictions about this player's future behaviors in the event he deviates from his rational plan. In light of this interpretation, forward induction is a restriction on the second part. It requires the opponents to believe whenever possible that the deviating player maximizes expected utility, knows the equilibrium outcome according to the rational plan, but might not know the off-equilibrium continuation.

Until now we have outlined the concept of forward induction in the language of extensive form games. Yet we wish to argue whether an outcome or an equilibrium satisfies forward induction should not depend on the particular way the game tree is drawn. That is, we propose that forward induction should be an invariant concept — it should depend only on the reduced normal form representation of the extensive form game. Consider Figure 1.2, the reduced normal form of the game tree in Figure 1.1. Player 2 could observe from the reduced normal form representation that his choice matters only when Player 1 chooses either T or B. If Player 2 considers what rational choices of Player 1 could give him an "non-trivial move", he should conclude that Player 1 is choosing T since B always gives strictly less than

	L	R
Out	2,2	2,2
T	3,1	0,0
B	0,0	1,3

FIGURE 1.2. Outside Option Game: Reduced Normal Form

2, the proposed equilibrium payoff. The best response to this belief is L, which upsets any equilibrium giving outcome $(2, 2)$[1].

Moreover, two finite extensive form games with perfect recall having the same reduced normal form can be transformed into one another through a sequence of strategically inessential transformations[2] (Thompson, 1952; Dalkey, 1953; Elmes and Reny, 1994). Thus invariance seems to be a basic property that should be respected by solution concepts motivated by decision theoretic principles. As we treat forward induction as a basic decision theoretic criterion, it is reasonable to combine the two criteria. Anyone contending that forward induction should not be invariant would need to point out which transformation(s) would affect the forward induction reasoning.

To construct an invariant concept of forward induction, we follow the definition outlined by Govindan and Wilson (2009b) in their Appendix B and build our definition on normal form perfect equilibrium. Reny (1992) shows that weakly sequential equilibrium is generically equivalent to normal form perfect equilibrium so this seems a natural extension. Using the representation of Blume, Brandenburger, and Dekel (1991), one could interpret a normal form perfect equilibrium as a Nash equilibrium under certain lexicographic belief system. The concept of forward induction is therefore a restriction on the set of possible lexicographic beliefs that players could hold. To be precise, take a normal form perfect equilibrium.

[1]Mailath, Samuelson, and Swinkels (1993, Section 9) make the same argument using normal form information sets.

[2]These transformation includes: inflation-deflation of information sets, addition of superfluous move, coalescing information sets, interchanging simultaneous moves, reducing simple lotteries to their expected values and adding a mixed strategy as a pure strategy. Dalkey (1953) suggests the first one. The first four transformations are considered by Thompson (1952). The last two transformations are added to the list by Kohlberg and Mertens (1986). Elmes and Reny (1994) modify the second one and delete the first one to preserve perfect recall.

Consider the set of all normal form perfect equilibria that are outcome equivalent to it. Call a pure strategy of a player "relevant" for this outcome if there exists a sequence of ε-perfect equilibria converging to a point in this set such that the pure strategy is a best response to this sequence. Otherwise the strategy is "irrelevant" for the outcome. Notice that a relevant strategy could be adopted by an optimizing player who knows the outcome but is uncertain about which equilibrium giving the outcome is "in effect". Forward induction posits that players believe relevant strategies to be infinitely more likely than irrelevant strategies. Thus we say an normal form equilibrium satisfies forward induction if it could be supported by a lexicographic belief system in which all strategy profiles putting positive probability only on relevant strategies occurs before all those putting positive probability on some irrelevant strategies.

As an illustration, consider again the Outside Option Game in Figure 1.2. All normal form perfect equilibria giving the outcome $(2, 2)$ are of the form $(\text{Out}, (\alpha, 1-\alpha))$ where $\alpha \in \left[0, \frac{2}{3}\right]$. Consider the sequence of ε-perfect equilibria $\left((1-4\varepsilon, 3\varepsilon, \varepsilon), \left(\frac{2}{3}, \frac{1}{3}\right)\right)$ converging to $\left(\text{Out}, \left(\frac{2}{3}, \frac{1}{3}\right)\right)$. Player 1's strategies Out and T are both best responses to this sequence; so are Player 2's L and R. On the other hand, B is not a best response to any sequence of ε-perfect equilibria as it is a strictly dominated strategy. Thus we say Out, T, L and R are relevant for the outcome $(2, 2)$ while B is irrelevant for the same outcome. However, there is no normal form perfect equilibrium with outcome $(2, 2)$ if T is infinitely more likely than B. Hence $(2, 2)$ is not a forward induction outcome and none of the normal form perfect equilibria giving this outcome is a forward induction equilibrium[3].

The definition by Govindan and Wilson (2009b) stops here but a consistency argument suggests that one should go further. Suppose we select a subset of normal form perfect equilibria (giving a certain outcome) that satisfies forward induction. We may find all strategies relevant for this subset. Call these strategies "second-order relevant" for the equilibrium

[3]This example also explains why we consider *all* normal form equilibria giving the same outcome rather than just a single equilibrium when we define relevant strategies. Consider the equilibrium (Out, R) in the Outside Option Game. If we define relevant strategies with respect to this single equilibrium, among Player 1's strategies only Out is relevant for this equilibrium. There is no way we could rule this equilibrium out. Yet it clearly violates the idea of forward induction!

outcome. We then require the opponents of a player to believe these "second-order relevant strategies" are infinitely more likely than the "first-order relevant strategies", which are infinitely more likely than the irrelevant strategies. This may strictly refine the set of equilibria selected (we give an example in Section 3.3). We can keep iterating our "forward induction algorithm" until no more strategy can be "pruned"[4]. The resulting set of forward induction equilibria can be considered as a "fixed point": Every equilibrium in the set can be supported by players believing that the deviators maximize expected utility but may be confused about which equilibria in the set is in effect. Having said that, the more rounds of iteration carried out, the more detailed an equilibrium plan players must be sharing. We do not want to assert that such a high level of sophisticated induction is always reasonable in all applications.

We then examine a few properties of forward induction equilibrium. We show that any strategically stable set (Kohlberg and Mertens, 1986) with a constant outcome contains a forward induction equilibrium. Since finite generic extensive form games with perfect recall have a stable outcome, forward induction equilibrium exists for all generic extensive form games[5]. We also explore the relationship between forward and backward induction. A forward induction equilibrium need not satisfy backward induction. Nevertheless, every generic extensive form game has a forward induction outcome which is an admissible invariant backward induction outcome. This result is closely related to that of Govindan and Wilson (2009b, Theorem 6.1) that any invariant backward induction outcome satisfies forward induction (using their definition).

Forward induction is often associated with the operation of iteratively eliminating strategies dominated at the equilibrium value (Kohlberg and Mertens, 1986, Proposition 6) in the literature. However, we show that our forward induction equilibrium does not survive iterative elimination of dominated strategies (weakly or strictly). Nor does it survive elimination of strategies dominated at the equilibrium value in *any* order. This discrepancy suggests

[4]In this paper, "pruning" a strategy means it is made arbitrarily less likely than the surviving ones. It does *not* mean eliminating the strategies. The two procedures are not equivalent. See Section 4.3 for details.
[5]By "generic" we mean the property that the tree possesses finitely many Nash equilibrium outcomes, see Kreps and Wilson (1982).

that forward induction is not equivalent to iteratively eliminating strategies dominated at the equilibrium value. There are two reasons for this: First, a strategy that is eliminated is different from a strategy receiving zero probability in equilibrium. When admissibility is required, the former does not receive any probability whereas the latter still receives arbitrarily small probability along the "trembling sequence". Thus conceptually we should not consider eliminating irrelevant strategies the same as making them very unlikely. Second, forward induction requires all irrelevant strategies to be infinitely less likely than all relevant strategies. Eliminating irrelevant strategies in an arbitrary order need not preserve equilibria satisfying this restriction. The concept of forward induction can at most justify deleting all irrelevant strategies for all players at the same time. Indeed, our forward induction equilibrium survives this elimination.

We do not want to claim that our definition is the only correct formalization for forward induction. After all, forward induction is not our innovation. Kohlberg and Mertens (1986, p. 1029) use the term "forward induction" to describe a property of the stable set. The concept has also been introduced in signaling games (Cho and Kreps, 1987; Banks and Sobel, 1987) and bargaining games (Rubinstein, 1985, Assumption B1). van Damme (1989) suggests a minimal criterion for a forward induction concept, upon which Al-Najjar (1995) builds a definition of forward induction for two-person multi-stage games. We compare our equilibrium concept with some of these precursors in Section 5. Our forward induction equilibrium implies some commonly known refinements in signaling games: the Intuitive Criterion, D1, D2 and Never a Weak Best Response (Cho and Kreps, 1987; Banks and Sobel, 1987). Any equilibrium outcome that could be deleted using these refinement fails our forward induction concept. In this sense our definition is a generalization of their concept.

The concept of extensive form rationalizability by Pearce (1984) is probably the earliest work on iteratively identifying "reasonable" beliefs in an extensive form game. Two other equilibrium concepts — justifiable equilibrium (McLennan, 1985) and explicable equilibrium (Reny, 1992) — apply iterative procedures based on variants of rationalizability to refine sequential and weakly sequential equilibrium. They are probably closest to forward

induction equilibrium. Loosely speaking, justifiable equilibrium iteratively prunes actions that give strictly lower payoffs than the remaining sequential equilibrium payoffs. Explicable equilibrium iteratively prunes strategies that are not best responses to any remaining weakly sequential equilibrium behavioral strategy profiles. One key difference between these concepts and forward induction equilibrium is that their iterative procedures do not anchor to a particular equilibrium outcome while our definition is outcome dependent. McLennan (1985) and Reny (1992) motivate their concepts by suggesting that a deviator knows he is in an equilibrium but is unsure about which one. We maintain a stronger assumption that a deviator knows the equilibrium outcome. Thus, when considering the confusion a deviator could have about the equilibrium plan, we assume the deviator knows the equilibrium path but might be confused over off-equilibrium continuations. Due to this difference, some justifiable or explicable equilibria do not satisfy forward induction, and some forward induction equilibria are not justifiable or explicable (see Section 5.2 for examples).

The rest of the paper is organized as follows: Section 2 contains the notations and the definition of weakly sequential equilibrium, which we will use throughout the paper. The formal definition of forward induction equilibrium is given in Section 3. We examine various properties of forward induction equilibrium in Section 4. Section 5 compares forward induction equilibrium to related equilibrium refinement concepts. Section 6 concludes.

2. Notations and Preliminaries

2.1. **Notations.** We consider only finite extensive form games with perfect recall. A typical game tree is denoted as Γ. The set of player is N. Let X be the set of nodes in the game. For each player $i \in N$, let X_i be the set of decision nodes at which i moves. Let H_i be the set of i's information sets (a partition of X_i). For $x \in X_i$, let $h(x) \in H_i$ be the unique information set of i containing x. The set of actions available to player i at information set $h_i \in H_i$ is given by $A_i(h_i)$. The set of all actions available to player i is given by $A_i = \cup_{h_i \in H_i} A_i(h_i)$. A pure strategy of player i is a function $\pi_i : H_i \to A_i$ with the restriction that $\pi_i(h_i) \in A_i(h_i)$ for all $h_i \in H_i$. Let Π_i be the set of all pure strategy of player i. A mixed strategy of player

i is a distribution over Π_i. A behavioral strategy of player i is a function $b_i : H_i \to \Delta(A_i)$ with the restriction that $\text{supp}\, b(h_i) \subseteq A_i(h_i)$ for all $h_i \in H_i$. The set of all behavioral strategies of player i is denoted as B_i. A belief of a player i is a function $\mu_i : X_i \to [0,1]$ such that $\sum_{x \in h_i} \mu_i(x) = 1$ for all $h_i \in H_i$. An assessment is a belief system and profile of behavioral strategies pair (μ, b). An assessment (μ, b) is *consistent* if there exists a sequence of assessments $\{(\mu^n, b^n)\}_n$ converging to (μ, b) such that for every n, b^n is completely mixed and μ^n is derived from b^n using Bayes' Rule.

Define the partial order $\prec$ on X such that $x \prec y$ if x is on the unique path from the root of the tree to node y. For a node $x \in h_i$ and an action $a \in A_i(h_i)$, write $(x,a) \prec y$ if $x \prec y$ and the unique path from the root to y requires player i to choose action a at information set h_i.

Adopting the notations in Govindan and Wilson (2009a), for each player i, any pure strategy $\pi_i \in \Pi_i$, and any node $y \in X$ (which could be terminal), write $\beta_i(y, \pi_i)$ as the probability that π_i does not exclude y. That is,

$$\beta_i(y, \pi_i) = \begin{cases} 1 & \text{if for each } (x,a) \prec y \text{ where } x \in X_i,\ \pi_i(h(x)) = a \\ 0 & \text{otherwise.} \end{cases}$$

If $\rho_i \in \Delta(\Pi_i)$ is a mixed strategy, extend the definition of β_i such that for any node $y \in X$,

$$\beta_i(y, \rho_i) = \sum_{\pi_i \in \Pi_i} \beta_i(y, \pi_i)\, \rho_i(\pi_i).$$

For a behavioral strategy b_i, $\beta_i(y, b_i)$ is the product of the probabilities that b_i puts on the actions on the path from the root of the tree to y. That is,

$$\beta_i(y, b_i) = \prod_{\substack{x \in X_i \\ (x,a) \prec y}} b_i(h(x))(a).$$

By perfect recall, for any player i, if $x, y \in h_i \in H_i$, then for any pure strategy π_i, $\beta_i(x, \pi_i) = \beta_i(y, \pi_i)$. In this case we will adopt the shorthand $\beta_i(h_i, \pi_i)$. Define similarly the same

shorthand for mixed and behavioral strategies. Finally, define $\beta_0(y)$ as the probability that nature does not exclude y (equal to 1 for all y if nature does not have a move).

Let Z be the set of all terminal nodes in the tree. For each player i, two pure strategies $\pi_i, \pi_i' \in \Pi_i$ are *equivalent* if $\beta_i(z, \pi_i) = \beta_i(z, \pi_i')$ for all $z \in Z$. Let $\{\Pi_i^k\}_{k=1}^K$ denote the partition of Π_i defined by this equivalence relation. For a mixed strategy $\rho_i \in \Delta(\Pi_i)$, the definition of $\beta_i(\cdot, \rho_i)$ means $\beta_i(z, \rho_i)$ depends only on the probability on each equivalence class, but not on how the probabilities are distributed within each equivalence class. Hence, we call two mixed strategies ρ_i, ρ_i' equivalent if they put the same probability on every equivalence class of pure strategies. Now let $s_i^k = \Pi_i^k$ and $S_i = \{s_i^1, \ldots, s_i^K\}$. Then every pure and mixed strategy of player i has an equivalent representation on $\Sigma_i = \Delta S_i$. The set of all mixed strategy profile is $\Sigma = \times_{i \in N} \Sigma_i$.

An *outcome* of the game is a distribution over Z. Given a profile of mixed strategy $\sigma \in \Sigma$, the probability that terminal node z results is given by

$$g_z(\sigma) = \beta_0(z) \prod_{i \in N} \beta_i(z, \sigma_i).$$

The vector $(g_z(\sigma))_{z \in Z}$ gives the outcome (as a distribution over Z) resulting from the strategy profile σ. When no confusion may arise, we will shorthand it as $g(\sigma)$. Let $u_i : Z \to \mathbb{R}$ be player i's utility function over terminal nodes. Using the terminology of Mailath, Samuelson, and Swinkels (1993), the *Pure Strategy Reduced Normal Form* game associated with Γ is $G = (S_i, v_i)_{i \in N}$ where $v_i : \Sigma \to \mathbb{R}$ is defined by $v_i(\sigma) = \sum_{z \in Z} u_i(z) g_z(\sigma)$ for all i.[6]

It is possible that, for some player i, some pure strategy $s_i \in S_i$ is outcome equivalent as a convex combination of other pure strategies. In this case we call s_i a *redundant* strategy. Using again the terminology of Mailath, Samuelson, and Swinkels (1993), the *Mixed Strategy Reduced Normal Form* of Γ is obtained from G by deleting all redundant strategies for all

[6]We define the utility function over mixed strategies profiles rather than pure strategy profiles. This is because utility is not multi-linear in all pure strategies. Nevertheless, given a strategy profile of all other players σ_{-i}, utility of Player i is linear in i's pure strategies as usual.

players. We refer to the mixed strategy reduced normal form when we use the phrase *Reduced Normal Form* in this paper.

Since we work with games with perfect recall, every mixed strategy $\sigma_i \in \Sigma_i$ has a behavioral strategy $b_i \in B_i$ that is realization equivalent to it (Kuhn, 1953). Moreover, if σ_i is completely mixed, we can construct from it an equivalent behavioral strategy b_i by defining: for all information sets $h_i \in H_i$, all actions $a_i \in A_i(h_i)$,

$$b_i(h_i)(a_i) = \frac{\sum_{s_i \text{ s.t. } \beta_i((x,a_i),s_i)>0 \text{ for } x \in h_i} \sigma_i(s_i)}{\sum_{s_i' \text{ s.t. } \beta_i(h_i,s_i)>0} \sigma_i(s_i')}.$$

In this case we say σ_i *induces* b_i. If $\{\sigma^n\}$ is a sequence of completely mixed strategy profiles converging to σ on Σ, we say $\{\sigma^n\}$ induces the consistent assessment (μ, b) where b is realization equivalent to σ if there exists a sequence of assessments $\{(\mu^n, b^n)\}_n \to (\mu, b)$ such that each b^n is induced by σ^n; and μ^n is derived from b^n using Bayes' Rule. Conversely, if b is a profile of behavioral strategies, we say it *corresponds* to a profile of mixed strategies $\sigma \in \Sigma$ if b is realization equivalent to σ.

2.2. **Weakly Sequential Equilibrium.** We now introduce the concept of weakly sequential equilibrium and a variant of it, which we adopt for the rest of this paper. First, we say that a behavioral strategy b_i of player i *excludes* an information set $h_i \in H_i$ of his if $\beta_i(h_i, b_i) = 0$. If instead $\beta_i(h_i, b_i) > 0$, we say b_i *enables* h_i.

Definition 2.1. (Reny, 1992) A consistent assessment (μ, b) of the extensive form game Γ is a *weakly sequential equilibrium* if, for all player $i \in N$, for all information sets $h_i \in H_i$ such that $\beta_i(h_i, b_i) > 0$,

$$b_i \in \arg\max_{b_i' \in B_i} \mathrm{E}_{\mu,(b_i', b_{-i})}\left[u_i(z) \mid h_i\right].$$

In other words, a weakly sequential equilibrium is a consistent assessment that constitutes an optimal continuation for each player at every information set enabled by his behavioral strategy. Notice however that the maximization is taken over all behavioral strategies, not just those agreeing with b_i at all enabled information sets. A weakly sequential equilibrium

differs from a sequential equilibrium (Kreps and Wilson, 1982) as a player's strategy need not be optimal at information sets that can never be reached given his equilibrium strategy. Every sequential equilibrium is weakly sequential, but the converse is not true unless no player moves twice along any path. For simultaneous move games, all Nash equilibria are weakly sequential.

Reny (1992) interprets a weakly sequential equilibrium by partitioning a player's behavioral strategy b_i into two parts: the first part is player i's *rational plan* of actions at all information sets that could be reached with positive probability given his plan; the second part is a *prediction* of player i's behavior at his excluded information sets — that is, after i has deviated from his rational plan. Of course, this prediction and i's opponents response to this prediction should in turn render i's equilibrium plan optimal.

Our reason for adopting weakly sequential equilibrium and hence abandoning backward induction is more practical than philosophical. In some sense, backward induction treats deviations as "mistakes" and players are still optimizing after the unlikely event of mistakes. Moreover, mistakes across information sets are uncorrelated. One mistake bears no information on the likelihood of future mistakes. Under this assumption, making inferences on perceived deviations is futile. This goes against the spirit of forward induction. Weakly sequential equilibrium offers the advantage of an equilibrium analysis without backward induction. Whether it is philosophically well-founded is beyond the scope of this paper.

The above definition of weakly sequential equilibrium, however, allows a player to use a weakly dominated behavioral strategy. This is undesirable. To achieve admissibility, we strengthen the definition of weakly sequential equilibrium to weakly quasi-perfect equilibrium as follows:

Definition 2.2. An assessment (μ, b) of an extensive form game Γ is a *weakly quasi-perfect equilibrium* if there exists a sequence of assessments with completely mixed behavioral strategies $\{(\mu^n, b^n)\}$ converging to (μ, b) such that:

(1) for all n, μ^n is derived from b^n using Bayes' Rule (that is, (μ, b) is consistent); and

(2) for all player $i \in N$, all information set $h_i \in H_i$ not excluded by b_i,

$$b_i \in \arg\max_{b_i' \in B_i} \mathrm{E}_{\mu^n, \left(b_i', b_{-i}^n\right)} \left[u_i\left(z\right) \mid h_i\right] \quad \text{for all } n.$$

That is, we strengthen the definition of a weakly sequential equilibrium by requiring the equilibrium strategy to be a best response to a sequence of assessments converging to the equilibrium assessment. We do not perturb nature's strategy in this definition. Nor do we perturb player i's own behavioral strategies at continuation information sets. The relationship between a weakly sequential equilibrium and a weakly quasi-perfect equilibrium is analogous to that between a sequential equilibrium (Kreps and Wilson, 1982) and a quasi-perfect equilibrium (van Damme, 1984).

The following equivalence between weakly quasi-perfect equilibrium and normal form perfect equilibrium (Selten, 1975) is demonstrated by Reny (1992):

Fact 2.3. *Let Γ be an extensive form game and G be its (reduced) normal form representation. Then*

(1) If σ is a normal form perfect equilibrium in G with justifying sequence $\{\sigma^n\}$, then there exists a sequence of assessments $\{(\mu^m, b^m)\}$ induced by a subsequence of $\{\sigma^n\}$, and $\{(\mu^m, b^m)\}$ justifies a weakly quasi-perfect equilibrium (μ, b) in Γ such that σ and b are realization equivalent.

(2) If (μ, b) is a weakly quasi-perfect equilibrium in Γ justified by a sequence of assessments $\{(\mu^n, b^n)\}$, then there is a sequence of completely mixed strategies $\{\sigma^n\}$ in G such that each σ^n corresponds to b^n and $\{\sigma^n\}$ justifies a normal form perfect equilibrium σ in G corresponding to b.

Proof. See Reny (1992, p. 632-633). □

Due to this equivalence, we state most of the definitions in this paper using only the reduced normal form representation. We trust that readers could readily translate our definitions into the language of extensive form representation if they so wish.

In addition to its equivalence to weakly quasi-perfect equilibrium, normal form perfect equilibrium has its own merits. As Blume, Brandenburger, and Dekel (1991) show (using the lexicographic belief system representation), a normal form perfect equilibrium is equivalent to an admissible Nash equilibrium satisfying the common prior and strong independence assumptions. It is also invariant to adding or deleting redundant strategies. Thus it is a mild refinement of Nash equilibrium simple enough to be the basic building block of other refinement concepts.

3. Definitions

3.1. **Definition by Govindan and Wilson.** Let us first review the definition of forward induction given by Govindan and Wilson (2009b) in their main text. Consider an outcome $\zeta \in \Delta(Z)$ of an extensive form game. A pure strategy π_i of player i is *relevant for* ζ if there is a weakly sequential equilibrium (μ, b) giving outcome ζ such that π_i constitutes an optimal continuation given μ_i and b_{-i} at every information set h_i enabled by π_i. Notice that a relevant strategy π_i may or may not be in the support of the mixed strategy realization equivalent to b_i. An information set h is *relevant for* ζ if it can be reached under some profile strategies relevant for ζ (not necessarily an equilibrium profile). A forward induction outcome is then defined as follows:

Definition 3.1. (Govindan and Wilson, 2009b) An outcome ζ satisfies *forward induction* if it results from a weakly sequential equilibrium (μ, b) such that, at every information set h_i relevant for ζ, the support of μ_i at h_i is confined to nodes that can be reached by profiles of other players' strategy π_{-i} that are relevant for ζ.

The intuition behind this definition merits some discussion. Recall our interpretation that a weakly sequential equilibrium strategy could be partitioned into two parts: a rational plan of a player and a prediction of his behavior had he deviates from his rational plan. Since a player is supposed to follow his rational plan, he knows the outcome arising from the rational plan. However, a player could potentially be confused about the prediction of

his behavior after his deviations and hence not his opponents' response after his deviation. In other words, he might have in mind another weakly sequential equilibrium with the same equilibrium path (hence outcome) but different off-equilibrium continuations. A pure strategy is relevant for this outcome if it is a best response to one of these other weakly sequential equilibria. Now if indeed Player i deviates, and the other players would like to maintain the view that: (1) Player i knows that they are playing a weakly sequential equilibrium with this particular outcome; and (2) Player i is an expected utility maximizer; then the other players should believe that Player i must be mixing between his relevant strategies. Of course, this restriction applies only if the information set could be reached with some relevant strategies, and that the opponent himself has not deviated (if he had, he could still maintain this belief but his behavior there need not be optimal given this belief anyway). Hence, forward induction — maintaining (1) and (2) whenever possible — can be expressed as a restriction on the support of beliefs at all relevant information sets. Finally, this belief system and the optimal response of other players should render it optimal for Player i to follow his equilibrium plan.

As an illustration, apply this definition to the Outside Option Game depicted in Figure 1.1. Consider the outcome $(2, 2)$ and let μ be the probability Player 2 puts on the node after T. Then any weakly sequential equilibrium assessment giving outcome $(2, 2)$ is of either the form $\langle (\text{Out},\ (\alpha, 1-\alpha)), \mu = \frac{3}{4} \rangle$ where $\alpha \leq \frac{2}{3}$; or $\langle (\text{Out}, R), \mu \rangle$ where $\mu \leq \frac{3}{4}$ (notice that Player 1 need not optimize at her second information set, as it is excluded by the strategy Out). For Player 1, both the pure strategies Out and T are relevant for $(2, 2)$: Out because it is the equilibrium strategy, T because it is a best response in $\langle \left(\text{Out},\ \left(\frac{2}{3}, \frac{1}{3}\right)\right), \frac{3}{4} \rangle$. The pure strategy B, however, is not relevant for $(2, 2)$ since Out always gives a strictly higher payoff. For Player 2, both of his pure strategies are relevant for $(2, 2)$ since they are both equilibrium strategies in some of these equilibria. Given the relevant strategies, both information sets are relevant: Player 1's information set is the root of the tree; Player 2's information set is enabled by the relevant strategy T. Now $(2, 2)$ does not satisfy forward induction, since in

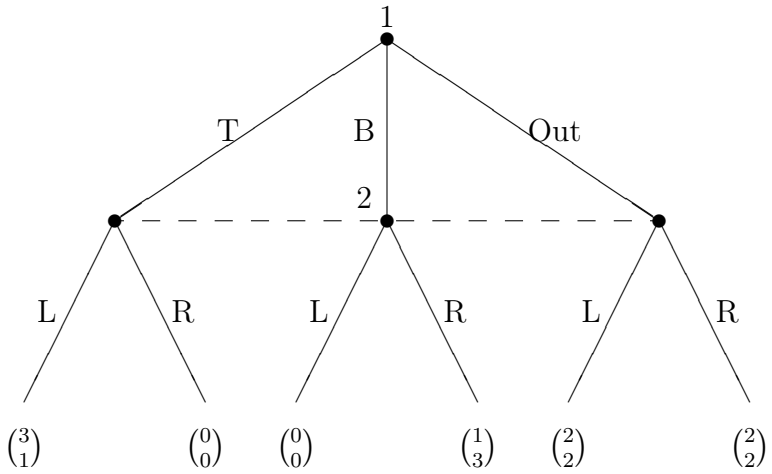


FIGURE 3.1. Outside Option Game: A Variant

every weakly sequential equilibrium giving $(2, 2)$, Player 2's belief at his relevant information set puts positive probability on the irrelevant strategy B.

Unfortunately, Definition 3.1 is not an invariant concept. That is, an outcome could satisfy forward induction in an extensive form game but not in another with the same reduced normal form. To see this, consider Figure 3.1, an extensive form game obtained from Figure 1.1 by coalescing the information sets of Player 1 and adding a superfluous move for Player 2. Both games have the same reduced normal form given in Figure 1.2. Again consider the outcome $(2, 2)$ (the two terminal nodes giving $(2, 2)$ are considered the same). Using almost the same argument as before, Out, T, L and R, as well as both information sets are all relevant for $(2, 2)$. Note that (Out, R), together with the belief putting probability 1 on Out, is a weakly sequential equilibrium. Since Out is a relevant strategy, this belief system satisfies the restriction in Definition 3.1. Therefore $(2, 2)$ is a forward induction outcome of this game.

In general, if we consider a finite extensive form game Γ with perfect recall, then there exists another extensive form game Γ' with the same pure strategy reduced normal form such that all Nash equilibrium outcomes of Γ' (hence that of Γ) satisfy forward induction according to Definition 3.1. To see this, consider the trivial simultaneous move extensive form game corresponding to the pure strategy reduced normal form. Since it is a simultaneous move

game, all Nash equilibria are sequential, hence weakly sequential. Consistency guarantees the belief system puts positive probability only on nodes reached by equilibrium strategies — which are all relevant for that equilibrium outcome — at every information set in every sequential equilibrium. Hence every Nash equilibrium outcome satisfies forward induction according to Definition 3.1. Forward induction loses all its refinement power in this extensive form game.

One may argue that this is not a defect of the definition, but forward induction itself should not be an invariant concept. After all, there is nothing "forward" for Player 2 to induct on in Figure 3.1. We do not find this argument convincing. First of all, while Player 2 would not be able to "observe the unexpected" in this simultaneous move game, he could realize that his choice matters only when Player 1 has chosen T or B. Player 2 can reason: what are the circumstances when my choice makes a difference so that I should choose carefully? Player 1 can get a payoff of 2 for sure if she follows her equilibrium plan. Thus there is no way for her to choose B rationally. Player 2 can then deduce that he should choose L in response to T. This line of reasoning is consistent with forward induction, and could be carried out even in the representation depicted by Figure 3.1. Thus forward induction reasoning is possible as long as players (or theorists) have information on the reduced normal form in this example. In general, one could utilize the normal form information sets defined by Mailath, Samuelson, and Swinkels (1993) to make similar forward induction arguments in normal form games. Since in many applications, it is not clear if economic agents have a particular extensive form representation in mind, an applicable definition of forward induction should not depend on the specific ways one presents an extensive form game.

3.2. **First-Order Forward Induction Equilibrium.** Careful readers may notice that, in the game depicted in Figure 3.1, there is no equilibrium with outcome $(2, 2)$ if Player 2 is optimizing against a sequence of completely mixed strategy of Player 1 such that the relevant strategy T is infinitely more likely than the irrelevant strategy B at the limit. This observation leads us naturally to adopt the stronger version of weakly quasi-perfect

equilibrium. Since weakly quasi-perfection on the extensive form is equivalent to normal form perfection (Fact 2.3), we have the luxury of working with the reduced normal form game. The following definitions follow those given by Govindan and Wilson (2009b) in their Appendix B.

Definition 3.2. Let $\zeta \in \Delta(Z)$ be a normal form perfect equilibrium outcome in the (reduced) normal form G. A pure strategy $s_i \in S_i$ of player i is (first-order) *relevant for* ζ if there exists a sequence of ε-perfect equilibrium $\{\sigma^n\}$ converging to a strategy profile $\sigma \in \Sigma$ giving outcome ζ, such that

$$v_i\left(s_i, \sigma_{-i}^n\right) \geq v_i\left(s_i', \sigma_{-i}^n\right) \quad \text{for all } s_i' \in S_i, \text{ for all } n.$$

In other words, a relevant strategy for an outcome is a strategy which could be adopted rationally and cautiously by a player who knows the equilibrium outcome but is unsure which normal form perfect equilibrium is in effect. Since we are requiring best response to a sequence of completely mixed strategies, no dominated strategy (weakly or strictly) could be relevant for any outcome. One may feel that, since relevant strategies are potential explanations for deviations, they need not satisfy as strict a requirement as equilibrium strategies. It is possible to weaken our definition by allowing all best responses in some normal form perfect equilibrium with a certain outcome to be relevant for that outcome. This would allow for more forward induction outcomes. We do not find this definition very useful, though, since in normal form games obtained from an extensive form game, many "suboptimal" strategies can be best responses to an equilibrium as long as the information set in which they matter are excluded by other players' strategies.

We now give the normal form version of Definition 3.1:

Definition 3.3. An outcome $\zeta \in \Delta(Z)$ satisfies (first-order) *forward induction* if there exists a sequence of ε-perfect equilibria $\{\sigma^n\}$ converging to a strategy profile $\sigma \in \Sigma$ with $g(\sigma) = \zeta$, such that for all player $i \in N$, for any $s_i, s_i' \in S_i$, if s_i is relevant for ζ and s_i' is

not, then

$$\lim_{n\to\infty} \frac{\sigma_i^n(s_i')}{\sigma_i^n(s_i)} = 0.$$

The requirement suggests that any relevant strategy is infinitely more likely than any irrelevant strategy at the equilibrium along the sequence of ε-perfect equilibria justifying σ. This supplement the intuition behind Definition 3.1 with a form of cautious behavior.

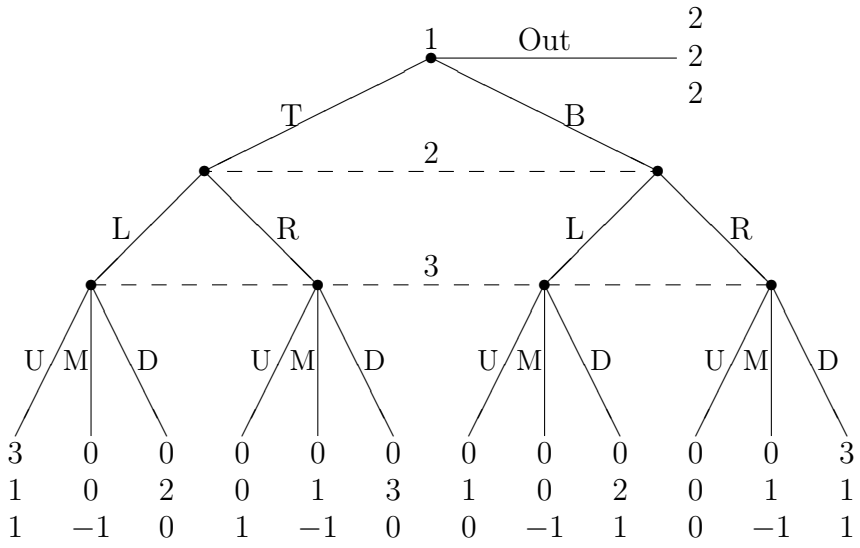
Definition 3.3 gives the condition for an *outcome* to satisfy forward induction. We might nonetheless wish to define forward induction as a property of an *equilibrium*. For this purpose we give the following definition of a forward induction equilibrium:

Definition 3.4. A normal form perfect equilibrium $\sigma \in \Sigma$ in G is a (first-order) *forward induction equilibrium* if there exists a sequence of ε-perfect equilibria $\{\sigma^n\}$ converging to σ such that for all player $i \in N$, all pairs of pure strategies $s_i, s_i' \in S_i$, if s_i is relevant for the outcome $g(\sigma)$ and s_i' is not, then

$$\lim_{n\to\infty} \frac{\sigma_i^n(s_i')}{\sigma_i^n(s_i)} = 0.$$

By definition, a forward induction outcome is the outcome of a forward induction equilibrium.

3.3. **An Iterative Definition.** The definition above suggests an algorithm for checking for forward induction: Take a set of normal form perfect equilibria with a certain outcome. Look for all the pure strategies that are best response to some justifying sequence of some equilibrium in the set. Call them relevant strategies for this outcome. Then select the equilibria that could be justified by some sequence in which any relevant strategy is infinitely more likely than any irrelevant strategy at the limit. We may wish, for a consistent motivation, that the set of all forward induction equilibria for a given outcome would be a fixed point of this procedure. Such a set would be a "fixed point" for forward induction reasoning: every equilibria in the set can be justified by all players believing that other players are infinitely



(a) Extensive Form

	L	R
Out	2,2,2	2,2,2
T	3,1,1	0,0,1
B	0,1,0	0,0,0

U

	L	R
Out	2,2,2	2,2,2
T	0,0,-1	0,1,-1
B	0,0,-1	0,1,-1

M

	L	R
Out	2,2,2	2,2,2
T	0,2,0	0,3,0
B	0,2,1	3,1,1

D

(b) Normal Form

FIGURE 3.2. Three-Person Outside Option Game

more likely to play best responses in (a neighborhood of) this set. Unfortunately, this is not true with Definition 3.4. To see this, consider the three-person game depicted in Figure 3.2.

In this game, the outcome $(2, 2, 2)$ satisfies first-order but not "second-order" forward induction. To see this, first note that Player 3 always prefers U to D if she believes that the relative probability of T to B is greater than $\frac{1}{2}$. Next, consider the sequence of ε-perfect equilibria $\langle (1-2\varepsilon, \varepsilon, \varepsilon), \left(\frac{2}{3(1-2\varepsilon)}, 1-\frac{2}{3(1-2\varepsilon)}\right), (\varepsilon, \varepsilon, 1-2\varepsilon)\rangle$. It converges to $\left(\text{Out}, \left(\frac{2}{3}, \frac{1}{3}\right), D\right)$. Out, B, L, R, U and D are all best responses to this sequence. There is another sequence of ε-perfect equilibria, $\langle (1-2\varepsilon, \varepsilon, \varepsilon), \left(\frac{2}{2+\varepsilon}, \frac{\varepsilon}{2+\varepsilon}\right), \left(\frac{2+\varepsilon}{3}, \frac{\varepsilon}{3}, \frac{1-2\varepsilon}{3}\right)\rangle$, which converges to $\left(\text{Out}, L, \left(\frac{2}{3}, 0, \frac{1}{3}\right)\right)$. Out, T, L, U and D are best responses to this sequence. Thus all pure strategies but M (which is dominated) are first-order relevant for the outcome $(2, 2, 2)$. The

second sequence of ε-perfect equilibria satisfies the restrictions on relative probabilities between relevant and irrelevant strategies. Therefore $(2,2,2)$ is a first-order forward induction outcome.

However, when M is infinitely less likely than U, Player 1 needs to play T with probability strictly higher than $\frac{1}{2}$ for R to be a best response. Yet whenever this is the case, Player 3 strictly prefers U to her other strategies and R is an inferior response to U. Thus in no first-order forward induction equilibrium can R receive positive probability. Since Player 1's strategy B gets a strictly lower payoff than Out unless Player 2 plays R, B cannot be second-order relevant. On the other hand, T remains second-order relevant since it is a best response to the second sequence of ε-perfect equilibria we constructed, which converges to a first-order forward induction equilibrium. Now if Player 3 believes T to be infinitely more likely than B, she plays U with probability 1, causing Player 2 to best respond by L. Player 1 can profitably deviate to T. Therefore $(2,2,2)$ does not satisfy second-order forward induction.

A natural response is to starting from the set of all normal form perfect equilibria giving a certain outcome and keep iterating our "forward induction algorithm" until no more strategies can be pruned. To formalize this idea, define, for any outcome $\zeta \in \Delta(Z)$,

$$\begin{aligned} F^{-1}(\zeta) &= \{\sigma \in \Sigma : g(\sigma) = \zeta\} \\ R_i^0(\zeta) &= S_i \qquad\qquad \text{for all } i \in N. \end{aligned}$$

Then for any $k \geq 0$, given F^l and R_i^{l+1} for all $i \in N$, all $l = -1, \ldots, k-1$, define inductively

$$F^k(\zeta) = \left\{ \sigma \in F^{k-1}(\zeta) \;\middle|\; \begin{array}{l} \exists \text{ a sequence of } \varepsilon\text{-perfect equilibria} \{\sigma^n\} \to \sigma \\ \text{with } g(\sigma) = \zeta \text{ such that:} \\ \forall l \in \{0, \ldots, k\}, \forall i \in N, \forall s_i, s_i' \in S_i, \\ \qquad s_i \in R_i^l \text{ and } s_i' \notin R_i^l \Rightarrow \frac{\sigma_i^n(s_i')}{\sigma_i^n(s_i)} \to 0. \end{array} \right\}; \tag{3.1}$$

and

$$R_i^{k+1}(\zeta) = \left\{ s_i \in R_i^k(\zeta) \;\middle|\; \begin{array}{l} \exists \text{ a sequence of } \varepsilon\text{-perfect equilibria} \{\sigma^n\} \to \sigma \\ \text{satisfying the conditions in (3.1), such that} \\ v_i\left(s_i, \sigma_{-i}^n\right) \geq v_i\left(s_i', \sigma_{-i}^n\right) \text{ for all } s_i' \in S_i, \text{ for all } n. \end{array} \right\}. \tag{3.2}$$

So $F^0(\zeta)$ is the set of all normal form perfect equilibria giving outcome ζ; $R_i^k(\zeta)$ is the set of player i's strategies kth-order relevant for ζ for all $k \geq 1$; and $F^k(\zeta)$ is the set of kth-order forward induction equilibria for outcome ζ. We define forward induction equilibrium and outcome as follows:

Definition 3.5. For all $k \geq 1$, a normal form perfect equilibrium $\sigma \in \Sigma$ is a *kth-order forward induction equilibrium* if $\sigma \in F^k(g(\sigma))$. It is a *forward induction equilibrium* if it is a kth-order forward induction equilibrium for all $k \geq 1$. An outcome ζ satisfies kth-order forward induction if it is the outcome of a kth-order forward induction equilibrium. It satisfies forward induction if it is the outcome of a forward induction equilibrium.

One can easily check that Definitions 3.3 and 3.4 are the $k = 1$ version of this definition.

Given an outcome, denote $F(\zeta) = \cap_{k \geq 1} F^k(\zeta)$ as the set of all forward induction equilibria with outcome ζ and likewise $R_i(\zeta) = \cap_{k \geq 1} R_i^k(\zeta)$ as all strategies of player i relevant for the outcome ζ. Since the game is finite, if $F(\zeta)$ is non-empty, then there is a finite K such that $F^K(\zeta) = F(\zeta)$ and $R_i^K(\zeta) = R_i(\zeta)$ for all player i. The set $F(\zeta)$ can therefore be viewed as a "fixed point" of the forward induction iteration defined by equations (3.1) and (3.2).

We wish to remark on our definition before proceeding to analyze the properties of forward induction equilibrium. Condition (b) in equation (3.1) imposes a sophisticated lexicographic ordering of a player's pure strategies that is commonly known to all players, including the

deviators. This assumption is rather strong[7]. There may be applications in which one would wish to stop after a few rounds of iteration instead.

4. Properties of Forward Induction Equilibrium

4.1. **Existence.** As the term "forward induction" is used by Kohlberg and Mertens (1986) to describe a property of their Stable Set, it is not surprising that there is a close relationship between Stable Set and Forward Induction Equilibrium. Recall that a closed set of Nash equilibria is said to be pre-stable in a game G if, for any $\varepsilon > 0$ there exists a $\delta_0 > 0$ such that, for any completely mixed strategy vector $(\sigma_1, \ldots, \sigma_N)$, and any vector of strictly positive real numbers $\delta = (\delta_1, \ldots, \delta_N)$ with $\|\delta\| < \delta_0$, the perturbed game obtained from replacing every pure strategy s_i of player i by $(1 - \delta_i)\, s_i + \delta_i \sigma_i$ has a Nash equilibrium ε-close to the set. A set of Nash equilibria is *Stable* if it is a minimal (by set inclusion) pre-stable set. If all equilibria in a stable set give the same outcome, we will call it a *constant outcome stable set* and the outcome a *stable outcome*. It turns out that every constant outcome stable set contains a forward induction equilibrium and every stable outcome is a forward induction outcome.

Proposition 4.1. *Every constant outcome stable set contains a forward induction equilibrium and every stable outcome is a forward induction outcome.*

Proof. Let E be a constant outcome stable set of a game G and ζ be the outcome of every equilibrium in E. For each player i, let Σ_i^0 be the set of all his completely mixed strategies.

[7]One might therefore be tempted to define a set of equilibria as "forward induction set" if it is a fixed point of our forward induction operator. This approach is unsatisfactory. If we choose a small set, we could admit too many outcomes: In the Outside Option Game example in Figure 1.2, the singleton set $\{(\text{Out}, R)\}$ would pass an analog of an iteration defined by equations (3.1) and (3.2). Yet we certainly would not want to call it a forward induction equilibrium. On the other hand, if we require the set to contain all normal form perfect equilibria with the same outcome, we can easily run into non-existence. It is hard to argue *a priori* why we should consider some sets but not others.

Take some $\varepsilon > 0$, define for each player i,

$$\begin{aligned}
\Sigma_i^0(\varepsilon) &= \Sigma_i^0 \\
\Sigma_i^k(\varepsilon) &= \left\{ \sigma_i \in \Sigma_i^0 \;\middle|\; \begin{array}{l} \text{For all } l \in \{1,\ldots,k\}, \text{ all } s_i, s_i' \in S_i, \\ \text{if } s_i \in R_i^l(\zeta) \text{ and } s_i' \notin R_i^l(\zeta), \\ \text{then } \sigma_i(s_i') < \varepsilon \sigma_i(s_i). \end{array} \right\} \text{ for all } k > 0.
\end{aligned}$$

For ε sufficiently small, as long as $R_i^k(\zeta)$ is non-empty, $\Sigma_i^k(\varepsilon)$ is non-empty. Moreover, with a proper rescaling of ε to ε', $\Sigma_i^{k+1}(\varepsilon') \subseteq \Sigma_i^k(\varepsilon)$ for all $k \geq 0$. Next define, for each $k \geq 0$, P^k as the set of sequences of perturbations of G, $\{G(\varepsilon)\}$ (with $\varepsilon \to 0$) such that, for each ε, $G(\varepsilon)$ is defined by replacing each pure strategy $s_i \in S_i$ by $(1-\varepsilon)s_i + \tilde{\sigma}_i(\varepsilon)$ where $\tilde{\sigma}_i \in \Sigma_i^k(\varepsilon)$ for all player $i \in N$. Again, as long as $\Sigma_i^k(\varepsilon)$ is non-empty, P^k is well-defined and $P^k \subseteq P^{k-1} \subseteq \cdots \subseteq P^0$.

Let $NE(G(\varepsilon))$ be the set of Nash equilibrium of the game $G(\varepsilon)$. Now define inductively

$$\begin{aligned}
E^0 &= E \\
E^k &= \left\{ e \in E^{k-1} \;\middle|\; \begin{array}{l} \exists \{G(\varepsilon)\} \in P^k, \text{ a sequence } \{\sigma(\varepsilon)\} \\ \text{with } \sigma(\varepsilon) \in NE(G(\varepsilon)) \text{ for all } \varepsilon \\ \text{such that } \sigma(\varepsilon) \to e. \end{array} \right\} \text{ for all } k > 0.
\end{aligned}$$

Since E is a stable set, for every sequence of perturbations in P^0, there exists an $e \in E = E^0$ such that e is the limit of a sequence of Nash equilibria of the perturbed games. Now if for all $l \leq k$, for every sequence of perturbations in P^l, there exists an $e \in E^l$ such that e is the limit of a sequence of Nash equilibria of the perturbed games, then since $P^{k+1} \subseteq P^k$ the same is true for $k+1$ (as long as $R_i^{k+1}(\zeta)$ is non-empty for all i so that P^{k+1} is well-defined).

We wish to show that for all $k \geq 0$, E^k is non-empty and $E^k \subseteq F^k(\zeta)$. For $k = 0$ this follows from the fact that all equilibria in a stable sets are normal form perfect equilibria. Now suppose for all $l = 1,\ldots,k$, E^l is non-empty and $E^l \subseteq F^l(\zeta)$. Then since $E^k \subseteq F^k(\zeta)$, and all equilibrium strategies of a kth-order forward induction equilibrium with outcome ζ are $(k+1)$th-order relevant for ζ, $R_i^{k+1}(\zeta)$ is non-empty for all i. So $\Sigma_i^{k+1}(\varepsilon)$ (for ε sufficiently

small) is non-empty for all i and $P^{k+1} \subseteq P^k$ is well-defined. By our earlier argument, E^{k+1} is non-empty. It remains to show that every equilibrium $e \in E^{k+1}$ is a $(k+1)$th-order forward induction equilibrium.

If $e \in E^{k+1}$, then there exists a sequence of Nash equilibria of perturbed games satisfying the conditions of P^{k+1} such that $\sigma(\varepsilon) \to e$. We claim that for each player i, for all ε small, $\sigma_i(\varepsilon)$ assigns zero probability to any pure strategy $s_i \notin R_i^{k+1}(\zeta)$. To see this, first note that any $s_i \in S_i \setminus R_i^1(\zeta)$ is not a best response to any justifying sequence for any normal form perfect equilibrium with outcome ζ. Thus s_i cannot be a best response in any Nash equilibrium of perturbed games sufficiently close to e. Hence $\sigma_i(\varepsilon)(s_i) = 0$. Write $\hat{\sigma}_i(\varepsilon)$ as the effective distribution induced by $\sigma_i(\varepsilon)$ on S_i (that is, for each $s_i \in S_i$, let $\hat{\sigma}_i(\varepsilon)(s_i) = (1-\varepsilon)\sigma_i(\varepsilon)(s_i) + \varepsilon\tilde{\sigma}_i(\varepsilon)(s_i)$). Take $l \leq k+1$. If for all $j < l$, $s_i \notin R_i^j(\zeta)$ receives zero probability according to $\sigma_i(\varepsilon)$, then $\{\hat{\sigma}(\varepsilon)\}$ is a sequence of ε-perfect equilibria converging to e such that: For all $i \in N$, all $j = 1, \ldots, l-1$, any $s_i, s_i' \in S_i$, if $s_i \in R_i^j(\zeta)$ and $s_i' \notin R_i^j(\zeta)$, then

$$\frac{\hat{\sigma}_i(\varepsilon)(s_i')}{\hat{\sigma}_i(\varepsilon)(s_i)} < \frac{\varepsilon\tilde{\sigma}_i(s_i)}{\tilde{\sigma}_i(s_i)} = \varepsilon \to 0.$$

By definition, any $s_i \notin R_i^l(\zeta)$ cannot be a best response to such a sequence and hence $\sigma(\varepsilon)(s_i) = 0$ for such s_i. Inducting on l gives a sequence of completely mixed strategy profiles $\hat{\sigma}(\varepsilon) \to e$ which satisfies the conditions in equation (3.1) for $F^{k+1}(\zeta)$. As the choice of $e \in E^{k+1}$ is arbitrary, $E^{k+1} \subseteq F^{k+1}(\zeta)$.

By now we have shown that

$$E^k \cap F^k(\zeta) = E^k \quad \text{for all } k \geq 0;$$

and $\left\{E^k\right\}$ is a decreasing (by set inclusion) sequence of non-empty, closed sets in a compact space. Therefore,

$$\bigcap_{k \geq 0} \left(E^k \cap F^k(\zeta)\right) \neq \emptyset.$$

Thus there exists a $e \in E$ such that $e \in \cap_{k \geq 0} F^k(\zeta) = \cap_{k \geq 0} F^k(g(e))$. By definition, e is a forward induction equilibrium and ζ is a forward induction outcome. $\square$

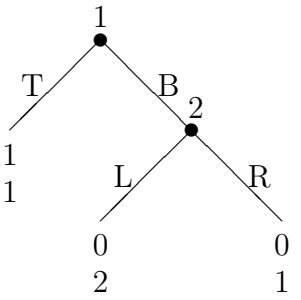
Since every finite generic extensive form game with perfect recall has a stable outcome (Kohlberg and Mertens, 1986), the existence of forward induction equilibrium in generic extensive form games is immediate:

Corollary 4.2. *Every finite generic extensive form game with perfect recall has a forward induction equilibrium, hence a forward induction outcome.*

Since every hyperstable set contains a fully stable set which contains a stable set (Kohlberg and Mertens, 1986), Proposition 4.1 remains true if we replace "stable" by "hyperstable" or "fully stable". Hence all strict Nash equilibria are forward induction equilibria (since they are singleton hyperstable sets), so are all completely mixed equilibria (since all strategies are relevant). However, a quasi-strict equilibrium (a strategy profile σ such that for all players i, the support of σ_i equals the set of pure best responses to σ_{-i}) may not be a forward induction equilibrium. In the Outside Option Game (Figure 1.2), $\left(\text{Out}, \left(\frac{1}{2}, \frac{1}{2}\right)\right)$ is a quasi-strict equilibrium but not a forward induction equilibrium.

Forward induction equilibria are admissible since they are normal form perfect equilibria. Our definition ensures that they are invariant — since normal form perfect equilibrium is invariant to adding a mixed strategy as a pure one and a mixed strategy can only be a best response if all pure strategies in its support are.

Fixing an outcome, the set of all forward induction equilibria giving that outcome need not be connected. Figure 4.2 provides an example. In this game, there are two disjoint components of Nash equilibria in which Player 1 chooses T, giving outcome $(2, 0, 0)$: one in which the product of Player 2's probability on L and Player 3's probability on l is strictly greater than that on R and r, and the expected payoff of Player 1 from going into the game is less than 2; another similar one but the product of probability on L and l is strictly less than that on R and r. It can be shown (see Section 4.2) that all equilibria in these two components are forward induction equilibria giving outcome $(2, 0, 0)$, but the two components are not connected.



(a) Extensive Form

	L	R
T	1,1	1,1
B	0,2	0,1

(b) Normal Form

FIGURE 4.1

It is also not true that if an equilibrium satisfies forward induction, then any equilibrium in the same component of Nash equilibria would satisfy forward induction. In Figure 4.1, the game has a single connected component of Nash equilibria, all of the form $(T, (\alpha, 1-\alpha))$ where $\alpha \in [0,1]$. However, only (T, L) among them is a forward induction equilibrium as all other equilibria put positive probability on the weakly dominated strategy R.

4.2. **Backward Induction.** Since a (constant outcome) stable set need not contain a sequential equilibrium nor gives a sequential equilibrium outcome, it follows from Proposition 4.1 that forward induction equilibrium and outcome need not satisfy backward induction. To be precise, a forward induction equilibrium of a normal form game G need not be a backward induction equilibrium on an extensive form game with reduced normal form G. There are also forward induction outcomes which are not sequential equilibrium outcomes on an extensive form with the same reduced normal form. Figure 11 in Kohlberg and Mertens (1986), which they attribute to Faruk Gul, gives one such example. For the readers' convenience the extensive form of the game is reproduced in Figure 4.2. In this game, the set of Nash equilibria $\{(T, L, l) \cup (T, R, r)\}$ is a stable set giving outcome $(2, 0, 0)$. By Proposition 4.1, $(2, 0, 0)$ is a forward induction outcome. Both equilibria in the set are forward induction equilibria (M is a best response to $\left((1-2\varepsilon, \varepsilon, \varepsilon), \left(\frac{1}{11}, \frac{10}{11}\right), \left(\frac{1}{5}, \frac{4}{5}\right)\right)$ and B is one to $\left((1-2\varepsilon, \varepsilon, \varepsilon), \left(\frac{25}{31}, \frac{6}{31}\right), \left(\frac{9}{10}, \frac{1}{10}\right)\right)$ so all strategies are relevant for $(2, 0, 0)$). However, in the

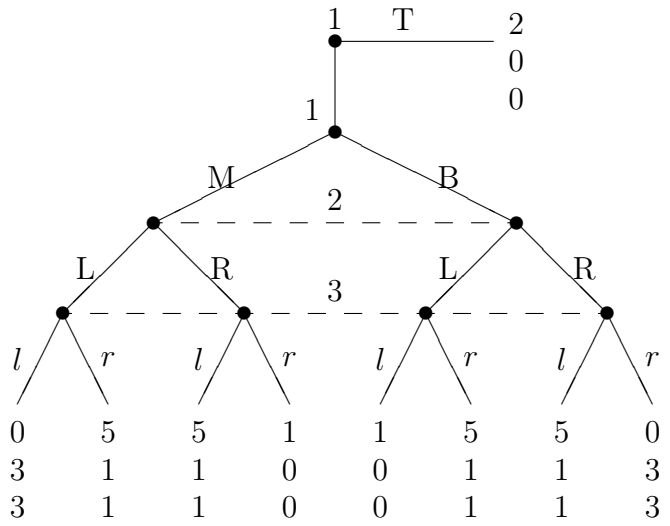


FIGURE 4.2. Gul's Example

representation in Figure 4.2, the unique sequential equilibrium is $\left(\left(0, \frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{2}, \frac{1}{2}\right)\right)$ with payoffs $\left(\frac{11}{4}, \frac{5}{4}, \frac{5}{4}\right)$.

One may object to our definition of forward induction based on this example. After all, we start with some weakly sequential equilibria (all those in which Player 1 plays T) which are not sequential in this tree. Take (T, L, l) as an example. Player 2 and 3 must believe M is relatively more likely than B for their actions to be optimal. Yet given (L, l), Player 1 strictly prefers B to M at her second information set. However, if Player 1 is maximizes her expected utility and knows the plan of (L, l), she should have chosen T in the first place. That Player 2 and 3 gets to move is an evidence that either Player 1 does not maximizes expected utility (in which case there is no need to require her to optimize) or she does not know the plan of 2 and 3 is (L, l). Since there is an equilibrium with outcome $(2, 0, 0)$ in which M is a best response, it is plausible that Player 2 and 3 could have conjectured the play M.

Put differently, if we hold onto the notion that players attach meaning to the "unexpected", a subgame cannot be treated as a game in its own right. The context of the larger game provides extra information on intended play in the subgame. Therefore, a forward induction solution of a game would not, on the ground of forward induction alone, induce the same

solution in any subgame; nor would any solution of a subgame necessarily be part of the solution of a game. Of course, one could resolve this tension between forward and backward induction by imposing stronger conditions on the solution concept. Yet nothing intrinsic in forward induction requires backward induction.

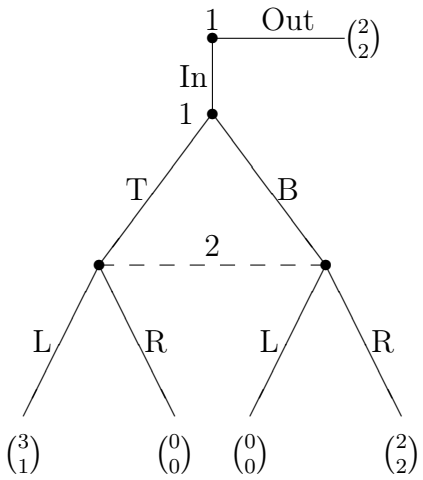
It is interesting to note, however, that every generic game tree has a forward induction outcome that is an admissible invariant backward induction outcome. Since every fully stable set of a tree contains a trembling hand perfect equilibrium of the tree (Kohlberg and Mertens, 1986, Proposition 3), a fully stable outcome is an admissible invariant backward induction outcome. By the fully stable version of Proposition 4.1 and the fact that every generic game tree has a fully stable outcome, every generic game tree has a forward induction outcome that is an admissible invariant backward induction outcome.

Govindan and Wilson (2009b, Theorem 6.1) prove that an invariant sequential equilibrium outcome of a generic[8] two-player game satisfies forward induction using their definition (c.f.: Definition 3.1). One may wonder if an analog of this theorem remains true under our definition. To be precise, we have the following conjecture:

Conjecture 4.3. *Let Γ be a two-person, finite extensive form game with perfect recall. Then there exists a full (Lebesgue) measure set of payoffs in $\mathbb{R}^{N\times|Z|}$ such that if the payoffs of Γ is in the set and G is the mixed strategy reduced normal form associated with Γ, then every invariant proper equilibrium outcome of G satisfies forward induction according to Definition 3.5.*

As we require admissibility, it is natural to strengthen the condition in the main theorem of Govindan and Wilson (2009b) from invariant sequential equilibrium to invariant quasi-perfect equilibrium. Since an invariant quasi-perfect equilibrium outcome is an invariant proper equilibrium outcome (see Mailath, Samuelson, and Swinkels (1997)), the conjecture is stated in terms of proper equilibrium to aid normal form analysis. Recall that proper

[8]Their generic requirement is stronger than ours. We require only that the game has finitely many Nash equilibrium outcomes. They impose an extra condition in addition to ours. See their Appendix A for a precise meaning of their generic games.



(a) Extensive Form

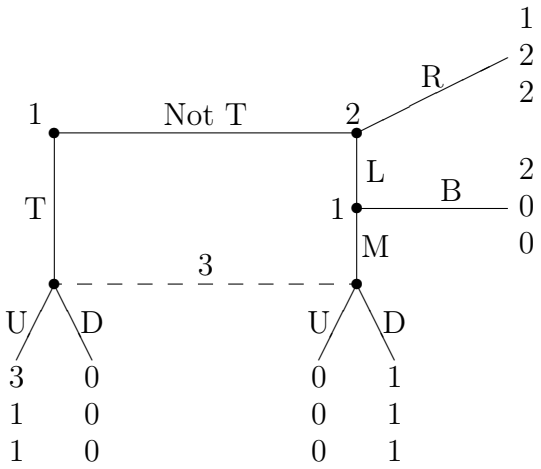
	L	R
Out	2,2	2,2
T	3,1	0,0
B	0,0	2,2

(b) Normal Form

FIGURE 4.3. A Two-Person Non-Generic Game

equilibrium is invariant with respect to the *pure strategy* reduced normal form game. Hence in the following discussion, it is without loss that we consider only the transformation of adding a mixed strategy as a pure strategy.

While the author is unable to prove Conjecture 4.3 or find a counter-example, all assumptions in the conjecture are indispensable. To see why genericity is necessary, consider Figure 4.3, an Outside Option Game with modified payoff. The payoff assignment is not generic in the sense that the following argument would not hold had we perturb the payoff to Player 1 after (B, R). The same argument as in the standard Outside Option Game shows that the outcome after Out does not satisfy forward induction (B is weakly dominated by Out). However, the outcome resulting from Out is an invariant proper equilibrium outcome. To see this, notice that (Out, R) is a proper equilibrium (the profile $((1, \varepsilon^2, \varepsilon), (\varepsilon, 1))$ is ε-proper). The only transformation that could potentially upset this proper equilibrium is the addition of a mixed strategy between T and Out (Player 2 has only two pure strategies; and adding other mixed strategies for Player 1 would not change the relative likelihood between T and B). So suppose we add a mixed strategy T^δ that plays T with probability δ and Out with probability $1 - \delta$. Unfortunately, B achieves the equilibrium payoff against R and performs strictly better than T. Thus for Player 2's strategies sufficiently close to R, B is also going to



(a) Extensive Form

	L	R
T	3,1,1	3,1,1
M	0,0,0	1,2,2
B	2,0,0	1,2,2

U

	L	R
T	0,0,0	0,0,0
M	1,1,1	1,2,2
B	2,0,0	1,2,2

D

(b) Normal Form

FIGURE 4.4

be a strictly better response than T^δ and T. Properness then requires B to be infinitely more likely than T^δ and T. To be precise, fix any $\delta \in (0,1)$ and denote Player 1's strategies in the game with the added mixed strategy in the form of $\left(\text{Out}, T, T^\delta, B\right)$. It could be verified that $\left(\left(1, \varepsilon^3, \varepsilon^2, \varepsilon\right), (\varepsilon, 1)\right)$ is an ε-proper equilibrium for $\varepsilon < \frac{2\delta}{2+3\delta}$. Conjecture 4.3 does not hold.

To see why two-person is necessary, consider Figure 4.4, a three-person game with perfect recall. Payoffs of this game can be perturbed to make it generic. The outcome $(1,2,2)$ does not satisfy forward induction: Player 1's strategy M is dominated by B and Player 2's strategy L is dominated by R. On the other hand, T is relevant for $(1,2,2)$ since it is a best response to the ε-perfect equilibrium $\langle\left(\varepsilon^2, \varepsilon, 1-\varepsilon-\varepsilon^2\right), (\varepsilon, 1-\varepsilon), \left(\frac{1}{3}, \frac{2}{3}\right)\rangle$. Yet when Player 2 plays R with sufficiently high probability and T is infinitely more likely than M, Player 3 plays U, causing Player 1 to deviate to T. Nevertheless, $(1,2,2)$ is an invariant proper equilibrium outcome. First note that $\langle\left(\varepsilon^2, \varepsilon, 1-\varepsilon-\varepsilon^2\right), (2\varepsilon, 1-2\varepsilon), (\varepsilon, 1-\varepsilon)\rangle$ is an

ε-proper equilibrium converging to (B, R, D) as $\varepsilon \to 0$. Adding mixed strategies for Player 2 or 3 is not going to upset this equilibrium since they both strictly prefer their equilibrium strategy along the sequence. The only potential way to upset this proper equilibrium is to add a mixture between T and B for Player 1. If the mixture could performs strictly better than M in the ε-proper equilibrium, then when $\sigma_2(L) \to 0$, Player 3 would strictly prefer U and upset any equilibrium with payoff $(1, 2, 2)$. Unfortunately this is impossible. Indeed, fix any $\delta \in (0, 1)$ and suppose a mixed strategy T^δ which puts probability δ on T and $1-\delta$ on B is added. Denote the strategies of Player 1 in the form of $\left(T, T^\delta, M, B\right)$. It could be verified that $\langle(\varepsilon^3, \varepsilon^2, \varepsilon, 1-\varepsilon-\varepsilon^2-\varepsilon^3), (2\varepsilon, 1-2\varepsilon), (\varepsilon, 1-\varepsilon)\rangle$ is ε-proper for $\varepsilon < \frac{-(\delta+2)+\sqrt{(\delta+2)^2+8\delta}}{4}$. Again Conjecture 4.3 does not hold.

Common to these examples is that an irrelevant strategy is achieving the equilibrium payoff while a relevant strategy is suboptimal in a proper equilibrium. As a result, any mixed strategy putting positive probability on the relevant strategy performs strictly worse than the irrelevant strategy at profiles sufficiently close to that proper equilibrium. This makes it possible for other players to believe that the irrelevant strategy is infinitely more likely than the relevant strategy at the proper equilibrium even after the transformations. More generally, adding a mixed strategy as a pure strategy is ineffective in changing the relative likelihood between a relevant and an irrelevant strategy when the payoff of the latter converges to the equilibrium payoff at a faster rate than the former. If both strategies reach an information set of an opponent player, the opponent may put positive probability on the node that is reached by the irrelevant strategy in a quasi-perfect equilibrium. It is unclear to the author how restricting attention to two-person and generic games could circumvent this difficulty in general.

4.3. **Iterative Dominance.** A forward induction equilibrium of a game G may fail to be one after deleting a dominated strategy. Consider the game in Figure 4.5. In this example, (T, L) is a forward induction equilibrium since $R_1^k((3,3)) = \{T\}$, $R_2^k((3,3)) = \{L, C\}$ for all $k \geq 1$ and the sequence $\langle(1-\varepsilon-\varepsilon^2, \varepsilon, \varepsilon^2), (1-\varepsilon-\varepsilon^2, \varepsilon, \varepsilon^2)\rangle$ justifies (T, L). However, if we

	L	C	R
T	3,3	2,3	1,0
M	0,2	0,1	0,0
B	0,1	0,2	4,4

FIGURE 4.5

delete the strictly dominated strategy M, (T, L) is no longer a forward induction equilibrium as L will then be weakly dominated by C. This example also shows a forward induction equilibrium may fail to survive iterative elimination of *strictly* dominated strategies, or an iterative elimination process that deletes *all* dominated strategies for all players at each round.

Even worse, a forward induction equilibrium could fail to survive elimination of strategies dominated at the equilibrium. Both M and B are irrelevant strategies for the equilibrium (T, L). Yet if we eliminate only M but not B, (T, L) is not a forward induction equilibrium after the elimination. Thus, forward induction equilibrium is weaker than Stable Set (Kohlberg and Mertens, 1986) in terms of iterative dominance. A stable set of a game contains a stable set of any game obtained by deletion of a dominated strategy or a strategy that is an inferior response in all the equilibria in the set (Kohlberg and Mertens, 1986, Proposition 6), regardless of the order of deletion. On the other hand, a forward induction equilibrium can only satisfy the following order-dependent version of equilibrium dominance:

Proposition 4.4. *A forward induction equilibrium σ of a normal form game G remains a forward induction equilibrium in the game obtained from G after deleting all strategies $s_i \notin R_i(g(\sigma))$ for all players $i \in N$.*

Proof. Let $G^* = (R_i(g(\sigma)), v_i^*)$ be the game obtained from deleting all irrelevant strategies for $g(\sigma)$ from G, where v_i^* is the utility function v_i restricted to the domain $\times_{i \in N} R_i(g(\sigma))$. We claim that if σ is a forward induction equilibrium of G, it is a normal form perfect equilibrium in G^*. Since σ is a forward induction equilibrium of G, there exists a sequence of ε-perfect equilibria $\{\sigma^n\}$ converging to σ such that, for all i, for all $s_i \in R_i(g(\sigma))$, all

$s_i' \notin R_i(g(\sigma))$,

$$\frac{\sigma_i^n(s_i')}{\sigma_i^n(s_i)} \to 0.$$

For each i, project σ_i^n onto $\Delta(R_i(g(\sigma)))$ and rescale the weights by a constant (so that they add up to one) to obtain a sequence of completely mixed strategies $\{\sigma_i^{*n}\}$ in $\Delta(R_i(g(\sigma)))$. Take $\sigma^{*n} = (\sigma_1^{*n}, \ldots, \sigma_N^{*n})$. We claim that this is a sequence of ε-perfect equilibria converging to σ in G^*. For if not, then there exists an $i \in N$, a strategy $s_i \in \text{supp}\,\sigma_i$, a strategy $s_i' \in R_i(g(\sigma))$ such that

$$v_i^*\left(s_i', \sigma_{-i}^{*n}\right) > v_i^*\left(s_i, \sigma_{-i}^{*n}\right)$$

for all n sufficiently large. Then for n sufficiently large it must also be the case that

$$v_i\left(s_i', \sigma_{-i}^{n}\right) > v_i\left(s_i, \sigma_{-i}^{n}\right),$$

contradicting the fact that $\{\sigma^n\}$ is a sequence of ε-perfect equilibria converging to σ in G.

Now the set of forward induction equilibria $F(g(\sigma))$ of G is contained in the set of normal form perfect equilibria of G^*. By definition, any strategy in $R_i(g(\sigma))$ is a best response to some justifying sequence of some equilibrium in $F(g(\sigma))$ so they are all first-order relevant for the outcome $g(\sigma)$ in G^*. Therefore all normal form perfect equilibria of G^* with outcome $g(\sigma)$, hence all forward induction equilibria with outcome $g(\sigma)$ in G are forward induction equilibria of G^*. □

The fact that forward induction equilibrium fails to survive iterative elimination of dominated strategies illuminates the nature of the latter procedure. In particular, there is a fundamental difference between an eliminated strategy and a strategy getting zero probability in equilibrium. The former cannot be used in any sense. The latter remains a threat as a potential deviation. This threat, even if it is never carried out *in equilibrium*, changes the strategic interaction. The difference between an eliminated strategy and a probability zero strategy is particularly pronounced when we require admissibility. A best response to a completely mixed strategy profile takes the latter but not the former into account. Since

	L	R
T	2,2	1,2
M	0,1	0,0
B	0,0	0,1

FIGURE 4.6

we are considering threats to a candidate equilibrium and are requiring admissibility, our forward induction equilibrium would be sensitive to this difference[9].

Even if we accept the procedure of eliminating a dominated strategy, the usual motivation for such a procedure may not justify eliminating strategies in *any* order. Consider the game in Figure 4.6, the standard example for demonstrating the order-dependence of iteratively eliminating weakly dominated strategies. If we eliminate M but not B, then L is weakly dominated so the only surviving outcome is $(1, 2)$. On the other hand, if we eliminate B but not M, then R is weakly dominated so the only outcome is $(2, 2)$. But is the procedure of "eliminating just M (or B)" well-justified? If M is eliminated based on the argument that a rational Player 1 would not choose a strictly dominated strategy, then B is subjected to the same reasoning and should be eliminated as well. Eliminating only M and arguing that R is a better choice for Player 2 because "in the unlikely event that Player 1 chooses B Player 2 does better by choosing R" is inconsistent — the exact argument for eliminating M means Player 2 would never choose B! Eliminating only M is like eliminating both M and B and then adding B back. Yet we certainly do not consider adding a dominated strategy a well-founded procedure (see Kohlberg and Mertens (1986, p. 1017) for a discussion). Indeed, if we eliminate both M and B on the grounds that they are both strictly dominated, both $(2, 2)$ and $(1, 2)$ survive.

The same argument should extend to the case of iteratively eliminating strategies dominated at a certain set of equilibria. In the game depicted in Figure 4.5, if M is eliminated based on the reason that Player 1 knows the outcome is $(3, 3)$ and M is not a best response to any equilibrium giving this outcome, on the same grounds should B be eliminated. If

[9]Apparently, other forward induction motivated concepts are also sensitive to this difference. McLennan (1985, p. 893) and Cho and Kreps (1987, p. 207-208) note the same issue; the former regarding justifiable equilibrium, the latter regarding D2 and Never a Weak Best Response.

both M and B (and to be consistent, R) are eliminated, (T, L) remains a normal form perfect and forward induction equilibrium of the resulting game. Forward induction allows for eliminating *all* irrelevant strategies for a certain outcome (as a rational player knowing the outcome would not choose them), rather than an arbitrary subset of them. Because of this, we feel it is inappropriate to equate forward induction with iterative elimination of dominated strategies.

Understandably, one might wish to argue that the strategies M and B in Figure 4.5 are not born equal — M is strictly dominated while B is not. Implicit in this argument is that (strictly) dominated strategies are further down the hierarchies of "rational choices" than strategies dominated only at a set of equilibria. We defer the discussion of such hierarchies to Section 5.2 when we compare forward induction equilibrium and explicable equilibrium.

5. Comparison with Other Forward Induction Concepts

5.1. **Signaling Games.** Forward induction is often applied to signaling games to select "reasonable" (often separating) equilibria. In this subsection we compare our general definition with some forward induction concepts in signaling games. A signaling game is a two-person extensive form game with perfect recall. Player 1 is the "sender" and Player 2 is the "receiver" of a "signal" or "message". Nature moves first, selects with a commonly known prior a type t of Player 1 from a finite type set T and reveals it to Player 1. Player 1, upon observing her type, chooses a "message" m from a finite message space M. Player 2 observes the message sent by Player 1 but not her type. Contingent on the message received, he then chooses an action a from a finite action set A.[10] A terminal node of the game is given by a type-message-action triple, (t, m, a). The utility function of Player i, $i = 1, 2$, is given by $u_i : T \times M \times A \to \mathbb{R}$. We restrict our attention to signaling games in which messages are "costly" for Player 1. This rules out cheap talk games (Crawford and Sobel, 1982) in which Player 1's utility is constant in message.

[10]It is possible to allow the message space M to depend on the type, and the action set to depend on the message. See Cho and Kreps (1987).

A pure strategy of Player 1 is a signaling function $s : T \to M$. A behavioral strategy of Player 1 is given by $\sigma : T \to \Delta(M)$ A pure strategy of Player 2 is a response function $r : M \to A$. Player 2's behavioral strategy is given by $\rho : M \to \Delta(A)$. Player 2's belief is given by a posterior probability μ over the type space T. Using the notations in Cho and Kreps (1987), write the set of best responses for Player 2 after observing m when his belief is μ as

$$BR(\mu, m) = \arg\max_{a \in A} \sum_{t \in T} u_2(t, m, a)\, \mu(t \mid m).$$

If $T' \subseteq T$, let $BR(T', m)$ denote the set of best responses by Player 2 to posteriors concentrated on the set T'. Formally,

$$BR(T', m) = \bigcup_{\mu : \mu(T' \mid m) = 1} BR(\mu, m).$$

And let $MBR(\mu, m)$ and $MBR(T', m)$ be the mixed best responses by Player 2 to belief μ and beliefs concentrated on the subset T' respectively.

In signaling games, a sequential equilibrium reduces to an assessment $\langle (\sigma, \rho)\, \mu \rangle$ such that: (1) For each type t, Player 1 maximizes her expected utility given the response of Player 2, that is,

$$\text{supp } \sigma(t) \subseteq \arg\max_{m \in M} \sum_{a \in A} u_1(t, m, a)\, \rho(a \mid m) \text{ for all } t \in T;$$

(2) For each message m, Player 2 maximizes his expected utility given μ:

$$\text{supp } \rho(m) \subseteq BR(\mu, m);$$

and (3) Player 2's belief is updated using Bayes' rule whenever possible.

Since both players move only once along every path in a signaling game, sequential equilibrium coincides with weakly sequential equilibrium. In addition, we are going to assume that the payoffs are generic so that weakly sequential equilibrium and weakly quasi-perfect (hence normal form perfect) equilibrium are equivalent. Due to this simplification, we are going to use Definition 3.1 without worrying about the caveats mentioned earlier.

Signaling games are typically plagued by multiple equilibria, many of which are supported by "unreasonable" off-equilibrium beliefs. There are forward induction motivated refinement concepts constructed particularly for signaling games. We review some of them here.

Intuitive Criterion: (Cho and Kreps, 1987) Take a sequential equilibrium outcome and let $u_1^*(t)$ be the payoff of a type t Player 1 in this equilibrium. For each out of equilibrium message m, form the set

$$T(m) = \left\{ t \in T : u_1^*(t) > \max_{a \in BR(T,m)} u_1(t, m, a) \right\}.$$

If, for any out of equilibrium message m there exists a type $t' \in T$ such that

$$u_1^*(t') < \min_{a \in BR(T \setminus S(m), m)} u_1(t'm, a),$$

then the equilibrium outcome fails the Intuitive Criterion.

Criterion D1: (Banks and Sobel, 1987) Fix a sequential equilirbium outcome and again denote the equilibrium payoff to type t of Player 1 as $u_1^*(t)$. For an out of equilibrium message m and for each type t, define the following two sets:

$$\begin{aligned} D_t &= \left\{ \alpha \in MBR(T, m) : u_1^*(t) < \sum_a u_1(t, m, r)\alpha(a) \right\} \\ D_t^0 &= \left\{ \alpha \in MBR(T, m) : u_1^*(t) = \sum_a u_1(t, m, r)\alpha(a) \right\} \end{aligned}$$

A type-message pair (t, m) could be pruned according to D1 if there exists a type $t' \neq t$ such that

$$D_t \cup D_t^0 \subseteq D_{t'}.$$

Criterion D2: (Banks and Sobel, 1987) Fix a sequential equilibrium outcome and define D_t and D_t^0 as above. A type-message pair (t, m) could be pruned according to D2 if

$$D_t \cup D_t^0 \subseteq \bigcup_{t' \neq t} D_{t'}.$$

Never a Weak Best Response: (Cho and Kreps, 1987) Again define D_t and D_t^0 as above. A type-message pair (t, m) could be pruned if

$$D_t^0 \subseteq \bigcup_{t' \neq t} D_{t'}.$$

It turns out that our forward induction concept implies these criteria:

Proposition 5.1. *Consider a generic signaling game* Γ *. Then under Definition 3.1,*

(1) If an equilibrium outcome of Γ *fails the Intuitive Criterion, it does not satisfy forward induction.*

(2) If, given an equilibrium outcome, a type-message pair (t, m) *can be pruned under Criteria D1, D2 or Never a Weak Best Response, then any pure strategy* s *such that* $s(t) = m$ *is irrelevant for that outcome.*

Proof. For the Intuitive Criterion: Given an equilibrium outcome, if a type t is in $T(m)$, then any pure strategy s such that $s(t) = m$ is irrelevant for that outcome since the strategy obtained from s by replacing $s(t)$ with the equilibrium message sent by type t performs strictly better than s in any (weakly) sequential equilibrium. Since there exists a type $t' \in T \setminus T(m)$ that could potentially achieve a payoff strictly higher than her equilibrium one by sending the message m, the information set in which Player 2 observes m, $h(m)$, is relevant for the outcome. However, since

$$u_1^*(t') < \min_{a \in BR(T \setminus T(m), m)} u_1(t', m, a),$$

there does not exist any sequential equilibrium giving the candidate equilibrium outcome when beliefs at the information set $h(m)$ puts zero probability on the set $T(m)$.

For D1, D2 and Never a Weak Best Response: Notice that any signaling function s with $s(t) = m$ is a best reply to a sequential equilibrium $\langle (\sigma, \rho), \mu \rangle$ giving the fixed outcome only when $\rho(m) \in D_t^0$. However, since $D_t^0 \subseteq \cup_{t' \neq t} D_{t'}$, whenever this is case there is always

another type having a strictly profitable deviation. Hence s is not a best reply to any sequential equilibrium giving the candidate outcome. □

Therefore, in a signaling game, any equilibrium that can be ruled out by the Intuitive Criterion, D1, D2 and Never a Weak Best Response can be ruled out by forward induction. Thus forward induction rejects the "both-quiche" outcome in the Beer-Quiche game and picks the Riley outcome in a Spence signaling game with finitely many types (appropriately discretized to make it finite) (Cho and Kreps, 1987). Yet since Never a Weak Best Response is strictly stronger than D1 and D2 (for an example, see Figure IV of Cho and Kreps (1987)), and that our iterative definition of forward induction requires *all* relevant strategies being infinitely more likely than *all* irrelevant strategies, our iterative definition need not correspond to Divinity and Universal Divinity, the iterative version of D1 and D2 respectively (Banks and Sobel, 1987).

5.2. **Justifiable and Explicable Equilibrium.** The program of identifying strategies that cannot be adopted by rational players who might potentially be confused about which equilibrium is "in effect" and iteratively "pruning" such strategies can be traced back to McLennan (1985). Working with extensive form games, McLennan calls an action (at an information set) first-order *useless* if it is an inferior reply in all sequential equilibria in that extensive form game. One then selects among sequential equilibria such that the belief system assigns positive probability to the node reached with the smallest number of useless actions at each information set. Now we can inductively define higher order useless actions as those that are inferior responses in all sequential equilibria remained after the last round of selection. Then select among the remaining equilibria by putting a lexicographic "uselessness condition" on the belief system. The resulting solution concept is known as *Justifiable Equilibrium.*

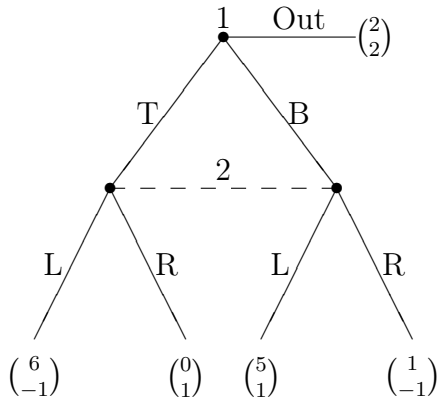
Since all justifiable equilibria are sequential while some forward induction equilibria are not, not all forward induction equilibria are justifiable. (Figure 4.2 provides some obvious examples.) Conversely, there are justifiable equilibria that do not satisfy forward induction. In our first presentation of the Outside Option Game (Figure 1.1), since $((\text{Out}, B), R)$ and

$((\text{In}, T), L)$ are both sequential (actually, trembling hand perfect) in this representation, none of the actions are useless. So the equilibrium $((\text{Out}, B), R)$ is justifiable. Yet it does not satisfy forward induction.

Another closely related concept is Explicable Equilibrium by Reny (1992). Again working on extensive form games, Reny takes a set of behavioral strategy profiles, B, which contains the set of all behavioral profiles from all weakly sequential equilibria. A behavioral strategy b_i of player i is *first-order best response relative to* B if there is a profile b in the convex hull of B, a belief system μ derived from b using Bayes' rule whenever possible, such that b_i constitutes an optimal continuation at every information set $h_i \in H_i$ enabled by b_i. A *first-order explicable equilibrium* is a consistent assessment such that, if a pure strategy s_i is first-order best response relative to B and s_i' is not, then the limiting probability (defined by the sequence giving the consistent assessment) on the set of all realization equivalent strategies of s_i' relatively to that of s_i is zero. One can iterate this procedure taking the behavioral strategies from all first-order explicable equilibrium to be the new choice of B. The resulting equilibrium concept is known as *Explicable Equilibrium*.

It would not be instructive to focus on the difference between explicable equilibrium and forward induction equilibrium arising from the difference between weakly sequential and normal form perfect equilibrium. To aid a meaningful comparison, we modify Reny's definition by replacing "weakly sequential equilibrium" with "weakly quasi-perfect equilibrium". In addition, we require a best response relative to B, b_i, to constitute an optimal continuation at every enabled information set against a sequence of completely mixed behavioral strategies converging to some b in the convex hull of B. This strengthened version of explicable equilibrium will be adopted throughout this subsection.

A key difference between explicable and forward induction equilibrium is that, when determining the "possible explanations" for deviations, the former allows for best responses with respect to *any* weakly quasi-perfect equilibrium (or equivalently, normal form perfect) whereas the latter allows only for best responses with respect to normal form perfect equilibria *with the same outcome*. Explicable equilibrium would be better motivated if we feel



(a) Extensive Form

	L	R
Out	2, 2	2, 2
T	6,-1	0, 1
B	5, 1	1,-1

(b) Normal Form

FIGURE 5.1. Reny's Example

the deviators do not know the outcome in addition to being confused about which weakly sequential equilibrium is in effect. Forward induction equilibrium, on the other hand, asserts that deviators know the outcome but could be confused about which normal form perfect equilibrium giving that outcome is in effect.

Due to this difference, it is not surprising that some explicable equilibria supported by beliefs that a deviator is best responding to a weakly quasi-perfect equilibrium with *another* outcome do not satisfy forward induction. Figure 5.1, a game taken from Reny (1992, Figure 4), provides one such example. Let μ be the belief Player 2 puts on the node after T in his information set. There is a weakly quasi-perfect equilibrium $\langle \left(0, \frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{2}, \frac{1}{2}\right), \mu = \frac{1}{2} \rangle$ giving expected payoffs $(3, 0)$. There is also a connected component which consists of weakly quasi-perfect equilibria of the form $\langle (\text{Out}, r), \mu \rangle$ with $\mu \geq \frac{1}{2}$ and those with $\langle (\text{Out}, (\alpha, 1-\alpha)), \frac{1}{2} \rangle$ where $\alpha \leq \frac{1}{4}$. All equilibria in this component give payoffs $(2, 2)$. Since all strategies are equilibrium strategies in some weakly quasi-perfect equilibrium, all strategies are first-order best responses (relative to any set of strategy profiles containing all weakly quasi-perfect equilibrium profiles). Hence all equilibria in the connected component with payoffs $(2, 2)$ are explicable. However, if we restrict attention to this connected component, only Out and B are relevant for the outcome $(2, 2)$: Out is an equilibrium strategy and B is a best response to the sequence of assessments $\langle (1 - 2\varepsilon, \varepsilon, \varepsilon), \left(\frac{1}{4}, \frac{3}{4}\right), \frac{1}{2} \rangle$. T is not relevant for the outcome

$(2,2)$: For T to achieve an expected payoff of 2, Player 2 needs to put probability at least $\frac{1}{3}$ on L. Yet in such cases, B yields a payoff strictly higher than 2. Thus in any forward induction equilibrium giving $(2,2)$ Player 2 needs to put probability 1 on B. But then L is a best response for 2 and Player 1 would deviate from Out. Therefore $(2,2)$ is not a forward induction outcome and none of the equilibria in the connected component satisfies forward induction.

How about the converse? Is it true that all forward induction equilibria are explicable? Again the answer is no. Recall the game in Figure 4.5. We have argued that (T, L) is a forward induction equilibrium. Yet it is not explicable in the strengthened sense: (B, R) is a strict Nash equilibrium so B is a best response relative to any set of strategy profiles containing all normal form perfect equilibrium profiles. M is a strictly dominated strategy so it cannot be a best response relative to any strategy profile. However, if the relative probability of M to B is zero, then (T, L) can no longer be a normal form perfect equilibrium[11].

We have argued in Section 4.3 that as long as Player 2 believes Player 1 knows the outcome is $(3,3)$, we should treat M and B equally as they are both inferior at any equilibrium giving $(3,3)$. On the other hand, one may feel that a strictly dominated strategy should not be treated the same way as a strategy which is a best response in some equilibrium. Very weak notion of rationality suffices to prune a strictly dominated strategy while a much stronger common belief requirement is required for pruning a strategy that is a best response in some equilibrium. Explicable equilibrium respects part of this intuition, while forward induction equilibrium asserts strongly that the equilibrium outcome is a common belief.

In general, one could come up with a "hierarchy of rational choices" and require that strategies down the hierarchy should be infinitely less likely than those higher up. But what is a "reasonable" ordering of strategies? Should the ranking be "local", that is, should the ranking pertain to the particular equilibrium in consideration (e.g.: proper equilibrium orders strategies by their costs at the equilibrium); or should it be "global" (e.g.: strictly

[11]Though it is still a weakly sequential equilibrium since it is a Nash equilibrium of a simultaneous move game.

dominated strategies are always infinitely less likely than undominated strategies)? It would be interesting to look for a well-founded ranking and construct an equilibrium concept in accordance with the hierarchy.

5.3. **Requirement by van Damme.** In an early paper on forward induction, van Damme (1989) proposes a property which "should be satisfied by any concept consistent with forward induction" (p. 485). This property requires that if in any generic two-person (extensive form) game Γ in which player i chooses between an outside option and entering a subgame of Γ which possess a unique equilibrium e^* (according to the concept) giving player i more than his outside option, then the concept should only admit equilibria of Γ in which player i enters the subgame and the equilibrium of the subgame is e^*. An example in which a stable outcome fails this property is given by van Damme (1989, p. 485-487, Figure 4). Since every constant outcome stable set contains a forward induction equilibrium (Proposition 4.1), our definition fails his requirement as well.

Does this mean we should rename our equilibrium concept? We feel the answer is no. Notice that van Damme's requirement asks for *backward induction* before forward induction — it demands a solution of a game to induce a solution in a subgame. This is explicit in the text: "forward induction can only determine which solution should be played in a subgame, but a solution of a game should always induce a solution in each subgame; *backwards induction ranks above forward induction*." (van Damme, 1989, p.485, emphasis added). As we have argued, how a subgame is reached conveys strategic information for players. To be consistent with forward induction, a subgame should not be treated as a game of its own right. The very fact that it is embedded in a larger game changes the strategic consideration players could have. Hence a solution concept consistent with forward induction may not respect backward induction. While we do not object to solution concepts capturing both backward and forward induction (and imposing a ranking between them), we disagree with the assertion that *any* concept consistent with forward induction should satisfy a property ranking backward induction above forward induction.

6. Conclusion

In this paper, we propose a definition of forward induction equilibrium and analyze its properties. However, there are aspects of forward induction we have not considered. In particular, we rule out the possibility that a player believes (correctly or incorrectly) that his opponents' strategies might be correlated. A strategy may be a best response to a correlated strategy profile of the opponents but fail to be relevant under our definition. Gul and Pearce (1996) suggest that forward induction reasoning loses its refinement power in stage games when public randomization is introduced. While Govindan and Robson (1998) point out that Gul and Pearce's argument relies on players using inadmissible strategies, it is true that expanding the set of relevant strategies by allowing correlation weakens a forward induction argument in general. Nevertheless, an extension of our definitions to games with correlation device is not trivial as they are often infinite games.

On a related note, we have insisted on the consistency of belief system (subsumed in the sequence of ε-perfect equilibria) in our definition. As consistency is implied by the independence of players' randomization and common beliefs (Kohlberg and Reny, 1997), we view it as the appropriate restriction on the belief system. However, a consistent assessment need not be structurally consistent, that is, it may not be possible for a player to find a single strategy profile as an alternative hypothesis for reaching his information sets (Kreps and Ramey, 1987). In the game depicted in Figure 1 of Kreps and Ramey (1987), there is no structurally consistent assessment supporting the unique forward induction outcome (which is also the unique Nash equilibrium outcome). While we do not view this as a flaw of our definition, one should be careful when interpreting the restrictions we put on off-equilibrium beliefs in our forward induction equilibrium.

References

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

Banks, J. S., and J. Sobel (1987): "Equilibrium Selection in Signaling Games," *Econometrica*, 55(3), 647–661.

Blume, L., A. Brandenburger, and E. Dekel (1991): "Lexicographic Probabilities and Equilibrium Refinements," *Econometrica*, 59(1), 81–98.

Cho, I.-K., and D. M. Kreps (1987): "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102(2), 179–221.

Crawford, V. P., and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50(6), 1431–1451.

Dalkey, N. (1953): "Equivalence of Information Patterns and Essentially Determinate Games," in *Contributions to the Theory of Games*, ed. by H. W. Kuhn, and A. W. Tucker, vol. 2, pp. 217–243, Princeton, New Jersey. Princeton University Press.

Elmes, S., and P. J. Reny (1994): "On the Strategic Equivalence of Extensive Form Games," *Journal of Economic Theory*, 62, 1–23.

Govindan, S., and A. J. Robson (1998): "Forward Induction, Public Randomization and Admissibility," *Journal of Economic Theory*, 82, 451–457.

Govindan, S., and R. Wilson (2009a): "Axiomatic Equilibrium Selection for Generic Two-Player Games," Working Paper, University of Iowa.

——— (2009b): "On Forward Induction," *Econometrica*, 77(1), 1–28.

Gul, F., and D. G. Pearce (1996): "Forward Induction and Public Randomization," *Journal of Economic Theory*, 70, 43–64.

Kohlberg, E., and J.-F. Mertens (1986): "On the Strategic Stability of Equilibria," *Econometrica*, 54(5), 1003–1038.

Kohlberg, E., and P. J. Reny (1997): "Independence on Relative Probability Spaces and Consistent Assessments in Game Trees," *Journal of Economic Theory*, 75, 280–313.

Kreps, D. M., and G. Ramey (1987): "Structural Consistency, Consistency, and Sequential Rationality," *Econometrica*, 55(6), 1331–1348.

Kreps, D. M., and R. Wilson (1982): "Sequential Equilibria," *Econometrica*, 50(4), 863–894.

Kuhn, H. W. (1953): "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games*, ed. by H. W. Kuhn, and A. W. Tucker, vol. 2, pp. 193–216, Princeton, New Jersey. Princeton University Press.

Mailath, G. J., L. Samuelson, and J. M. Swinkels (1993): "Extensive Form Reasoning in Normal Form Games," *Econometrica*, 61(2), 273–302.

——— (1997): "How Proper is Sequential Equilibrium?," *Games and Economic Behavior*, 18, 193–218.

McLennan, A. (1985): "Justifiable Beliefs in Sequential Equilibrium," *Econometrica*, 53(4), 889–904.

Pearce, D. G. (1984): "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52(4), 1029–1050.

Reny, P. J. (1992): "Backward Induction, Normal Form Perfection and Explicable Equilibria," *Econometrica*, 60(3), 627–649.

Rubinstein, A. (1985): "A Bargaining Model with Incomplete Information About Time Preferences," *Econometrica*, 53(5), 1151–1172.

Selten, R. (1975): "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4(1), 25–55.

Thompson, F. B. (1952): "Equivalence of Games in Extensive Form," Discussion Paper RM 759, The Rand Corporation.

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

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

Department of Economics, University of Chicago

E-mail address: ptyman@uchicago.edu