ON FORWARD INDUCTION

SRIHARI GOVINDAN AND ROBERT WILSON

ABSTRACT. We examine Hillas and Kohlberg's conjecture that invariance to the addition of payoff-redundant strategies implies that a backward induction outcome survives deletion of strategies that are inferior replies to all equilibria with the same outcome. Although it suffices in simple games to interpret backward induction as a subgame-perfect or sequential equilibrium, to obtain general theorems we use a quasi-perfect equilibrium, i.e. a sequential equilibrium in strategies that are admissible continuations from each information set. Using this version of backward induction, we prove the Hillas-Kohlberg conjecture for two-player extensive-form games with perfect recall. We also prove an analogous theorem for general games by using the property of a proper equilibrium that it is equivalent to a quasi-perfect equilibrium of every extensive form with the same normal form, provided beliefs are justified by perturbations invariant to inessential transformations of the extensive form. For a two-player game we prove that if a set of equilibria includes a proper equilibrium of every game with the same reduced normal form then it satisfies forward induction, i.e. it includes a proper equilibrium of the game after deleting strategies that are inferior replies to all equilibria in the set. We invoke slightly stronger versions of invariance and properness to handle nonlinearities in an N-player game.

1. INTRODUCTION

Our purpose is to address suggestions by Hillas [14] and Kohlberg [17] and the following summary observation by Hillas and Kohlberg [16] in their survey of equilibrium refinements:

> "... there appears to be a relationship between backward and forward induction. In many examples—in fact, in all of the examples we have examined—a combination of the invariances we have discussed and backward induction gives the results of forward induction arguments ... ." [16, §13.6]

Their examples and others in the literature are two-player games in extensive form with perfect recall that are typically either outside-option games or signaling games. The usual assumptions are that:

- Backward induction means that the outcome results from a subgame-perfect or sequential equilibrium.

- Invariance means survival of the backward induction outcome when payoff-redundant strategies are adjoined to the game.
- Forward induction means survival of the backward induction outcome after deletion of strategies that are inferior replies to all equilibria with that outcome.

In §2 we review the motivation for forward induction and reprise two standard examples using these same assumptions.

However, a sequential equilibrium is an insufficient representation of backward induction in games more complicated than the usual motivating examples. Here we obtain general theorems by using a quasi-perfect equilibrium. One can interpret van Damme's [6] definition of quasi-perfect equilibrium as the refinement of sequential equilibrium that requires each player's strategy to provide an admissible continuation from each information set. Using this version of backward induction, Theorem 3.5 verifies the Hillas-Kohlberg conjecture for two-player games in extensive form with perfect recall.

In §4 and §5 we develop formulations of backward and forward induction adapted to general games, including games in normal form. In §6 and §7 we prove analogs of the Hillas-Kohlberg conjecture for general two-player and N-player games.

The paper is divided into two parts that are largely independent and can be read separately. In §2 and §3 we focus on games in extensive form, and in §4 – §7 on games in normal form. A more stringent version of forward induction proposed by van Damme [8] is addressed briefly in §8.

2. Forward Induction in Extensive-Form Games

In this section we review in §2.1 and §2.2 the motivation for forward induction in extensive-form games. The main ideas are illustrated in §2.3 by two examples, one an outside-option game and the other a signaling game.

2.1. **Background.** Kohlberg and Mertens [18] introduce forward induction as a criterion for selecting among the Nash equilibria of a game. They do not provide an explicit definition, relying instead on motivating examples and the cryptic label to their theorem that:

> "(Forward Induction) A stable set contains a stable set of any game obtained by deletion of a strategy which is an inferior response in all the equilibria of the set." [18, Proposition 6]

This property—that a subset of a selected set of equilibria survives deletion of inferior strategies—is seen by subsequent authors as the crucial *test* for forward induction. However, the relevance of this test is not immediately obvious from the *motivation* for forward induction. The motivation is summarized by Hillas and Kohlberg [16]:

> "... a self-enforcing assessment of the game must not only be consistent with deductions based on the opponents' rational behavior in the future (backward induction) but it must also be consistent with deductions based on the opponents' rational behavior in the past (forward induction)." [16, §42.11]
>
> "Forward induction involves an assumption that players assume, even if they see something unexpected, that the other players chose rationally in the past." [16, §42.13.6]

In the next subsection we explain the motivation for forward induction in more detail and show how one is led to test for forward induction by considering the effects of deleting inferior strategies.

2.2. **Motivation for Forward Induction in a Generic Extensive-Form Game.** The literature includes no formal definition of forward induction; e.g. Hillas and Kohlberg [16, §42.11] say that, "A formal definition of forward induction has proved a little elusive."

Here we review the basic ideas in the context of a game in extensive form with perfect recall and generic payoffs, which includes the motivating examples in the literature. For such a game, all Nash equilibria in a connected set induce the same outcome, viz. the same probability distribution on terminal nodes of the game tree [10]. For simplicity in this subsection and the next, we assume that backward induction is satisfied by a sequential equilibrium of the extensive form, and by an outcome we shall mean an outcome resulting from some sequential equilibrium. That is, within the components of Nash equilibria with that outcome, some equilibria are sequential.[1]

Recall that a sequential equilibrium requires that, from each of his information sets, a player's strategy is an optimal continuation in reply to other players' strategies. Optimality is based on some consistent beliefs; i.e. for each information set, on conditional probabilities of the histories that reach it, even for those information sets not reached with positive probability by the equilibrium strategies. The sequential equilibria with the same outcome differ only in their beliefs and behaviors at information sets not reached by equilibrium play.

---

[1]It is often the case that some non-sequential equilibria in such a component become sequential if sufficient payoff-redundant strategies are adjoined to the extensive form—see the example in [12, §2.3]. However, to address the Hillas-Kohlberg conjecture without confusing the issue it is preferable not to invoke invariance to payoff-redundant strategies in the motivation or definition of forward induction. For the same reason we defer consideration of van Damme's [8] interpretation of forward induction to §8.

The motive for forward induction is to enforce some discipline on beliefs and hence behaviors at unreached information sets, and thereby to select among the equilibrium outcomes. The examples in the literature suggest two ways.

(1) Along the boundary of the set of equilibria inducing a given outcome there is often some player who is indifferent between his equilibrium strategy and a particular deviation. That is, for some equilibrium in the set the deviation is an optimal reply. One can therefore require that, at another player's information set that might have been reached due to this deviation, his belief should assign a greater likelihood to this deviation than to errors with no rational explanation—and therefore his continuation strategy should be an optimal reply to this belief. Note that this approach stems from consideration of *sets* of equilibria, a perspective that Kohlberg and Mertens [18] and Hillas and Kohlberg [16] emphasize is intrinsic to forward induction. See Example 1 in §2.3 for an illustration.

(2) A more pragmatic perspective argues that other players' beliefs should allow interpretation of a player's actions as credible signals of private information or future intentions. That is, a player attempting to signal should not be stymied by others' beliefs that are blind to the implications of observed actions. For instance, Kohlberg and Mertens [18, p. 1013] assert that, "a subgame should not be treated as a separate game, because it was preceded by a very specific form of preplay communication—the play leading to the subgame." Based on the motivating examples of sender-receiver signaling games, the typical source of an intransigent belief at an information set off the path of equilibrium play is the receiver's insistence on ascribing positive conditional probability to strategies of the sender that are dominated or otherwise inferior when other strategies are more likely sources of deviant behavior by the sender. Again the suggested discipline is to require beliefs to recognize the possible rational explanations for deviations. A typical application of this approach to signaling games is Cho and Kreps' [4] 'equilibrium dominance' and 'intuitive' criteria, which restrict the support of the receiver's belief to those types of the signaler who might benefit from deviating if they anticipate the receiver's optimal reply to this belief. See Example 2 in §2.3 for an illustration.

Both (1) and (2) suggest a minimal test, the one used by Kohlberg and Mertens and again by Hillas and Kohlberg in their analyses of examples. If some equilibrium with the given outcome is sensitive to the presence of a deviant action, in the sense that its incentive compatibility constraint is binding, then possibly there is a rational explanation for the deviation. Conversely, if no rational explanation is possible then presumably this constraint is

nowhere binding. Therefore, the same outcome should survive when that action is excluded. By 'excluded' one can mean restricting the action's probability to be zero in the belief of any other player or equivalently, as we do here, deleting the pure strategies that use that action.

One can interpret this test as implementing the following definition of forward induction in the context of this section.

Definition 2.1 (Test for Forward Induction)**.** An outcome satisfies the test for forward induction if it remains an equilibrium outcome after deleting actions that are inferior replies to every equilibrium with that outcome.

Here, an action at an information set is an inferior reply if every strategy that does not exclude the information set and that chooses the action is an inferior reply in every equilibrium having the specified outcome; and deleting an action entails deleting all paths that follow it in the game tree.

If an outcome of a sequential equilibrium passes this test for forward induction then it can be supported with beliefs that assign zero probability to inferior replies to all equilibria with that outcome, and conversely. This test is consistent in major respects with the applications of forward induction to outside-option games by van Damme [7] and Hauk and Hurkens [13]; weaker variants of forward induction to signaling games by Banks and Sobel [1], Cho and Kreps [4], and Cho and Sobel [5]; and others reviewed in surveys by Fudenberg and Tirole [9, §11], Hillas and Kohlberg [16], and Kreps and Sobel [19].

2.3. **Examples.** In this subsection we use two standard examples to illustrate how the test for forward induction can reject some equilibria. We also use these examples to illustrate that one obtains the same result when backward induction is complemented by invariance to the addition of payoff-redundant strategies, which anticipates Theorem 3.5 below.

Example 1 — An Outside-Option Game. The top panel of Figure 1 displays the extensive and normal forms of a two-player game consisting of a subgame with simultaneous moves that is preceded by an outside option initially available to player I. As in case (1) the component of equilibria in which player I chooses his outside option includes an equilibrium in which player II's strategy has probability 2/3 of his left column and therefore player I is indifferent about deviating to his top row in the subgame, whereas there is no such equilibrium justifying deviating to the bottom row. Or as in case (2) player I might anticipate that player II will recognize rejection of the outside option as a signal that player I will choose the top row and therefore II should respond with the left column. To apply the test for forward induction one deletes the inferior strategy in which I's rejection of the outside option is followed by his choosing the bottom row in the subgame. In fact, this component

Extensive Form   Strategic Form

II

3,1	0,0
0,0	1,3

2,2	2 2
3,1	0,0
0,0	1,3

II

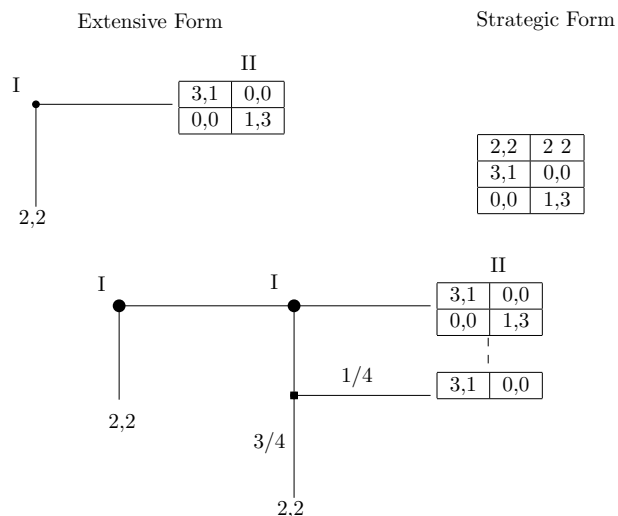|	
3,1	0,0
0,0	1,3

3,1	0,0

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

fails the test, since in the pruned subgame player II's dominant strategy is to play left, and anticipating this, player I rejects the outside option.

As in Hillas [14, Figure 2], one can invoke invariance and backward induction to obtain this conclusion. The bottom panel of Figure 1 shows the extensive form after adjoining a redundant strategy in which, after tentatively rejecting the outside option, player I randomizes between the outside option and the top row of the subgame with probabilities 3/4 and 1/4. Player II does not observe which strategy of player I led to rejection of the outside option. In the unique subgame-perfect equilibrium of this equivalent game player I rejects both the outside option and the randomization and then chooses the top row of the final subgame.

Example 2 — A Signaling Game. The top panel of Figure 2 displays the two-player game Beer-Quiche studied by Cho and Kreps [4] and discussed further by Kohlberg and Mertens [18, §3.6.B] and Fudenberg and Tirole [9, §11.2]. Consider the component of sequential equilibria with the outcome Q-R; that is, both types W and S of player I choose Q and player II responds with R. The equilibria in this component are sustained by player II's belief after observing B that I's type W was as likely to have deviated as type S. In all these equilibria, B is an inferior choice for type W. But in the equilibrium for which player II assigns equal probabilities to W and S after observing B and mixes equally between F and R, type S is indifferent between Q and B, as in case (1) above, and if II recognizes this as the source of I's deviation then he will infer after observing B that I's type is S and therefore choose R. Alternatively, as in case (2), if player I's type is S then he might deviate to B in hopes that this action will credibly signal his type, since his equilibrium payoff is 2 from Q but he obtains 3 from player II's optimal reply R if the signal is recognized, but type W has
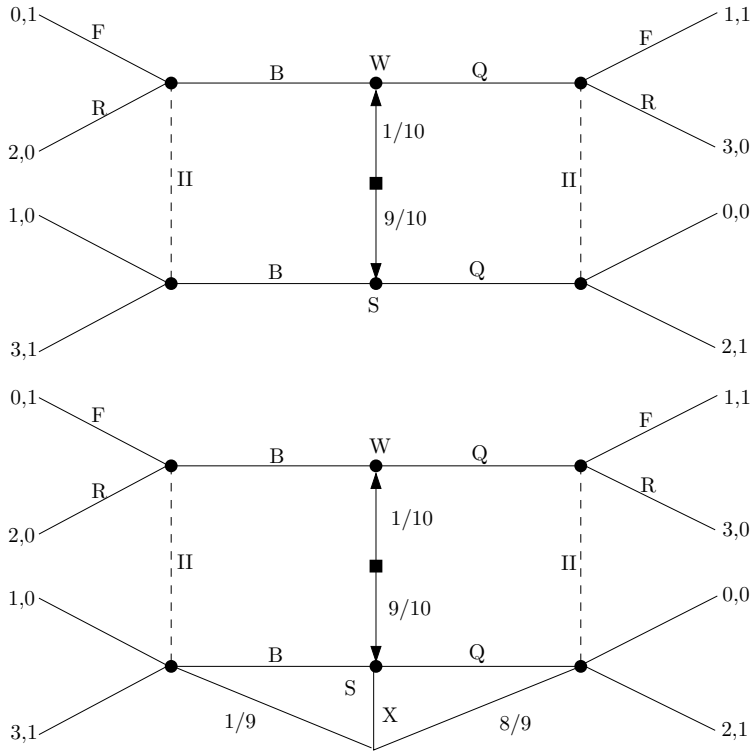
FIGURE 2. Two versions of the Beer-Quiche game

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

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

no comparable incentive to deviate. One can therefore apply the test for forward induction by considering the 'pruned' game obtained by deleting player I's action B when his type is W, or in the normal form by deleting player I's pure strategies that choose B when his type is W. In fact, the sequential equilibria that choose Q do not survive in the pruned game, since player II's optimal response to B is then R, which makes it advantageous for player I's type S to deviate by choosing B. Thus the sequential equilibria with the outcome Q-R do not satisfy the test for forward induction. This leaves only the component of sequential equilibria with the outcome B-R in which both types of player I choose B.

As in Example 1, one can obtain this same conclusion by invoking invariance and backward induction. The bottom panel of Figure 2 shows the extensive form after adjoining a redundant action X for type S of player I that produces a randomization between B and Q with probabilities 1/9 and 8/9. Note that X is an optimal action for I's type S at some equilibria in the component. Denote by BQ player I's pure strategy that chooses B if his type is W and chooses Q if his type is S, and similarly for his other pure strategies. The normal form of this expanded game is shown in Table 1 with all payoffs multiplied by 10. Now consider the following extensive form that has the same normal form. Player I initially chooses whether or not to use his pure strategy QQ, and if not then subsequently he chooses among his other pure strategies BB, BQ, QB, and QX. After each of these five pure strategies, the extensive form in the bottom panel ensues, but with I's action dictated by his prior choice of a strategy. That is, nature chooses I's type to be W or S, the selected pure strategy dictates the subsequent choice of B or Q, and then player II (still having observed only which one of B or Q was chosen) chooses F or R. At player I's information set where, after rejecting QQ, he chooses among his other pure strategies, a sequential equilibrium requires that he assigns zero probability to BQ, since it is strictly dominated by QX in the continuation. At player II's information set after observing B, a sequential equilibrium requires that his behavioral strategy is an optimal reply to some consistent belief about those strategies and types of player I that reach this information set. But every mixture of I's pure strategies BB, QB, and QX implies that, given his choice of B, the conditional probability that his type is S exceeds 9/10. Therefore, player II's reply to B must be R in every sequential equilibrium of this extensive form. Hence the component with outcome Q-R is inconsistent with invariance and backward induction, in agreement with its failure to satisfy the test for forward induction.

In Examples 1 and 2 it is sufficient to interpret backward induction as requiring a sequential equilibrium. However, more complicated examples show that this interpretation is insufficient in general extensive-form games, such as those in which the players alternative moves repeatedly along a single path of play.

For the remainder of the paper we interpret backward induction as requiring a quasi-perfect equilibrium. By enforcing admissibility of continuation strategies at each information set, a quasi-perfect equilibrium enforces a more stringent version of rationality than does a sequential equilibrium. Most relevant for forward induction, however, is that quasi-perfection ensures that the beliefs that support a sequential equilibrium "respect preferences" as defined by Blume, Brandenberger, and Dekel [2], as we elaborate in §4.2.

3. A Version of the Hillas-Kohlberg Conjecture for Two-Player Games

In this section we prove a version of the Hillas-Kohlberg conjecture for any two-player game in extensive form with perfect recall. In this context, by an outcome we mean a probability distribution on the terminal nodes of the game tree induced by the players' strategies in some equilibrium. Because the game has perfect recall, Kuhn's theorem implies that the possible outcomes from mixed and behavioral strategies are the same.

3.1. **Quasi-Perfection.**

As mentioned, by a backward induction outcome we mean one induced by a quasi-perfect equilibrium. Van Damme [6] defines a quasi-perfect equilibrium as a sequential equilibrium that satisfies a weak version of conditional admissibility of a player's continuation strategy from each information set; viz., at each of his information sets a player's action must initiate a continuation strategy that is optimal against a shrinking sequence of perturbations of other players' strategies.

Definition 3.1 (Quasi-Perfect). A quasi-perfect equilibrium is a limit point of a sequence of ε-quasi-perfect profiles as $\varepsilon \downarrow 0$, where a profile b^ε of completely mixed behavioral strategies is ε-quasi-perfect if at each information set of a player n the probability of choosing an action exceeds ε only if there is an optimal continuation strategy in reply to $(b_m^\varepsilon)_{m \neq n}$ that chooses that action.

3.2. **Formulation.**

Next we establish the formulation used for Theorem 3.5 below.

For the remainder of this section, let Γ be a two-player game in extensive form with perfect recall. Our notation for a generic player is n and we use m for his opponent. For each player n let S_n, Σ_n, and B_n be his sets of pure, mixed, and behavioral strategies, respectively, and let S, Σ, and B be the corresponding product sets of profiles of players' strategies. If a strategy for player n excludes one of his information sets, his choice at that information set is irrelevant. Therefore, the normal form of Γ we adopt in this section is the simplified normal form obtained from the full normal form by treating two pure strategies of a player as equivalent if they exclude the same information sets and prescribe the same choices at information sets they do not exclude. For each n, each pure strategy in S_n is thus an equivalence class of his pure strategies.[2]

Since we do not exclude the possibility that the payoffs in Γ are non-generic, in Definition 3.2 below we use a definition of forward induction that applies to sets of equilibrium

---

[2] The simplified normal form need not be the same as the reduced normal form in which one also deletes payoff-redundant pure strategies when the payoffs are nongeneric. We use the simplified normal form only to simplify exposition in this section. Our results remain valid if one uses the full normal form of Γ.

outcomes—not just one outcome.[3] Therefore, let P be a closed set of Nash equilibrium outcomes and let $B(P)$ and $\Sigma(P)$ be the sets of equilibria in behavioral and mixed strategies that induce outcomes in P.

Let U_n be the collection of player n's information sets. For each information set $u_n \in U_n$ let $A_n(u_n)$ be the set of n's available actions at u_n. An action $a_n \in A_n(u_n)$ is P-inferior (or just *inferior*) if every pure strategy of player n that does not exclude u_n and that chooses a_n is not an optimal reply to any equilibrium with an outcome in P. Clearly, each outcome in P assigns zero probability to each terminal node that follows an inferior action. Let $A_n^0(u_n)$ be the set of n's inferior actions at u_n and use $A_n^1(u_n) \equiv A_n(u_n) \backslash A_n^0(u_n)$ to denote n's non-inferior actions at u_n. For $i = 0, 1$ let U_n^i be the collection of information sets u_n for which $A_n^i(u_n) \neq \emptyset$ and let $A_n^i = \cup_{u_n \in U_n^i} A_n^i(u_n)$, assuming all actions are labeled differently. (To simplify notation we omit mention of the dependence of these sets on P.)

If $u_n \in U_n^0 \setminus U_n^1$ (i.e. all actions at u_n are inferior) then necessarily u_n has a predecessor and the action a_n' at its nearest predecessor u_n' that leads to u_n is also inferior, so continuing backward in the game tree one finds the unique last predecessor of u_n in $U_n^0 \cap U_n^1$.

Say that a pure strategy of player n is inferior if it chooses an inferior action at some information set that it does not exclude. Let S_n^0 be the set of n's inferior pure strategies. By the previous paragraph, each $s_n \in S_n^0$ chooses an inferior action at some $u_n \in U_n^0 \cap U_n^1$ that it does not exclude. Use $S_n^1 = S_n \setminus S_n^\circ$ to denote n's non-inferior strategies. No strategy in S_n^0 is an optimal reply to any equilibrium in $\Sigma(P)$, and the support of an equilibrium in $\Sigma(P)$ is contained in $S_1^1 \times S_2^1$.

3.3. **Definition of Forward Induction in Extensive-Form Games.** The pruned game $\Gamma(P)$ is obtained from Γ by deleting every path that includes a P-inferior action, i.e. an action in $A_1^0 \cup A_2^0$. Observe that $\Gamma(P)$ is well-defined, i.e. deleting paths that include inferior actions yields a game tree. Indeed, as above, each information set $u_n \in U_n^0$ where an inferior action is available is either in U_n^1 or has a unique last predecessor in $U_n^0 \cap U_n^1$. Therefore $\Gamma(P)$ is obtained by deleting the inferior continuations from each $u_n \in U_n^0 \cap U_n^1$. In particular, the nodes of $\Gamma(P)$ are the nodes of Γ that are preceded only by non-inferior actions of both players. Let $U_n(P)$ be the subcollection of n's information sets in Γ that contain at least one of these nodes. Then $U_n(P)$ is a subset of U_n^1 and it corresponds one-to-one to his collection of information sets in $\Gamma(P)$, so we use the same notation for both games.

---

[3]If Γ has generic payoffs then all equilibria in any connected set of the Nash equilibria have the same outcome [10, 18, 20]. In this case it suffices to consider the singleton set of the unique outcome from a component of the Nash equilibria.

Since an outcome in P assigns probability zero to terminal nodes that follow inferior actions, its projection to the terminal nodes of the pruned game $\Gamma(P)$ is well defined. Use $\mathrm{proj}(P)$ to denote the set of projections of outcomes in P to the terminal nodes of $\Gamma(P)$. One sees easily that $\mathrm{proj}(P)$ is a set of Nash equilibrium outcomes of $\Gamma(P)$.

The following definition modifies Definition 2.1 to allow P to be a set of outcomes.

Definition 3.2 (Extensive-Form Forward Induction)**.** The set P of outcomes of Γ satisfies forward induction if $\mathrm{proj}(P)$ includes a backward induction outcome of the pruned game $\Gamma(P)$.

Later we use the following sufficient condition for forward induction.

Lemma 3.3. *The set P of outcomes of Γ satisfies forward induction if there exists a sequence b^ε of behavioral strategy profiles converging to an equilibrium $b \in B(P)$ and a sequence of equivalent mixed strategies σ^ε such that*

(1) *For each n and $u_n \in U_n(P)$, b prescribes an optimal action against the sequence b^ε.*
(2) *Each strategy $s_n \in S_n^1$ is in the support of σ_n^ε for all ε and $\lim_{\varepsilon \downarrow 0} \sigma_{n,s_n^0}^\varepsilon / \sigma_{n,s_n}^\varepsilon = 0$ for all $s_n^0 \in S_n^0$.*

Proof. By Blume, Brandenberger, and Dekel [2, Proposition 2] we can replace σ^ε with a convergent subsequence to construct a lexicographic probability system [LPS] $(\sigma_n^0, \ldots, \sigma_n^{K_n})$ for each player n and a corresponding sequence $(\lambda_n^1(\varepsilon), \ldots, \lambda_n^{K_n}(\varepsilon)) \in (0,1)^{K_n}$ converging to the origin as $\varepsilon \downarrow 0$ such that the sequence σ_n^ε is expressible as the nested combination $(1 - \lambda_n^1(\varepsilon))\sigma_n^0 + \lambda_n^1(\varepsilon)((1 - \lambda_n^2(\varepsilon))\sigma_n^1 + \cdots \lambda_n^{K_n}(\varepsilon)\sigma_n^{K_n})$. Obviously σ_n^0 is the limit of σ_n^ε and it is equivalent to b_n. Let k_n^* be the smallest integer $k \geqslant -1$ such that σ_n^{k+1} has some strategy in S_n^0 in its support. Since σ_n^0 is equivalent to b_n and b belongs to $B(P)$, $k_n^* > -1$. Also, by Assumption (2) of the lemma, every pure strategy in S_n^1 is assigned a positive probability by some $i \leqslant k_n^*$. Thus the union of the supports of σ_n^i, $0 \leqslant i \leqslant k_n^*$, is S_n^1.

For each n, let $\bar{\sigma}_n^\varepsilon$ be the sequence $d(\varepsilon) \sum_{i=0}^{k_n^*} \varepsilon^i \sigma_n^i$ where $d(\varepsilon)$ is the normalizing factor. Let $\bar{b}_n^\varepsilon$ be an equivalent sequence of behavioral strategies converging to some $\bar{b}_n$. As we saw above, the support of $\bar{\sigma}_n^\varepsilon$ is S_n^1 all along the sequence; therefore, at each information set in U_n^1 (and hence also in $U_n(P)$) $\bar{b}_n^\varepsilon$ mixes completely over the actions in $A_n^1(u_n)$. Moreover, we claim that $\bar{b}_n$ and b_n agree at each $u_n \in U_n(P)$. Indeed, for each $u_n \in U_n(P)$ the set $S_n^1(u_n)$ of pure strategies in $s_n \in S_n^1$ that choose all the actions preceding u_n is nonempty; therefore the smallest integer $0 \leqslant i \leqslant K_n$ such that σ_n^i contains a strategy in $S_n^1(u_n)$ in its support is no more than k_n^*. The mixture prescribed both $\bar{b}_n$ and b_n coincide with that prescribed a behavioral strategy equivalent to σ_n^i, which proves our claim.

All along the sequence, $\bar{b}_n^\varepsilon$ mixes completely over actions in $A_n^1(u_n)$ for each $u_n \in U_n(P)$, and hence it induces a sequence $\hat{b}_n^\varepsilon$ of completely mixed behavioral strategies in $\Gamma(P)$ whose limit $\hat{b}_n$ is induced $\bar{b}_n$ (or equivalently b_n). By passing to a subsequence we will assume that for each n and each $u_n \in U_n(P)$, the set of optimal actions against $\hat{b}_n^\varepsilon$ is constant across the sequence. To finish the proof it is sufficient to show that at each $u_n \in U_n(P)$ the mixture prescribed by $\hat{b}_n$ is optimal against the sequence $\hat{b}_m^\varepsilon$. Suppose a_n is an optimal action at an information set $u_n \in U_n(P)$ against the sequence $\hat{b}_m^\varepsilon$ and suppose a_n' is another action there that is not. There exists $0 \leqslant i \leqslant k_m^*$ such that: σ_m^i does not exclude u_n and a_n is a better action than a_n' against it and for each $i' < i$, either $\sigma_m^{i'}$ excludes u_n or both actions are optimal continuations against it. Given this property, and given the nestedness property of σ_m^ε, which is equivalent to b_m^ε, a_n is a better action against b_m^ε for all small ε than a_n'. By Assumption (1) of the lemma, a_n' is therefore assigned zero probability by b_n. Hence in $\hat{b}_n$ it is assigned zero probability as well. Thus $\hat{b}_n$ is optimal against the sequence $\hat{b}_m^\varepsilon$. $\square$

Remark 3.4. Looking at the beliefs induced by the sequence b^ε, one sees that for each n and each information set $u_n \in U_n(P)$ the limit of the beliefs assigns zero probability to nodes that follow inferior actions. Thus, the lemma shows the relation between deleting inferior strategies and directly restricting beliefs in the original game, as discussed in the motivation for the earlier Definition 2.1 of the test for forward induction in §2.

To invoke invariance in Theorem 3.5 below, we define two extensive-form games that are equivalent to Γ in that they have the same reduced normal forms. The first is called the *splintered* version of Γ, and the second, called a *test game*, adjoins payoff-redundant strategies to the normal form. These are defined in the next two subsections.

3.4. **The Splintered Version.** Define the splintered version $\tilde{\Gamma}$ to be the same as Γ except that each player n chooses an action at each information set $u_n \in U_n^0 \cap U_n^1$ by first choosing whether or not to play an inferior action and then choosing an action from the chosen subset $A_n^0(u_n)$ or $A_n^1(u_n)$. Specifically, such a u_n is separated into three information sets $\tilde{u}_n$, u_n^0, and u_n^1 such that at $\tilde{u}_n$ he decides between two actions $\alpha^0(\tilde{u}_n)$ and $\alpha^1(\tilde{u}_n)$ and then: (0) choosing $\alpha^0(\tilde{u}_n)$ leads to u_n^0 where only actions in $A_n^0(u_n)$ are available, or (1) choosing $\alpha^1(\tilde{u}_n)$ leads to u_n^1 where only actions in $A_n^1(u_n)$ are available. Let $\tilde{S}_n$, $\tilde{B}_n$, and $\tilde{\Sigma}_n$ be the sets of pure, behavioral and mixed strategies for player n in the splintered version $\tilde{\Gamma}$.

As with Γ, we use the simplified normal form of $\tilde{\Gamma}$, i.e. a pure strategy is an equivalence class of strategies that agree both on the information sets they exclude and on the actions at those they do not. Under this assumption, the two extensive-form games $\tilde{\Gamma}$ and Γ have

the same normal form. In order to make clear to which game we are referring, we use $\tilde{S}_n$ to denote the set of pure strategies in $\tilde{\Gamma}$. Let $\tilde{S}_n^0$ be n's set of pure strategies in $\tilde{S}_n$ that choose the action $\alpha^0(\tilde{u}_n)$ at some non-excluded $\tilde{u}_n$. And let $\tilde{S}_n^1 = \tilde{S}_n \backslash \tilde{S}_n^0$. Then for $i = 0, 1$, $\tilde{S}_n^i$ corresponds to the set S_n^i in Γ.

3.5. **Test Games.** The gist of the Hillas-Kohlberg conjecture is that backward induction outcomes that survive addition of redundant strategies (i.e. satisfy invariance) must also survive deletion of inferior strategies. Theorem 3.5 below establishes that testing whether the outcomes in P satisfy forward induction (i.e. some outcome in $\mathrm{proj}(P)$ is a backward induction outcome of the pruned game) is equivalent to checking whether some outcome in P survives as a backward induction outcome of each game in a sequence of *canonical test games* parameterized by $\delta \in (0, 1)$ as $\delta \downarrow 0$. Each test game adjoins payoff-redundant pure strategies, so its normal form is larger, in contrast to the pruned game whose normal form is smaller.

Each test game $\tilde{\Gamma}(P, \delta)$ treats some mixed strategies of Γ as additional pure strategies in the resulting normal form. These payoff-redundant strategies are constructed as follows. For each function $\pi_n : U_n^1 \to A_n^1$ such that $\pi_n(u_n) \in A_n^1(u_n)$, and $0 < \delta < 1$, let $b_n(\pi_n, \delta)$ be a behavioral strategy in Γ that at each $u_n \in U_n^1$ chooses $\pi_n(u_n)$ with probability $1 - \delta$ and with probability δ mixes uniformly over the actions in $A_n^1(u_n)$. Let $\sigma_n(\pi_n, \delta)$ be an equivalent mixed strategy. By construction, the support of $\sigma_n(\pi_n, \delta)$ is S_n^1. Let $T_n(P, \delta)$ be the collection of these mixed strategies obtained from all possible functions π_n. Similarly, in the splintered version $\tilde{\Gamma}$, for each $\sigma_n(\pi_n, \delta) \in T_n(P, \delta)$ let $\tilde{\sigma}_n(\pi_n, \delta)$ be the equivalent mixed strategy and let $\tilde{T}_n(P, \delta)$ be the collection of these mixed strategies.

The following two facts about the strategies in $T_n(P, \delta)$ (and analogously in $\tilde{T}_n(P, \delta)$) are important to our construction. First, for every mixed strategy $\sigma_n(\pi_n, \delta) \in T_n(P, \delta)$ the probability of each pure strategy $s_n \in S_n^1$ is at least $c\delta^{|U_n^1|}$, where c is a positive constant that is independent of π_n and δ. Second, there exists $\delta > 0$ such that if a pure strategy s_n is an optimal reply to an equilibrium with an outcome in P then for each $\delta' < \delta$ there exists a strategy in $T_n(P, \delta')$ that is a better reply against that equilibrium than each strategy in S_n^0; viz., this strategy puts greater weight on the optimal actions prescribed by s_n at those $u_n \in U_n^1$.

A test game $\tilde{\Gamma}(P, \delta)$ is an extensive-form game played in two stages, constructed as follows.

Stage 1: The first stage consists of simultaneous moves by the two players in which each
  player n chooses a strategy in $\tilde{S}_n^1 \cup \tilde{S}_n^0 \cup \tilde{T}_n(P, \delta)$ using a two-step procedure. Player n
  first chooses between two actions x_n and y_n. If he chooses x_n then that completes the first

stage for him. If he chooses y_n then subsequently (at a second information set) he chooses a strategy from the set $\tilde{S}_n^0 \cup \tilde{T}_n(P, \delta)$ to complete the first stage.

Thus for player n the first stage ends with a choice of either x_n or $(y_n, \tilde{s}_n)$ for some $\tilde{s}_n \in \tilde{S}_n^0 \cup \tilde{T}_n(P, \delta)$. Player m does not learn which strategy n implemented in the first stage; viz., his information sets in the second stage reveal exactly the same information as in the original game Γ and its splintered version $\tilde{\Gamma}$.

Stage 2: In the second stage, after each pair of choices by the two players in the first stage, there follows a modified copy of the splintered game $\tilde{\Gamma}$. The modifications are as follows. In each copy of $\tilde{\Gamma}$ that follows the choice x_n in the first stage, nature automatically chooses $\alpha^1(\tilde{u}_n)$ for each $u_n \in U_n^0 \cap U_n^1$, leaving player n to make a choice at each of his other information sets. In each copy that follows a choice $(y_n, \tilde{s}_n)$ by player n in the first stage, all his choices are implemented automatically by nature using the strategy $\tilde{s}_n$.

In the test game $\tilde{\Gamma}(P, \delta)$ the set of n's pure strategies that choose x_n at the first stage correspond exactly to the set $\tilde{S}_n^1$ in the splintered version. Thus, his pure strategy set is $\tilde{S}_n(P, \delta) \equiv \tilde{S}_n^1 \cup \tilde{S}_n^0 \cup \tilde{T}_n(P, \delta)$. Since $\tilde{T}_n(P, \delta)$ consists of mixed strategies available in $\tilde{\Gamma}$, $\tilde{\Gamma}(P, \delta)$ has the same reduced normal form as the splintered version $\tilde{\Gamma}$ and hence Γ itself. Letting $\tilde{\Sigma}_n(P, \delta)$ be n's mixed strategies in the test game there is a well-defined linear map $\tilde{f}_n^\delta : \tilde{\Sigma}_n(P, \delta) \to \Sigma$ that sends each mixed strategy in $\tilde{\Gamma}(P, \delta)$ to the implied mixture over strategies in S_n. Define $\tilde{P}(\delta)$ as the outcomes of $\tilde{\Gamma}(P, \delta)$ resulting from the equilibria in $(\tilde{f}^\delta)^{-1}(\Sigma(P))$.

3.6. **Proof of the Conjecture.** The following theorem verifies our version of the Hillas-Kohlberg conjecture that invariance and backward induction imply forward induction. Recall that we implement backward induction by a quasi-perfect equilibrium.

Theorem 3.5. *P satisfies forward induction if, for each small $\delta > 0$, $\tilde{P}(\delta)$ contains a backward induction outcome of the canonical test game $\tilde{\Gamma}(P, \delta)$ that has the same reduced normal form as Γ.*

Proof. For each small $\delta > 0$ let $\tilde{b}(\delta, \varepsilon)$ be a sequence of ε-quasi-perfect profiles converging as $\varepsilon \downarrow 0$ to a quasi-perfect equilibrium $\tilde{b}(\delta)$ of the test game $\tilde{\Gamma}(P, \delta)$ that induces an outcome in $\tilde{P}(\delta)$. Let $\tilde{\sigma}(\delta, \varepsilon)$ be a (sub)sequence of equivalent mixed strategies in $\tilde{\Gamma}(P, \delta)$ converging to $\tilde{\sigma}(\delta)$. For each n, express $\tilde{\sigma}_n(\delta, \varepsilon)$ as the convex combination $p_{x_n}(\delta, \varepsilon)\tilde{\tau}_n^1(\delta, \varepsilon) + p_{y_n}(\delta, \varepsilon)[r_n(\delta, \varepsilon)\tilde{\tau}_n^0(\delta, \varepsilon) + (1 - r_n(\delta, \varepsilon))\tilde{\tau}_n^2(\delta, \varepsilon)]$, where $p_{x_n}(\delta, \varepsilon)$ and $p_{y_n}(\delta, \varepsilon)$ are the probabilities of the two actions x_n, y_n at player n's first information set in $\tilde{\Gamma}(P, \delta)$ under $\tilde{b}(\delta, \varepsilon)$; and the

supports of $\tilde{\tau}_n^1(\delta, \varepsilon)$, $\tilde{\tau}_n^0(\delta, \varepsilon)$, and $\tilde{\tau}_n^2(\delta, \varepsilon)$ are contained in $\tilde{S}_n^1$, $\tilde{S}_n^0$, and $\tilde{T}_n(P, \delta)$ respectively, which determines $r_n(\delta, \varepsilon)$.

Translate these results to the original game Γ as follows. For $i = 0, 1, 2$ let $\tau_n^i(\delta, \varepsilon) = \tilde{f}_n^\delta(\tilde{\tau}_n^i(\delta, \varepsilon))$ and let $\sigma_n(\delta, \varepsilon) = \tilde{f}_n^\delta(\tilde{\sigma}(\delta, \varepsilon))$. By the linearity of $\tilde{f}_n^\delta$, $\sigma_n(\delta, \varepsilon) = p_{x_n}(\delta, \varepsilon)\tau_n^1(\delta, \varepsilon) + p_{y_n}(\delta, \varepsilon)[r_n(\delta, \varepsilon)\tau_n^0(\delta, \varepsilon) + (1 - r_n(\delta, \varepsilon))\tau_n^2(\delta, \varepsilon)]$. Let $\tau_n^i(\delta)$ be the limit of $\tau_n^i(\delta, \varepsilon)$ as $\varepsilon \downarrow 0$. Let $b(\delta, \varepsilon)$ be a sequence of behavioral profiles equivalent to $\sigma(\delta, \varepsilon)$ and let $b(\delta)$ be its limit as $\varepsilon \downarrow 0$.

If δ is small enough then some strategy in $T_n(P, \delta)$ is a better reply against $b_m(\delta)$ than each strategy in S_n^0; therefore, the corresponding strategy in $\tilde{T}_n(P, \delta)$ is a better reply against $\tilde{b}_m(\delta)$. Optimal continuation from player n's information set following the choice of y_n therefore requires that, for all small δ, $r_n(\delta, \varepsilon)$ converges to zero as $\varepsilon \downarrow 0$.

By the sequential rationality of player n's decision following his choice of x_n in the first stage of the test games, $\tilde{\tau}_n^1(\delta)$ is at least as good a reply as any strategy in $\tilde{S}_n^1$ against the sequence $\tilde{\sigma}_m(\delta, \varepsilon)$. In Γ, therefore, for each $u_n \in U_n(P)$, if u_n is not excluded by strategies in a sequence $\tau_n^1(\delta, \varepsilon)$ then the behavioral randomization implied by $\tau_n^1(\delta)$ is optimal against the sequence $b_m(\delta, \varepsilon)$. By the sequential rationality of playing $\tilde{\tau}_n^2(\delta)$ at the node following y_n the total probability of actions that are not optimal against the sequence $\tilde{b}_m(\delta, \varepsilon)$ under $\tilde{\tau}_n^2(\delta)$ is at most δ. In Γ, therefore, the total probability under $b_n(\delta)$ of actions that are suboptimal at $u_n \in U_n(P)$ against the sequence $b_m(\delta, \varepsilon)$ is no more than δ. Choose now a sequence of δ's converging to zero such that $b(\delta)$ converges to an equilibrium $b \in B(P)$ of Γ. For each δ in the sequence choose $\varepsilon(\delta)$ such that $r_n(\delta, \varepsilon(\delta)) \leqslant \delta^{(|U_n^1|+1)}$ for $n = 1, 2$.

We now prove that the corresponding sequence $b(\delta, \varepsilon(\delta))$ in Γ satisfies the two conditions of Lemma 3.3, which completes the proof of the theorem. Regarding condition (1), for each n and each $u_n \in U_n(P)$, if an action at u_n is not optimal against a subsequence of $b_m(\delta, \varepsilon(\delta))$ then as we saw above its probability under $b_n(\delta)$ is at most δ and hence its probability is zero in b_n. Regarding condition (2), $\tilde{\tau}_n^2(\delta, \varepsilon(\delta))$ is a mixture over strategies in $\tilde{T}_n(P, \delta)$, each of which (as a mixed strategy in Γ) has support S_n^1 and assigns probability at least $c\delta^{|U_n^1|}$ to each strategy in S_n^1. Therefore the probability of each strategy in $\tilde{S}_n^1$ is at least $p_{y_n}(\delta, \varepsilon(\delta))[1 - r_n(\delta, \varepsilon(\delta))]c\delta^{|U_n^1|}$. The probability of strategies in $\tilde{S}_n^0$ is at most $p_{y_n}(\delta, \varepsilon(\delta))c\delta^{|U_n^1|+1}$. Hence the limit of the ratios of these probabilities is zero. In Γ, therefore, the corresponding limit of the ratios of probabilities of strategies in S_n^1 and S_n^0 is also zero, which is condition (2) of Lemma 3.3. $\qquad\square$

In the following sections we extend the above result to general games, first for two players in §6 and then for N players in §7.

4. Forward Induction in General Games

In this section we first provide in §4.1 an overview of the assumptions and results obtained in the remainder of the paper, and then in §4.2 we justify the formulation of backward induction in terms of a proper equilibrium that we use for a game in normal form.

4.1. **Summary of Assumptions and Results.**
We assume equivalence of the extensive and normal forms of a game, and we invoke invariance with respect to payoff-redundant strategies in the normal form. Thus we consider all games in normal or extensive form (with perfect recall) having the same reduced normal form to be strategically equivalent.

We interpret backward induction as requiring a quasi-perfect equilibrium in every extensive form with the same normal form. Backward induction is also required to be consistent with invariance in the following sense. The quasi-perfect equilibria in extensive forms that differ only by inessential transformations (i.e., have the same normal form) should be supported by beliefs generated by the same perturbations; that is, by the same perturbations of strategies in the normal form. To represent this version of backward induction in the normal form of a game we rely on the characterization by Hillas [15] and Mailath, Samuelson, and Swinkels [21]. They show that a proper equilibrium of a normal form is the limit of a sequence of ε-proper profiles as $\varepsilon \downarrow 0$ if and only if in every extensive form with that normal form there is a quasi-perfect equilibrium that is the limit of this same sequence. Thus, a proper equilibrium is precisely the right normal-form representation of backward induction when it is required to induce a quasi-perfect equilibrium in every extensive form with that normal form, and conversely. See §4.2 below for further discussion.

We therefore interpret forward induction as follows. A subset of the Nash equilibria of a game satisfies forward induction if its projection contains a proper equilibrium of the game obtained by deleting each player's pure strategies that are inferior replies to every equilibrium in the subset.

For a two-player game we prove in §6 that if a set of equilibria includes a proper equilibrium for every equivalent game then this set satisfies forward induction. In §7 we use slightly stronger versions of invariance and properness for a game with more than two players.

4.2. **Forward Induction for a Game in Normal Form.**
The formulation of backward and forward induction in terms of sequential equilibria of the extensive form used in §2 is not directly usable here. Our aim is to verify a general version of the Hillas-Kohlberg conjecture that invariances and backward induction imply forward induction. The invariances they (and we) invoke are the equivalence of the extensive and normal forms of a game, and the equivalence of all normal-form games having the same reduced normal form after deleting

payoff-redundant pure strategies. Backward induction must therefore be formulated in terms of equilibria of the normal form. We argue below that a proper equilibrium is the best representation of backward induction in the normal form. Although the subsequent theorems do not depend on this interpretation, we offer it to establish a connection to the analogous Theorem 3.5 for games in extensive form.

The first step is to recognize that a sequential equilibrium is not generally an adequate representation of backward induction in an extensive form. As observed by Kohlberg and Mertens [18, §2.4], a sequential equilibrium can use inadmissible strategies and it need not survive inessential transformations of the extensive form. At a minimum, therefore, the formulation should avoid these deficiencies. The apparently weakest refinement of sequential equilibrium that assures admissibility is a quasi-perfect equilibrium of the extensive form, as defined by van Damme [6]. Regarding invariance to inessential transformations, our approach is to enforce this property directly, as follows. Recall that a sequential or quasi-perfect equilibrium specifies for each player a pair comprising a mixed (or behavioral) strategy and a consistent belief that is the limit of the conditional probability system obtained from a convergent sequence of perturbed strategies. We require similarly that a representation of backward induction in the normal form specifies such pairs for every extensive form with the same normal form. Because we assume invariance, moreover, we require that the *same* sequence of perturbed strategies induces the beliefs in each equivalent extensive form.

To implement this requirement we imitate the proof in Hillas [15]. This uses the formulation of equilibrium in terms of lexicographic probability systems, as defined by Blume, Brandenberger, and Dekel [2]. Considering only a two-player game for simplicity, a lexicographic equilibrium is specified by a lexicographic probability system [LPS] for each player n that is a sequence of mixed strategies, say $\mathcal{L}_n = (\sigma_n^0, \ldots, \sigma_n^{K_n})$ such that each of his pure strategies is assigned a positive probability by some level of $\mathcal{L}_n$. A sequential equilibrium has a lexicographic representation using a weak version of optimality, namely, at each of his information sets player n's equilibrium strategy σ_n^0 is an optimal reply to the first level of the other's LPS that does not exclude reaching that information set. A quasi-perfect equilibrium requires further that his strategy is a lexicographically optimal reply to the ensuing subsequence of the other's LPS. Now suppose one insists further that the same LPS for each player should provide a quasi-perfect equilibrium in every extensive form with the same normal form. Then independently of the extensive form, each LPS respects preferences [2]: for any two pure strategies s_n, s_n' of player n, if s_n is a lexicographic better reply than s_n' against his opponent's LPS then s_n is infinitely more likely than s_n' according to $\mathcal{L}_n$, i.e. if s_n' is assigned a positive probability by σ_n^k then s_n is assigned a positive probability by

σ_n^j for some $j < k$. This is precisely Blume, Brandenberger, and Dekel's [2, Proposition 8] characterization of a proper equilibrium of the normal form.

This is essentially the result obtained by Hillas [15] and Mailath, Samuelson, and Swinkels [21]. They characterize a sequence of ε-proper profiles and the proper equilibrium that is its limit as $\varepsilon \downarrow 0$. The following summary version is stated by Hillas and Kohlberg [16, Theorem 7]: "An equilibrium σ of a normal-form game G is supported as a proper equilibrium by a sequence of completely mixed strategies $\{\sigma^k\}$ with limit σ if and only if $\{\sigma^k\}$ induces a quasi-perfect equilibrium in any extensive-form game having the normal form G." Thus, a proper equilibrium is precisely the right representation of backward induction in the normal form when backward induction in any extensive form with that normal form is represented by a quasi-perfect equilibrium, and the perturbations that justify beliefs are invariant across all these extensive forms.[4]

In §7 we use a strengthened version of properness for games with more than two players. That this is necessary can be seen in Example 2, the Beer-Quiche game. Of the two components of equilibria of this game, only the one in which both types of player I choose B contains a proper equilibrium of every equivalent game. But if this game is interpreted as a game with three players by treating the two types of player I as distinct players then both components have such equilibria; e.g., both components have proper equilibria (Kohlberg and Mertens [18, §3.6.B]). When there are more than two players, therefore, properness must be strengthened to establish an analogous version of the Hillas-Kohlberg conjecture. The stronger version we use is designed to handle the nonlinearities that occur in N-player games, and in particular to control the relative magnitudes of the probabilities of inferior strategies in the sequence of ε-proper profiles whose limit is a proper equilibrium.

The main tool in the two-player case is Blume, Brandenberger, and Dekel's [2, Proposition 8] characterization of a proper equilibrium by a lexicographic probability system. The analysis of the N-player case uses Lojasiewicz's inequality [3, Corollary 2.6.7].

5. Definitions for Games in Normal Form

We consider a finite game G. The set of players is $\mathcal{N} = \{1, \ldots, N\}$. For each player $n \in \mathcal{N}$, let S_n and Σ_n be his sets of pure and mixed strategies, respectively, and interpret S_n as the vertices of the simplex Σ_n. Let $\Sigma = \prod_n \Sigma_n$.

---

[4]Van Damme [6] and Kohlberg and Mertens [18, Appendix A] prove that a proper equilibrium of the normal form induces a sequential equilibrium in every extensive form with that normal form. But the converse is false: Hillas and Kohlberg [16, Figure 23] provide an example of an improper equilibrium that induces a sequential (in fact, quasi-perfect) equilibrium in each extensive-form representation with the same normal form. This example does not contradict the results of Hillas [15] and Mailath, Samuelson, and Swinkels [21] because beliefs are generated by perturbations that vary among equivalent extensive forms.

Player n's expected payoff from the profile $\sigma \in \Sigma$ is $G_n(\sigma)$, and $G_n(\sigma_{-n}, \tau_n)$ is his expected payoff if everyone else plays according to σ and he plays $\tau_n \in \Sigma_n$.

For each $n \in \mathcal{N}$, $0 < \delta < 1$, s_n, and $s'_n \in S_n$, let $\sigma_n(s_n, s'_n, \delta)$ denote the mixed strategy of player n that randomizes between s_n and s'_n with probabilities δ and $1 - \delta$.

5.1. **Invariance.** We invoke two invariance principles that exclude some presentation effects. The first, equivalence of the extensive and normal forms of a game, is implemented by casting our formulation entirely in the normal form. The second requires invariance to addition or deletion of redundant pure strategies. Say that:

(1) A pure strategy is payoff-redundant if its payoffs (for all players, and all pure strategies of other players) are replicated by the expected payoffs from some mixture of the player's other pure strategies.

(2) Two games are equivalent if they have the same reduced normal form (apart from labeling of strategies) obtained by deleting payoff-redundant pure strategies, and

(3) Mixed strategies in equivalent games are equivalent if they induce the same mixed strategy in the reduced normal form.

Specifically, let the columns of the matrix A_n represent player n's pure strategies in game G as mixed strategies in its reduced normal form G^*, and similarly $\tilde{A}_n$ represents his pure strategies in an equivalent game $\tilde{G}$. Then his mixed strategies σ_n, $\tilde{\sigma}_n$, and σ_n^* in the games G, $\tilde{G}$, and G^* are equivalent if $A_n \sigma_n = \tilde{A}_n \tilde{\sigma}_n = \sigma_n^*$. Note that for each player n, $G_n(\sigma) = G_n^*(\sigma^*)$ since G differs from G^* only by adjoining payoff-redundant strategies.

5.2. **Proper Equilibrium.** The definition of a proper equilibrium is due to Myerson [23]. Given $0 < \varepsilon < 1$, a profile $\sigma \in \Sigma$, $\sigma \gg 0$, of completely mixed strategies is ε-*proper* if for each $n \in \mathcal{N}$ and $s_n, t_n \in S_n$,

$$G_n(\sigma_{-n}, s_n) < G_n(\sigma_{-n}, t_n) \quad \text{only if} \quad \sigma_{n,s_n} \leqslant \varepsilon \sigma_{n,t_n}.$$

A profile $\sigma \in \Sigma$ is a *proper equilibrium* if there exists a sequence of positive ε's converging to zero and a corresponding sequence of ε-proper equilibria converging to σ.

5.3. **Deletion of Never Weak Best Replies.** Let Σ^* be a subset of the Nash equilibria of G. Suppose for each n, we are given a subset S_n° of S_n such that each $t_n \in S_n^\circ$ is an inferior reply against every equilibrium in Σ^*. That is,

$$(\forall\, n \in \mathcal{N}, t_n \in S_n^\circ, \sigma \in \Sigma^*) \quad G_n(\sigma_{-n}, t_n) < G_n(\sigma).$$

Consider now the game $\hat{G}$ obtained by deleting the strategies in S_n° for each n. Since all the equilibria in Σ^* assign zero probability to the pure strategies in S_n° for each n, we can view

Σ^* as a subset of the equilibria of $\hat{G}$ by just dropping the coordinates corresponding to the deleted strategies—that is, by projecting Σ^* into the set $\hat{\Sigma}$ of profiles of mixed strategies in $\hat{G}$. In Definition 5.1 below, and throughout, forward induction is applied in this way.

5.4. **Forward Induction in Normal-Form Games.** Our main results establish sufficient conditions for the following version of forward induction.

Definition 5.1 (Forward Induction). A subset Σ^* of the Nash equilibria satisfies forward induction if it includes a proper equilibrium of the game obtained by deleting each player's pure strategies that are inferior replies to every equilibrium in Σ^*.

6. Two-Player Games in Normal Form

Recall that the Hillas-Kohlberg conjecture asks whether invariance and backward induction imply forward induction. The following theorem verifies this conjecture for our normal-form versions of backward and forward induction.

Theorem 6.1. *If a closed subset of the Nash equilibria of a two-player game includes for every equivalent game an equilibrium equivalent to a proper equilibrium of that game then it satisfies forward induction.*

An extensive-form version is the following. If a subset of the Nash equilibria of a two-player game induces a quasi-perfect equilibrium for every extensive form with the same reduced normal form, with beliefs justified by perturbations invariant across extensive forms with the same normal form, then also for the normal-form game obtained by deleting players' strategies that are inferior replies to every equilibrium in the subset, an equilibrium in the subset induces a quasi-perfect equilibrium for every extensive form with that normal form, with beliefs justified by perturbations invariant across these extensive forms.

In outline, the following proof (1) invokes the hypothesized existence of a proper equilibrium for the equivalent game obtained by adjoining redundant pure strategies that are mixtures of used and non-inferior strategies, (2) observes that if the used strategies have large probabilities in these mixtures then the inferior pure strategies have positive probability only in mixed strategies in the tail of the lexicographic probability system [LPS] that characterizes this proper equilibrium, and then (3) chops off the tail to obtain a truncated LPS that characterizes a proper equilibrium for the 'pruned' game obtained by deleting the inferior strategies.

Proof of Theorem 6.1. Let Σ^* be a closed subset of the Nash equilibria of G that satisfies the hypothesis of the theorem. Suppose for each player n that S_n° is a subset of pure strategies

that are inferior replies by player n against every equilibrium in Σ^*. Let $\hat{G}$ be the game obtained from G by deleting the strategies in S_n° for each player n. We show that Σ^* contains a proper equilibrium of $\hat{G}$.

For $n = 1, 2$, let S_n^* be the set of pure strategies that are in the support of some equilibrium in Σ^*. From G construct the equivalent game $\bar{G}$ by adjoining the following mixed strategies as pure strategies: for each $s_n \in S_n^*$ and each $s_n' \in S_n \setminus S_n^\circ$, adjoin the mixed strategy $\sigma_n(s_n, s_n', \delta)$, where $0 < \delta < 1$ is sufficiently large that if s_n is a best reply against a strategy profile $\sigma^* \in \Sigma^*$ then $\sigma_n(s_n, s_n', \delta)$ is a better reply against σ^* than each $t_n \in S_n^\circ$. The games G and $\bar{G}$ are obviously equivalent.

By assumption, Σ^* includes an equilibrium σ^* that is equivalent to a proper equilibrium $\bar{\sigma}^*$ of $\bar{G}$. As in Blume, Brandenberger, and Dekel [2, Proposition 5], there exists for each n a lexicographical probability system [LPS] $\bar{\mathcal{L}}_n = (\bar{\sigma}_n^0, \ldots, \bar{\sigma}_n^{K_n})$ over his strategies in $\bar{G}$ such that: (i) $\bar{\sigma}_n^* = \bar{\sigma}_n^0$; (ii) $\bar{\mathcal{L}}_n$ has full support in the sense that each pure strategy is assigned a positive probability by some level of $\bar{\mathcal{L}}_n$; and (iii) $\bar{\mathcal{L}}_n$ respects preferences: for any two pure strategies $\bar{s}_n, \bar{s}_n'$ of player n in the game $\bar{G}$, if $\bar{s}_n$ is a lexicographic better reply than $\bar{s}_n'$ against his opponent's LPS in $\bar{G}$, then $\bar{s}_n$ is infinitely more likely than $\bar{s}_n'$ according to $\bar{\mathcal{L}}_n$, i.e. if $\bar{s}_n'$ is assigned a positive probability by $\bar{\sigma}_n^k$ for some k, then $\bar{s}_n$ is assigned a positive probability by $\bar{\sigma}_n^j$ for some $j < k$.

For each n let k_n^* be smallest integer such that for each $s_n' \in S_n \setminus S_n^\circ$, either s_n' or one of the new "pure" strategies $\sigma_n(s_n, s_n', \delta)$ for some $s_n \in S_n^*$ has a positive probability under $\bar{\sigma}_n^k$ for some $k \leqslant k_n^*$. We claim now that, for each $k \leqslant k_n^*$, $\bar{\sigma}_n^k$ assigns zero probability to every strategy in S_n°. Indeed, choose s_n in the support of σ_n^*; then for each $s_n' \in S_n \setminus S_n^\circ$, the mixed strategy $\sigma_n(s_n, s_n', \delta)$ is a better reply against σ^* than each $t_n \in S_n^\circ$; hence, in the game $\bar{G}$, the pure strategy $\sigma_n(s_n, s_n', \delta)$ is a better reply than each such t_n against $\bar{\sigma}^*$, which is equivalent to σ^*. Since $\bar{\mathcal{L}}_n$ respects preferences, $\sigma_n(s_n, s_n', \delta)$ is infinitely more likely than t_n; therefore, by the definition of k_n^*, t_n is assigned zero probability by level $k \leqslant k_n^*$ of $\bar{\mathcal{L}}_n$.

For each n and $0 \leqslant k \leqslant k_n^*$, let σ^k be the mixed strategy in G that is equivalent to $\bar{\sigma}_n^k$ and let $\mathcal{L}_n$ be the LPS $(\sigma_n^0, \ldots \sigma_n^{k_n^*})$. Using our claim in the previous paragraph, we see that for each k, the support of σ_n^k is contained in $S_n \setminus S_n^\circ$. Therefore $\mathcal{L}_n$ can be viewed as an LPS in the game $\hat{G}$. Moreover, by the definition of k_n^*, each strategy in $S_n \setminus S_n^\circ$ is assigned a positive probability by some level of $\mathcal{L}_n$. Therefore, $\mathcal{L}_n$ has full support in $\hat{G}$. We claim now that it respects preferences as well. Indeed, suppose s_n is a better reply than s_n' against $\mathcal{L}_n$, and s_n' is assigned a positive probability by some level k of $\mathcal{L}_n$. Then s_n is a better reply against $\bar{\mathcal{L}}_n$ than s_n', and either: (a) s_n' or (b) some $\sigma_n(s_n'', s_n', \delta)$ is assigned a positive probability by level k of $\bar{\mathcal{L}}_n$. Since $\bar{\mathcal{L}}_n$ respects preferences and s_n is a better reply than s_n' against $\bar{\mathcal{L}}_n$ as

well, s_n is assigned a positive probability by $\bar{\sigma}_n^j$ for some $j < k$ if (a) holds and $\sigma_n(s_n'', s_n, \delta)$ is assigned positive probability by $\bar{\sigma}_n^j$ for some $j < k$ if (b) holds. Either way, s_n is assigned a positive probability by σ_n^j for some $j < k$. Thus $\mathcal{L}_n$ respects preferences.

Since $\mathcal{L}_n$ has full support and respects preferences the projection of σ^* is a proper equilibrium of $\hat{G}$, using the characterization by Blume, Brandenberger, and Dekel [2, Proposition 8]. □

Observe that in the above proof, the set S_n^0 of strategies that were deleted for player n could be a proper subset of the set of his strategies that are inferior replies against every equilibrium in Σ^*. Thus the theorem shows that Σ^* satisfies forward induction in a slightly stronger form. The same is true of its N-player analog in the next section.

6.1. **Existence.** Kohlberg and Mertens [18, Proposition 5] prove that every game has a fully stable set of equilibria, and each fully stable set satisfies invariance and includes a proper equilibrium for every equivalent game. Hence the hypothesis of Theorem 6.1 is satisfied by a fully stable set of equilibria. Stable sets as defined by Mertens [22] and metastable sets [11] also satisfy the hypothesis.[5]

7. N-Player Games in Normal Form

In this section we strengthen the formulations of invariance and backward induction to obtain an analog of Theorem 6.1 for games with more than two players. We mentioned in §4.2 the necessity of a strengthened version of properness. In fact, the insufficiency of a proper equilibrium stems from its definition as the limit of a sequence of ε-proper profiles of mixed strategies for which no positive lower bounds are imposed on the probabilities of pure strategies. Due to the nonlinearities that occur in an N-player game, our method of proof requires positive lower bounds to control the relative magnitudes of the probabilities of inferior strategies. The version used here, called 'factorial properness,' suffices because it enforces lower bounds in the definition of the analogous version, called an $\varepsilon!$-proper profile.

7.1. **Factorial Properness.** A strategy profile is $\varepsilon!$-*proper* for $0 < \varepsilon < 1$ if it is an equilibrium of the game $G^{\varepsilon!}$ obtained restricting each player's set of mixed strategies to the polyhedron whose vertices are the $|S_n|!$ permutations of the mixed strategy $(1, \varepsilon, \ldots, \varepsilon^{|S_n|-1})/d(\varepsilon)$, where $d(\varepsilon) = [1 - \varepsilon^{|S_n|}]/[1 - \varepsilon]$ normalizes the probabilities. Say that $\sigma \in \Sigma$ is *proper!* if there exists a positive sequence of ε's converging to zero and a corresponding sequence of

---

[5]In [12] we show that invariance and a 'truly perfect' version of quasi-perfection imply that a set of equilibria is stable as defined by Kohlberg and Mertens [18], and therefore satisfies forward induction in the original sense of Kohlberg and Mertens [18, Proposition 6].

equilibria of $G^{\varepsilon!}$ converging to σ. This stronger version of properness was first considered by Kohlberg and Mertens [18, Proposition 5], although they did not employ this terminology.

7.2. **Strong Invariance.** For $\eta > 0$, say that a pure strategy s_n' is an η-duplicate of s_n if, for all s_{-n}, $G_m(s_{-n}, s_n) = G_m(s_{-n}, s_n')$ for all $m \neq n$, and $G_n(s_{-n}, s_n) = G_n(s_{-n}, s_n') + \eta$. Thus, s_n' is an exact duplicate of s_n from the viewpoint of n's opponents, but for player n himself s_n' is a dominated strategy. Say that a pure strategy is a *near-duplicate* if it is either payoff-redundant or is an η-duplicate of some other pure strategy for some $\eta > 0$. Requiring that a game is equivalent to the game obtained by adjoining near-duplicate strategies is a slight strengthening of invariance as defined in §5.1.

If $\bar{G}$ is a game obtained from G by adding near-duplicate strategies then there is a map π from the strategy set $\bar{\Sigma}$ of $\bar{G}$ to Σ that sends each $\bar{\sigma} \in \bar{\Sigma}$ to the strategy profile $\sigma \in \Sigma$ under which for each n and $s_n \in S_n$, the probability of s_n is the total probability under $\bar{\sigma}$ of the subset of strategies in $\bar{G}$ that are near-duplicates of s_n. Say that $\sigma \in \Sigma$ is equivalent to $\bar{\sigma}$ if $\sigma = \pi(\bar{\sigma})$ and for each n and s_n, $\bar{\sigma}$ assigns zero probability to strategies that are η-duplicates of s_n for some η.

7.3. **An N-Player Version of the Hillas-Kohlberg Conjecture.** We now prove a version of the Hillas-Kohlberg conjecture for an N-player game, using still the formulation of forward induction in Definition 5.1.

Theorem 7.1. *If a closed subset of the Nash equilibria includes, for every game obtained by adjoining near-duplicate strategies, an equilibrium equivalent to a proper! equilibrium of that game, then it satisfies forward induction.*

Proof. Let Σ^* be a closed subset of the Nash equilibria of a game G that satisfies the hypothesis. For each n, let S_n° be a subset of pure strategies that are inferior at each equilibrium in Σ^*. Let $\hat{G}$ be the game obtained from G by deleting the strategies in S_n° for each n, and let $\hat{S}_n$ and $\hat{\Sigma}_n$ be n's pure and mixed strategy spaces in $\hat{G}$. We argue by contradiction that Σ^* must include a proper equilibrium of $\hat{G}$.

Suppose Σ^* does not include a proper equilibrium of $\hat{G}$. Then there exists a closed neighborhood V of Σ^* in $\hat{\Sigma}$ and an $\hat{\varepsilon} > 0$ such that V does not contain an $\hat{\varepsilon}$-proper equilibrium of $\hat{G}$. Take a sufficiently fine triangulation of $\hat{\Sigma}$ such that the set U consisting of the simplices of this triangulation that intersect Σ^* are contained in V. U is then a closed semialgebraic neighborhood of Σ^* that is contained in V. Since U does not contain an $\hat{\varepsilon}$-proper equilibrium of $\hat{G}$, for every completely mixed strategy profile $\sigma \in U$ there exist n and $s_n, s_n' \in \hat{S}_n$ such that $\sigma_{n,s_n'} > \hat{\varepsilon}\sigma_{n,s_n}$ but s_n' is an inferior reply against σ compared to s_n. Therefore the

function $f : U \to \mathbb{R}_+$ given by

$$f(\sigma) = \max_n \max_{s_n, s'_n \in \hat{S}_n} (G_n(\sigma_{-n}, s_n) - G_n(\sigma_{-n}, s'_n))^+ \times (\sigma_{n,s'_n} - \hat{\varepsilon}\sigma_{n,s_n})^+$$

is strictly positive on $U \setminus \partial\hat{\Sigma}$. Define $g : U \to \mathbb{R}$ by $g(\sigma) = \min_{n,s_n} \sigma_{n,s_n}$. Then $f^{-1}(0) \subseteq g^{-1}(0)$ and, by Lojasiewicz's inequality (Bochnak et al. [3, Corollary 2.6.7]), there exists a positive real number c and a positive integer p such that $f \geqslant cg^p$.

Construct a strongly equivalent game $\bar{G}$ by adding the following strategies for each player n as pure strategies:

- For each $s_n \in \hat{S}_n$, $kp - |\hat{S}_n|$ copies of the η-duplicate s_n^η of s_n, where $k \equiv \max_{n'}(|\hat{S}'_n| + |\hat{S}'_n|^2) - 1$ and η is sufficiently small such that if s_n is an optimal reply against an equilibrium in Σ^* then s_n^η is a better reply than each $s_n^\circ \in S_n^\circ$.
- For $s_n, s'_n \in \hat{S}_n$, the strategy $\sigma_n(s_n, s'_n, \delta)$ where $0 < \delta < 1$ is sufficiently large that, if s_n is optimal against an equilibrium in Σ^*, $\sigma_n(s_n, s'_n, \delta)$ is a better reply than the η-duplicate t_n^η for each $t_n \in \hat{S}_n$.

Suppose $\bar{G}$ has a proper! equilibrium equivalent to an equilibrium in Σ^*. Then there exists a sequence of ε_t's converging to zero and a corresponding sequence $\bar{\sigma}^t$ of $\varepsilon_t!$-proper equilibria converging to a point $\bar{\sigma}^*$ that has an equivalent strategy profile $\sigma^* \in \Sigma^*$ in the game G. By replacing the sequence with an appropriate subsequence, we can assume that the preference ordering over the pure strategies in $\bar{G}$ for each n when his opponents play according to $\bar{\sigma}^t$ is independent of t. For each n fix a pure strategy s_n in the original game G that is optimal all along the sequence $\bar{\sigma}^t$. For each $s'_n \in \hat{S}_n$ now, the only strategies in $\bar{G}$ that are possibly at least as good a reply as $\sigma_n(s_n, s'_n, \delta)$ against $\bar{\sigma}^*$ are the strategies in $\hat{S}_n$ and those of the form $\sigma_n(t_n, t'_n, \delta)$ for t_n, t'_n in $\hat{S}_n$—and hence this property holds all along the sequence $\bar{\sigma}^t$ as well. The number of strategies that are possibly no worse replies against $\bar{\sigma}^t$ than $\sigma_n(s_n, s'_n, \delta)$ is therefore $|\hat{S}_n| + |\hat{S}_n|^2 - 1$. Since $\bar{\sigma}^t$ is a sequence of $\varepsilon_t!$-proper equilibria, we then have that for each $s'_n \in \hat{S}_n$, $\varepsilon_t^k = O(\sigma_{n,\sigma_n(s_n,s'_n,\delta)}^t)$, i.e. there exists $C > 0$ such that $\varepsilon_t^k \leqslant C\sigma_{n,\sigma_n(s_n,s'_n,\delta)}^t$ for all t. The strategies that are better replies than the η-duplicate s_n^η of s_n include s_n itself and the strategies $\sigma_n(s_n, s'_n, \delta)$ for $s'_n \in \hat{S}_n$. Hence, $\bar{\sigma}_{n,s_n^\eta}^t = O(\varepsilon_t^{|\hat{S}_n|+1})$ for each copy of the the η-duplicate s_n^η. s_n^η is a better reply than each $s_n^\circ \in S_n^\circ$ against $\bar{\sigma}^*$ and, hence, against the sequence $\bar{\sigma}^t$. Since there are $kp - |\hat{S}_n|$ copies of s_n^η, $\varepsilon_t!$-properness of $\bar{\sigma}^t$ implies that for each $s_n^\circ \in S_n^\circ$, $\bar{\sigma}_{n,s_n^\circ}^t = O(\varepsilon_t^{kp+1})$.

For each element of the sequence $\bar{\sigma}^t$ and each player n, let $\bar{\tau}_n^t$ be the conditional distribution over $\bar{S}_n \setminus S_n^\circ$, viewed as a strategy in $\bar{G}$ by letting the probability of strategies in S_n° be zero.

For each n and $\bar{s}_n \in \bar{S}_n$, $|\bar{\sigma}^t_{n,\bar{s}_n} - \bar{\tau}^t_{n,\bar{s}_n}| = O(\varepsilon_t^{kp+1})$ because $\bar{\sigma}^t_{n,s_n^\circ} = O(\varepsilon_t^{kp+1})$ for all $s_n^\circ \in S_n^\circ$. Therefore, for each player n, if a pure strategy $\bar{s}_n$ is at least as good a reply as another $\bar{s}'_n$ against the sequence $\bar{\sigma}^t$, then it either continues to be so against $\bar{\tau}^t_n$ or $\bar{G}_n(\bar{\tau}^t_{-n}, \bar{s}'_n) - \bar{G}_n(\bar{\tau}^t_{-n}, \bar{s}_n) = O(\varepsilon_t^{kp+1})$.

Let $\hat{\sigma}^t$ be the sequence in Σ that is the image of the sequence $\bar{\tau}^t_n$ under the map π that was defined in §4.2. The support of each element of the sequence is $\hat{S}_n$ and therefore it can viewed as a sequence in $\hat{\Sigma} \backslash \partial \hat{\Sigma}$. In particular, the sequence is eventually in $U \backslash \partial \hat{\Sigma}$. Replace $\hat{\sigma}^t$ with a subsequence that is contained in U and such that there exists n and $s_n, s'_n \in \hat{S}_n$ such that $f(\hat{\sigma}^t) = (G_n(\hat{\sigma}^t_{-n}, s_n) - G_n(\hat{\sigma}^t_{-n}, s'_n))(\hat{\sigma}^t_{n,s_n} - \hat{\varepsilon}\hat{\sigma}^t_{n,s'_n})$ all along the subsequence. We finish the proof of the theorem by showing the following two facts about the sequence $\hat{\sigma}^t$: (1) $\varepsilon_t^k = O(g(\hat{\sigma}^t))$; and (2) $O(\varepsilon_t^{kp+1}) = f(\hat{\sigma}^t)$. Points (1) and (2) contradict our earlier conclusion that $f \geqslant cg^p$ and, therefore, indeed finishes the proof of the theorem.

Point (1) is true because $\varepsilon_t^k = O(\bar{\sigma}_{n,\sigma_n(s_n,s'_n,\delta)})$ and $|\bar{\tau}_{n,\sigma_n(s_n,s'_n,\delta)} - \bar{\sigma}_{n,\sigma_n(s_n,s'_n,\delta)}| = O(\varepsilon_t^{kp+1})$. We turn now to point (2). Since $f(\hat{\sigma}^t) = (G_n(\hat{\sigma}^t_{-n}, s_n) - G_n(\hat{\sigma}^t_{-n}, s'_n))(\hat{\sigma}^t_{n,s_n} - \hat{\varepsilon}\hat{\sigma}^t_{n,s'_n}) \leqslant G_n(\hat{\sigma}^t_{-n}, s_n) - G_n(\hat{\sigma}^t_{-n}, s'_n)$, it is sufficient to prove that $G_n(\hat{\sigma}^t_{-n}, s_n) - G_n(\hat{\sigma}^t_{-n}, s'_n) = O(\varepsilon_t^{kp+1})$. We argue by contradiction. Suppose $G_n(\hat{\sigma}^t_{-n}, s_n) - G_n(\hat{\sigma}^t_{-n}, s'_n) \neq O(\varepsilon_t^{kp+1})$. Then $0 < G_n(\bar{\tau}^t_{-n}, s_n) - G_n(\bar{\tau}^t_{-n}, s'_n) \neq O(\varepsilon_t^{kp+1})$. By what we saw earlier, this implies that s_n is a better reply than s'_n against the sequence $\bar{\sigma}^t$. By ε_t!-properness of the sequence $\hat{\sigma}^t$, therefore: (i) $\hat{\sigma}^t_{n,s'_n} = O(\varepsilon_t\hat{\sigma}^t_{n,s_n})$; (ii) $\hat{\sigma}_{n,\sigma_n(s''_n,s'_n\delta)} = O(\varepsilon_t\hat{\sigma}_{n,\sigma_n(s''_n,s_n,\delta)})$ for each $s''_n \in \hat{S}_n$; and (iii) $\bar{\sigma}^t_{n,s'^\eta_n} = O(\varepsilon_t\bar{\sigma}^t_{n,s_n^\eta})$. These three properties continue to hold for the sequence $\bar{\tau}^t_n$. Hence, in the sequence $\hat{\sigma}^t_n$, $\hat{\sigma}^t_{n,s'_n} = O(\varepsilon_t\hat{\sigma}^t_{n,s_n})$. But that would imply that sufficiently far out in the sequence $f(\hat{\sigma}^t) = (G_n(\hat{\sigma}^t_{-n}, s_n) - G_n(\hat{\sigma}^t_{-n}, s'_n))(\hat{\sigma}^t_{n,s'_n} - \hat{\varepsilon}\hat{\sigma}^t_{n,s_n})^+ = 0$, which is impossible since $\hat{\sigma}^t$ is a sequence of completely mixed strategies in U and f is positive on the whole sequence. This establishes the contradiction and proves point (2). $\square$

8. Van Damme's Interpretation of Forward Induction

The theorems in §3, §6, and §7 do not address the stronger version of forward induction considered by van Damme [8]. Van Damme interprets forward induction as a property of a solution concept (such as stability) rather than a set of equilibria:[6]

> "Kohlberg and Mertens argue that a solution of a game should ... be independent of irrelevant alternatives, ... [this] requirement states that strategies which certainly will not be used by rational players can have no influence

---

[6]An explicit formulation in these terms is attributed to van Damme by Fudenberg and Tirole [9, Definition 11.8].

on whether a solution is self-enforcing; it is the formalisation of the forward induction requirement ... ." [8, §41.4]

That is, a solution concept maps each game into selected subsets of its equilibria, and van Damme sees forward induction as requiring a certain consistency across games, akin to the way independence of irrelevant alternatives is a property of an individual or social choice function. Taking this approach, and supposing that a solution concept satisfies invariance, one can establish analogous results for the following strengthening of forward induction:

> A subset Σ^* of equilibria satisfies strong forward induction if, for every game $\tilde{G}$ equivalent to the game obtained by deleting each player's pure strategies that are inferior replies to every equilibrium in Σ^*, it includes an equilibrium equivalent to a proper equilibrium of $\tilde{G}$.

That is, just as the hypothesis of Theorem 6.1 assumes that Σ^* includes an equilibrium equivalent to a proper equilibrium of *every game equivalent to* the given game, so too the conclusion of the analogous theorem asks for a proper equilibrium of *every game equivalent to* the pruned game. This analogous theorem has a similar proof that was included in a previous version of this paper. It is omitted here to avoid tying backward induction to invariance or tying forward induction to a particular solution concept the way Kohlberg and Mertens [18, Proposition 6] tied it to stability.

References

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

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

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

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

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

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

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

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

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

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

[11] Govindan, Srihari, and Robert Wilson (2006a), "Metastable Equilibria," Research Report 1934, Stanford Business School.

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

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

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

[15] Hillas, John (1996), "On the Relation Between Perfect Equilibria in Extensive Form Games and Proper Equilibria in Normal Form Games," University of Auckland.

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

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

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

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

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

[21] Mailath, George, Larry Samuelson, and Jeroen Swinkels (1997), "How Proper is Sequential Equilibrium?", *Games and Economic Behavior*, 18: 193–218.

[22] Mertens, Jean-François (1989), "Stable Equilibria—A Reformulation, Part I: Definition and Basic Properties," *Mathematics of Operations Research*, 14: 575–625.

[23] Myerson, Roger (1978), "Refinement of the Nash Equilibrium Concept," *International Journal of Game Theory*, 7: 73–80.

DEPARTMENT OF ECONOMICS, UNIVERSITY OF IOWA, IOWA CITY IA 52242, USA.
E-mail address: srihari-govindan@uiowa.edu

STANFORD BUSINESS SCHOOL, STANFORD, CA 94305-5015, USA.
E-mail address: rwilson@stanford.edu