*Econometrica*, Vol. 54, No. 5 (September, 1986), 1003-1037

# ON THE STRATEGIC STABILITY OF EQUILIBRIA[1]

BY ELON KOHLBERG AND JEAN-FRANCOIS MERTENS

A basic problem in the theory of noncooperative games is the following: which Nash equilibria are strategically stable, i.e. self-enforcing, and does every game have a strategically stable equilibrium? We list three conditions which seem necessary for strategic stability—backwards induction, iterated dominance, and invariance—and define a set-valued equilibrium concept that satisfies all three of them. We prove that every game has at least one such equilibrium set. Also, we show that the departure from the usual notion of single-valued equilibrium is relatively minor, because the sets reduce to points in all generic games.

KEYWORDS: Nash equilibrium, stable equilibrium.

## 1. INTRODUCTION

THE CONCEPT OF EQUILIBRIUM, as defined by Nash (1951), is central in the theory of noncooperative games. It reduces the set of all possible strategic choices by the players to a much smaller set of those choices that are stable in the sense that no player can increase his payoff by unilaterally changing his strategy.

One might be tempted to conclude that Nash equilibria must actually be "strategically stable" (self-enforcing). However, such a conclusion would be false, as the example in Figure 1 demonstrates: $(T, R)$, i.e. "$T$" for player I and "$R$" for player II, is a Nash equilibrium. Yet, even though communication is impossible, player II will obviously deviate to $L$ whenever he has to play, thus upsetting the equilibrium.

Since not all Nash equilibria are strategically stable, the natural question that arises is: which ones are?

In the context of games in extensive form, Selten (1975) has proposed the concept of "perfect" equilibrium. Kreps and Wilson (1982) have proposed a variant—"sequential" equilibrium. Both concepts restrict attention to those Nash equilibria that can be obtained by a process of backwards induction (like $(M, L)$ in the example).

But while the restriction to "backwards induction equilibria" is necessary for strategic stability, it is far from being sufficient: One can construct simple examples of perfect or sequential equilibria which, at least in our opinion, are not

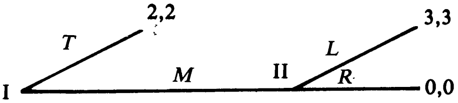


FIGURE 1

[1] This research has been supported in part by National Science Foundation Grant SES-8312910 and by NATO Grant 159-81. Support by the Harvard Business School and by the Institute for Advanced Studies at Hebrew University is also gratefully acknowledged.

strategically stable. Further, both these concepts are flawed in that the equilibria they define may change when the description of the game tree is changed in an irrelevant way (for instance, when a three way choice of some player is replaced by two consecutive binary choices) or when a strictly dominated move is deleted.

In the context of games in normal form, it has long been recognized that the concept of Nash equilibrium is not restrictive enough in that it admits the use of (weakly) dominated strategies. Luce and Raiffa (1957) suggested one might restrict attention to Nash equilibria of the reduced normal form that is obtained by a process of iterated elimination of dominated strategies. Myerson (1978) and Kalai and Samet (1984) have proposed other ideas for restricting the set of Nash equilibria in normal form games ("proper" and "persistent" equilibria). But all these normal form concepts are far from implying strategic stability, and they too suffer from basic flaws.

Furthermore, the two developments—one of restricting the set of Nash equilibria in extensive games, and the other of restricting the set of Nash equilibria in normal form games, have been quite separate (currently prevalent opinion being that normal form analysis cannot capture the essence of backwards induction). Our view is that they should be unified: that a good concept of "strategically stable equilibrium" should satisfy both the backwards induction rationality of the extensive form and the iterated dominance rationality of the normal form, and at the same time be independent of irrelevant details in the description of the game. Our object in this paper is to define an equilibrium concept which satisfies all these requirements.

Section 2 contains a discussion, while Section 3 contains a formal development. Each section can be read independently of the other.

## 2. DISCUSSION OF REQUIREMENTS FOR STRATEGIC STABILITY

### 2.1. *Approach and Point of View*

A noncooperative game is played without any possibility of communication between the players. However, we may think of the actual play as being preceded by a more or less explicit process of preplay communication (the course of which has to be common knowledge to all players), which gives rise to a particular choice of strategies. Loosely speaking, such a prescription of strategies, one for each player, is "strategically stable"[2] if in any actual play of the game, no player will ever have an incentive to deviate from his prescribed strategy.[3]

[2] Note that strategic stability is quite distinct from dynamic stability, which is a property usually associated with an adjustment process, or from evolutionary stability (in the sense of Maynard Smith (1976)).

We also wish to stress that our sole concern is with the strategic stability of a given prescription of strategies, and not with the process of arriving at such a prescription. For example, consider the game below:

|  | $L$ | $R$ |
|---|---|---|
| $T$ | 3, 3 | 0, 0 |
| $B$ | 0, 0 | 1, 1 |

Whatever "incentive to deviate" may mean, once the players expect the strategies $B$, $R$ to be played, For any given game, there is often widespread agreement about which equilibria are more stable than others—although the exact extent of the strategically stable equilibria may be in part a matter of taste. We agree that an ideal way to discuss which equilibria are stable, and to delineate this common feeling, would be to proceed axiomatically. However, we do not yet feel ready for such an approach; we think the discussion in this section will abundantly illustrate the difficulties involved.[4] Instead, we will heuristically discuss some "necessary conditions" for strategic stability, and try to motivate the type of equilibrium concept one is led to if one wishes to satisfy them.

### 2.2. *Review of Definitions*

The discussion in this paper is mostly informal, and can be followed without familiarity with the details of the various equilibrium concepts appearing in the literature. For the sake of completeness, however, we provide below a brief review of the relevant definitions.

We will use the term "(game) tree" for the extensive form of a game with perfect recall (i.e., where every player remembers whatever he knew previously, including his past actions).

The *agent normal form* (Selten) of a tree is the normal form of the game between *agents*, obtained by letting each information set be manned by a different agent, and by giving any agent of the same player that player's payoff.

neither player will have an incentive to unilaterally deviate. So the equilibrium "1, 1" is strategically stable. We will have nothing to say about the distinction between "1, 1" and the more attractive (strategically stable) equilibrium "3, 3." In our mind, such a distinction deals with the pre-play bargaining game, and hence with cooperative theory, rather than with the game itself.

[3] We adhere to the classical point of view that the game under consideration fully describes the real situation—that any (pre)commitment possibilities, any repetitive aspect, any probabilities of error, or any possibility of jointly observing some random event, have already been modelled in the game tree. In particular, the "incentive to deviate" must be understood in the context of a one-shot play of the game itself, and probabilities of error—such as those appearing in the definition of "perfect equilibrium," must not be interpreted as probabilities that the players will actually err in choosing their strategies. Also, no random event (not described in the extensive form) can be observed by a player, except if it is completely independent of any random event observed by any other player and of the moves of nature in the tree. Indeed, any such additional observation would lead to the "extensive form correlated equilibria" of F. Forges (1984). Even before the start of the game such observations have to be forbidden (except for random variables that are common knowledge to all players, if also the analysis is done conditionally to those random variables). Indeed such observations would lead to the "normal form correlated equilibria" (cf. loc. cit.). We must therefore think of the game being played as follows: the referee (or experimenter) starts to select players who do not know each other, and puts them in separate cubicles, with no means of communication to the outside world—not even a window—other than a computer terminal. The players first get from the terminal a full description of this setup, and of the game they are going to play, and next they are told a recommended mixed strategy vector, and that it is a stable equilibrium, expected to be adhered to by all participants. Finally, the computer makes them play the game (informing them whenever they reach an information set, and asking them for their moves—with additional precautions if the game does not have perfect recall). In principle, in situations where those restrictions are not met, the game tree is just used as a shorthand notation for the rules of a much bigger "extended game" (cf. loc. cit.), and it is the stability of the equilibria of the extended game that has to be analyzed.

[4] See also Appendix E.

A *behavioral strategy* of a player in a tree is a list of (mixed) strategies, one for each of his agents. Kuhn (1953) has shown that every mixed strategy of a player in a tree is equivalent to some behavioral strategy, in the sense that both give the same probability distribution on the endpoints whatever be the strategies of all opponents.

A *sequential equilibrium* (Kreps-Wilson) of an $n$-player tree is an $n$-tuple of behavioral strategies which is the limit of a sequence $(\sigma_m)$ of completely mixed (i.e., strictly positive) behavioral strategies, such that every agent maximizes his expected payoff given the strategies of all other agents and given the limiting conditional probability distribution on his information set implied by $(\sigma_m)$.

An *$\varepsilon$-perfect equilibrium* of a normal form game (Selten) is a completely mixed strategy vector, such that any pure strategy which is not a best reply has weight less than $\varepsilon$.

An *$\varepsilon$-proper equilibrium* of a normal form game (Myerson) is a completely mixed strategy vector, such that whenever some pure strategy $s_1$ is a worse reply than some other pure strategy $s_2$, the weight on $s_1$ is smaller than $\varepsilon$ times the weight on $s_2$.

A *perfect (proper) equilibrium of a normal form* game is a limit ($\varepsilon \to 0$) of $\varepsilon$-perfect (proper) equilibria.[5]

A *perfect (proper) equilibrium of a tree* is a perfect (proper) equilibrium of its agent normal form.

It is evident that "proper" is a stronger requirement than "perfect." It is also easy to verify that a perfect equilibrium of a tree is sequential (Kreps-Wilson).

Existence theorems have been proved for all the above concepts (Kreps-Wilson, Myerson, Selten).

## 2.3. *Backwards Induction Rationality*

One necessary condition for "strategic stability" (of an $n$-tuple of prescribed strategies) is that, at every point during any play of the game, each player must believe that his prescribed strategy will maximize his expected payoff in the remainder of the game, i.e., *a strategically stable equilibrium must conform with backwards induction.*

In games of perfect information, the meaning of this requirement is clear (Zermelo (1912)). But in games of imperfect information the meaning is ambiguous at best: for example, is $(T, R)$ a "backwards induction equilibrium" in the following game? (Dotted lines denote information sets.)

[5] An equivalent (in fact, Selten's (1975) original) definition of perfect equilibrium is as follows: $e$ is a perfect equilibrium of a normal form game if for any $\varepsilon > 0$, there exists a vector of positive numbers $\delta_1, \ldots, \delta_n$ ($n$ players), and a vector of completely mixed strategies $\sigma_1, \ldots, \sigma_n$, such that the perturbed game where every strategy $s$ of player $i$ is replaced by $(1-\delta_i)s + \delta_i\sigma_i$ has an equilibrium $\varepsilon$-close to $e$.

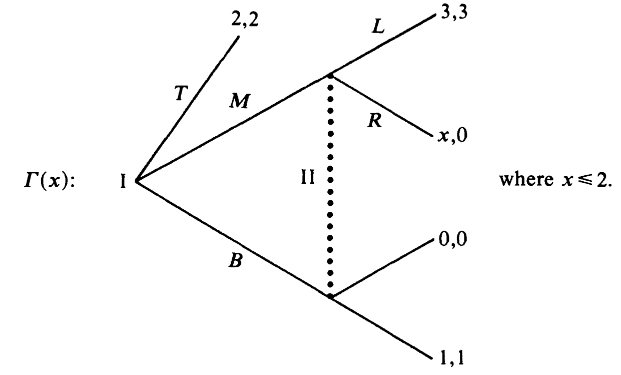


FIGURE 2

The answer depends on II's assessment of the conditional probabilities of the two points in his information set. (It is "yes" if he assesses the conditional probability of the bottom point to be at least 3/4.) But what is a reasonable assessment of the conditional probabilities of two points, each having probability zero?

So the interpretation of "backwards induction" hinges on what assessments of probabilities are considered reasonable. One quite natural interpretation appears to be that of "sequential equilibrium": any assessment is reasonable as long as it is consistent with the probabilities implied by the equilibrium strategies, and as long as assessments in different information sets do not contradict one another. In the example, if I chooses $T$, then any assessment by II is reasonable; so $(T, R)$ is a sequential equilibrium. (Formally, let $(1\text{-}10\varepsilon, \varepsilon, 9\varepsilon)$ and $(\varepsilon, 1-\varepsilon)$ be completely mixed strategies for players I and II, respectively. They converge to "$T$" for player I and "$R$" for player II, and the conditional probability on the bottom point of the information set converges to .9.)

Observe, however, that $(T, R)$ is strategically unstable: player II knows that I will never choose $B$, which is strictly dominated by $T$ (and also by $M$ for $x>1$) so if II sees he has to play, he should deduce that I, who was supposed to play $T$ and was sure to get 2 in this way, certainly did not choose $B$, where he was sure to get less than 2; player II should thus infer that I had in fact played $M$, betting on a chance to get more than 2 (and on the fact that II would understand this signal); and so player II should play $L$, and hence player I should play $M$, deviating from the equilibrium prescription. (Moreover, for $x=2$, player I has an additional, more direct reason to deviate from $T$ to $M$: $T$ is dominated by $M$.)

We see then that conformity with backwards induction, while being necessary for strategic stability, is not sufficient.[6]

[6] This conclusion would remain unchanged if instead of using sequentiality we used some other formal interpretation of "backwards induction," e.g. "perfectness". Indeed, for $x<2$, the "bad" equilibrium "2, 2" is perfect in $\Gamma(x)$ (because the strategies $(1-10\varepsilon, \varepsilon, 9\varepsilon)$, $(\varepsilon, 1-\varepsilon)$ are $(9\varepsilon)$-perfect).

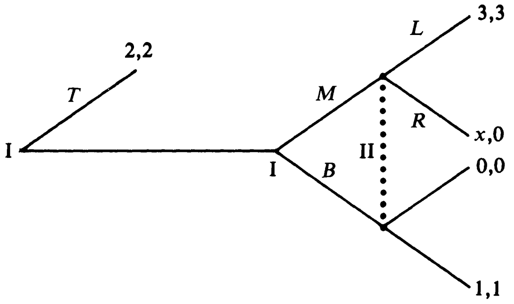


FIGURE 3

## 2.4. *Invariance*

Consider again the game $\Gamma(x)$, with $1<x\leqslant 2$. We saw that the "bad" equilibrium "2, 2" was sequential; however, it is no longer sequential in the above presentation of the same game (Figure 3). (Because here, at the second information set of Player I, his choice $M$ strictly dominates his choice $B$, so that player I has to choose $M$ there rather than $B$ in any sequential equilibrium, and therefore player II should assign probability one to the position following choice $M$. Thus player II has to choose $L$ and therefore player I has to choose $M$: the only sequential equilibrium is "3, 3".)

This example highlights a basic flaw in the concept of "sequential equilibrium": it depends on all the arbitrary details with which the tree was drawn.[7] Thus, if one would like a "strategically stable" equilibrium to be free of the same flaw, one must require that it remain sequential regardless of the irrelevant details in the presentation of the tree. (As we have seen, such a requirement would reduce the set of equilibria in our example to the "good" equilibrium "3, 3".)

The papers by Thompson (1952) and Dalkey (1953) show that if two game trees without moves of nature have the same normal form (i.e., up to duplicated pure strategies) then one may be transformed into the other by a sequence of completely inessential transformations of the game tree, like the one we made for $\Gamma(x)$ ("coalescing of moves"—the other 3 basic transformations being inflation-deflation,[8] addition of superfluous moves,[9] and interchange of simultaneous moves[10]). The same result easily implies its own extension to games with moves of nature, the only additional basic transformation required being that, whenever

[7] "Perfect" equilibrium suffers from the same flaw: As we have seen in footnote 6, for $x<2$, the equilibrium "2, 2" is perfect in the first version of $\Gamma(x)$; however, it is not perfect in the second version (we even know that it is not sequential).

[8] I.e., splitting an information set in two parts if the player can anyway deduce in what part he is from his knowledge of the strategy he is using.

[9] A superfluous move is a move such that its outcome does not affect at all the rest of the game (including terminal payoffs), and such that no player is informed of its outcome.

[10] To represent a pair of simultaneous moves of two players in the tree, one has to draw first the move of one of them, next the move of the other (uninformed of the result of the first choice). One asks that the two different representations thus obtained by considered as equivalent.

a move by nature leads only to terminal nodes, it is equivalent to a terminal node with the corresponding expected payoffs.[11] So any solution that is independent of those 5 categories of irrelevant details of the tree must depend only on the normal form.

In particular, then, a strategically stable equilibrium of a game tree must be sequential in any other game tree having the same normal form.

One may wonder whether this requirement is not so restrictive as to rule out existence. The answer is "no", as is shown by the following proposition, whose proof is given in Appendix A (recall that every normal form game has a proper equilibrium).[12]

PROPOSITION 0: *A proper equilibrium of a normal form is sequential in any tree with that normal form.*

In other words, given a game tree, a proper equilibrium of its normal form will give a sequential equilibrium in any variant of that tree obtained by applying any of the above-mentioned inessential transformations.

Yet even a proper equilibrium may be strategically unstable. For example, the bad equilibrium "2, 2" is proper in $\Gamma(0)$ (because $(1-\varepsilon-\varepsilon^2, \varepsilon^2, \varepsilon), (\varepsilon, 1-\varepsilon)$ is $\varepsilon$-proper).

It appears then that being sequential in any variant of the tree does not guarantee strategic stability. But have we considered all possible variants? Note, in particular, that players are explicitly allowed to randomize between moves (or between strategies) in a game—for instance, player I is allowed to toss some coin to decide between $T$ and $M$ in $\Gamma(0)$. The result of the coin toss is a choíce by nature, so a fully equivalent description of the same game is shown in Figure 4. Yet in this tree the only sequential equilibrium remaining is the "good" equilibrium "3, 3." (At player I's second information set, his choice to toss a coin strictly dominates his choice $B$. So in any sequential equilibrium, he has to choose there either the coin toss or $M$. Therefore the conditional distribution on player II's information set has to assign zero probability to the position following $B$. Thus player II has to choose $L$ and therefore player I has to choose $M$: the only sequential equilibrium is "3, 3.")

If this form of adding or deleting superfluous moves is added to the list of inessential transformations,[13] then two game trees can be transformed one into the other if and only if they have the same normal form, modulo adding or

[11] Use first the above result, treating nature like any other player and handling the probabilities of nature's choices properly, to transform by the first four operations the given tree to a tree with simultaneous moves corresponding to the normal form, with nature moving last. Use then the additional operation to get rid of nature, next again the first four operations to clean the resulting tree of any duplicate strategies, etc.

[12] This proposition first appeared in Kohlberg and Mertens (1982). A refinement is given by van Damme (1984).

[13] If one wants a formal definition in the tree, this sixth elementary transformation says that, in an information set that is followed by no other information set, one may add or delete moves that lead in effect to a lottery between other moves.

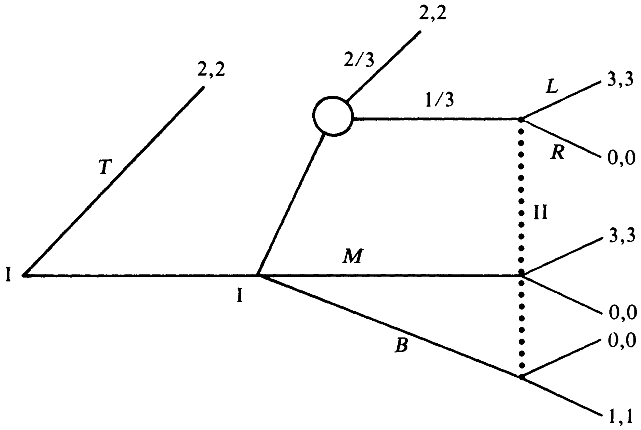


FIGURE 4

deleting pure strategies that are equivalent to convex combinations of other pure strategies (i.e., give the same payoffs to all players whatever be the strategies of all opponents). It is this concept of *reduced normal form* that we will use: where all pure strategies that are convex combinations of other pure strategies have been deleted.

We are thus led to the following invariance requirement, in addition to our previous backwards induction requirement: *The set of "strategically stable" equilibria should depend only on the reduced normal form of the game.*[14]

Putting the two requirements together, a strategically stable equilibrium should conform with backwards induction (e.g., be sequential) in any game tree with the same reduced normal form as that of the given game.

Thus it is natural to ask whether every game has an equilibrium which is sequential in any tree with the same reduced normal form, and whether such an equilibrium must in fact be strategically stable. The first question is addressed in Section 2.8, and the second in Appendix E.

## 2.5. *Why Invariance?*

Selten (1975) writes: "...it is clear that for the purpose of the investigation of the problem of perfectness, the normal form is an inadequate representation of the extensive form." Kreps and Wilson (1982) write: "Analyses that ignore the role of beliefs, such as analysis based on normal form representation, inherently ignore the role of anticipated actions off the equilibrium path.... This lacuna often weakens the normative implications of the analysis, and in the extreme yields Nash equilibria that are patently implausible."

[14] As we have seen, the known solution concepts that satisfy backwards induction (sequential, perfect, proper) fail to be invariant; conversely, the invariant solution concepts (Nash, Nash in undominated strategies, normal form perfect, persistent, etc.) do not yield backwards induction.

In other words, according to this point of view, any equilibrium concept defined on the normal form will miss the essence of backwards induction in the extensive form. We disagree with such beliefs: in fact, Proposition 0 shows them to be false.[15]

More generally, we believe that elementary transformations, like those suggested by Thompson and Dalkey, are irrelevant for correct decision making: after all, the transformed tree is merely a different presentation of the same decision problem, and decision theory should not be misled by presentation effects. We contend that to hold the opposite point of view is to admit that decision theory is useless in real-life applications, where problems present themselves without a specific formalism such as a tree. The results of Thompson and Dalkey therefore imply that any solution concept (under our caveat of footnote 3) should only depend on the normal form.

For an equilibrium solution, one can give an additional argument, which does not rely so much on step by step equivalences of games (more or less independently of the solution concept), but argues instead from the solution concept: In essence, an equilibrium is just a simultaneous solution of each player's individual decision problem. Therefore, since the normal form is sufficient in individual decision theory, and since the normal form of the game allows the recovery of all conceivable such one person normal forms (i.e. for any prior on the strategies of the opponents), it should be sufficient for equilibria.

Put slightly differently, no reasonable definition of rationality could imply a different behavior for the strategist when he has to give instructions to his agents in advance of the play, as compared to the situation where he would have to carry out those instructions himself.

In some sense, the fact that the reduced normal form captures all the relevant information for decision purposes results directly from the (almost tautological) fact that what matters for decision purposes in an outcome is only the corresponding utility vector (and not e.g., the particular history leading to that outcome). This is the point of view we adopt in the whole paper (e.g., when interpreting the outcomes as the corresponding utility vectors in the result of Thompson (1952)).

This is also the argument for adding the fifth elementary transformation (for games with moves of nature) to those of Dalkey and Thompson, and the specific argument for adding the sixth is quite similar.

Indeed, the extensive form is only an abbreviated notation for a fuller description where any additional choice of a lottery between several actions is explicitly available to the players.[16]

[15] Proposition 0 seems the most striking consequence of properness and follows almost immediately from the definition. Was it the above mentioned point of view that prevented it from being discovered immediately?

[16] Even physically if so desired; e.g., in the parable of the strategist and the agents, by instructing the agent to make that specific lottery; or, in our parable of the player sitting in a cubicle (footnote 3), where he has for instance to push one of two buttons to indicate his move, by having the player build for himself a small mechanical device with three buttons, two of which make the device push the corresponding button of the terminal and the third makes the device randomize between the two. More realistically, a company might use agents. In fact, many organizations set policies in one decision center, and have other levels of the organization implement them.

This argument implies more generally a second aspect of the invariance requirement, namely that *one should treat mixed strategies just like pure strategies.* In particular, one should therefore also identify any two "duplicate" mixed strategies. It is in this sense that we will interpret the reduced normal form *strategies* (i.e., as the equivalence classes given this identification).[17]

Anyway, the sufficiency of the normal form is the traditional position, and we think it is sufficiently well founded that, in order to substantiate beliefs like the ones cited above and to reject this classical point of view, one would need an example of two game trees with the same normal form, and whose "reasonable equilibria" are completely different—say every equilibrium payoff whatsoever is patently implausible in one of the two trees. Preferably, both trees should in addition be generic.[18]

No examples approaching anything of this nature were ever produced. Indeed, if one interprets "patently implausible" as "nonsequential," then Proposition 0 already implies the impossibility of such an example; even if one has much more stringent requirements for plausible equilibria, and even if one interprets the "normal form" as the "reduced normal form," this paper will imply the impossibility of such an example.

### 2.6. *Remarks on Backwards Induction*

In games of perfect information, the idea of backwards induction may be characterized by the following properties:

(BI0) a solution of a one-player game should be consistent with payoff maximization;

(BI1) a solution of a game induces a solution in any subgame;

(BI2) any solution of a subgame is part of a solution of the game;

(BI3) a solution of a game remains a solution when a subgame is replaced by a terminal position at which the players receive their expected payoffs (according to this solution) in the subgame.

Sequential equilibrium seems to be the direct generalization to games of imperfect information, because it satisfies all four properties (Nash equilibrium does not satisfy BI1, perfect and proper equilibria do not satisfy BI3; it is not clear whether proper equilibrium satisfies BI2).

But, while in (generic) games of perfect information, all these properties seem to conform with the idea of strategic stability, this is no longer the case in games of imperfect information. Specifically, we claim that one should not insist on BI2, and perhaps not on BI3, as properties of a "strategically stable" equilibrium.

[17] For instance, all mixed strategies that give rise to the same behavioral strategy will be equivalent. This identification makes the interpretation of statements like Proposition 0, or the comparison of the solution of two games having the same reduced normal form, both easier and more natural.

[18] Also, one would then want a precise statement about which one exactly of the elementary transformations is objectionable and what should be the influence of this transformation on a "good" solution concept. The solution concept exhibiting this specific influence should then still be invariant under the other elementary transformations.

(One should certainly insist on BI0; as for BI1, it—like Selten's "subgame perfectness"—follows immediately from the idea of strategic stability.)

Regarding BI2, our requirement that the solution be invariant under inessential transformations of the tree contradicts it. For example, this invariance requirement implies, as we have seen, that only the equilibrium "3, 3", but not the (strategically stable) equilibrium "1, 1" of the subgame be part of a strategically stable equilibrium of $\Gamma(0)$ (Figure 3).

Essentially what is involved here is an argument of "forward induction": 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. In the above example, it is common knowledge that, when player II has to play in the subgame, preplay communication (for the subgame) has effectively ended with the following message from player I to player II: "Look, I had the opportunity to get 2 for sure, and nevertheless I decided to play in this subgame, and my move is already made. And we both know that you can no longer talk to me, because we are *in* the game, and my move is made. So think now well, and make your decision."

Thus, in some sense, just as much as we would like each strategically stable equilibrium to conform with backwards induction, we would like the solution concept (as a correspondence) not to satisfy all the backwards induction properties: it should violate BI2 to save the forward induction.

Regarding BI3, it is incompatible with admissibility (i.e. the restriction to undominated strategies), which is a much more basic requirement following from strategic stability (see the section below). For example, consider the game shown in Figure 5. If the top subgame is replaced by a terminal position with payoffs (1, 1), admissibility would force player I to choose top in any solution. Similarly he should choose bottom in any solution. Hence the contradiction. (The same example shows that neither perfect nor proper equilibria satisfy BI3.)

## 2.7. *Iterated Dominance Rationality*

### A. *Admissibility*

Basic decision theory postulates that a player will never actually choose an inadmissible, i.e. (weakly) dominated, strategy (e.g., Luce and Raiffa (1957, p.

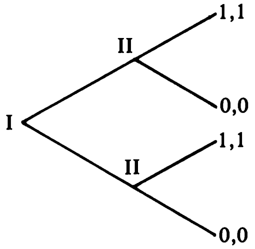


FIGURE 5

287, Axiom 5)). Thus *in a strategically stable equilibrium, the players' strategies must be admissible.* Because the admissibility requirement is central in what follows (e.g., it is the reason we discard an otherwise satisfactory concept, "fully stable equilibrium"), we would like to assuage any doubts the reader might have regarding its exclusion of all dominated strategies, and not only of strictly dominated ones:

First, we note that the founders of decision theory were clearly well aware that an even simpler and in some sense more "elegant" theory could be obtained using a less restrictive axiom involving strict dominance. (For instance, one would then always have the equality of the Bayes procedures and the admissible procedures—while statisticians from Wald on have to make do with statements like "the admissible procedures are the proper Bayes procedures and some of their limits.") Yet they still insisted on requiring admissibility. The requirement goes back to Wald; according to Arrow, this "rule ... is extremely reasonable" (K. J. Arrow (1951, p. 429)), and Luce and Raiffa add (loc. cit., p. 307) that it remains "equally acceptable" in the context of games.

In addition, we may observe that exclusion of dominated strategies follows from the exclusion of strictly dominated moves and from our invariance requirement: assume player I should use strategy $s$ in equilibrium, and $s$ is dominated by $t$. Assume first also that the other players' payoff is the same in rows $s$ and $t$ wherever player I's payoff is so. Then we can draw a tree for this normal form where player I first chooses either the pair $(s, t)$ or any of his other strategies, next player II makes his choice, and finally, player I is asked to choose between $s$ and $t$ only if it matters. In such a game, to choose $s$ is a strictly dominated move, so whatever be player I's beliefs on the others' strategy choices, he should use $t$ rather than $s$. Changing now the others' payoffs to revert to the given game can change player I's beliefs about their strategy choices, but cannot change the fact that, whatever be those beliefs, he should use $t$ rather than $s$.

For those reasons, we follow the above cited authors in considering admissibility of the players' strategies as a basic requirement for strategic stability.

### B. *Iterated Dominance*

One might argue that, since dominated strategies are never actually chosen, and since all players know this, then deletion of such strategies can have no impact on strategic stability. This would lead to requiring that a strategically stable equilibrium remain so when a dominated strategy is deleted (and hence, when the deletion is done iteratively).

Unfortunately, however, this requirement is incompatible with existence. For example, any strategically stable equilibrium in the game

|  | *L* | *R* |
|---|---|---|
| *T* | 3, 2 | 2, 2 |
| *Ω*: *M* | 1, 1 | 0, 0 |
| *B* | 0, 0 | 1, 1 |

should also be strategically stable in the game

|  | *L* | *R* |
|---|---|---|
| *T* | 3, 2 | 2, 2 |
| *M* | 1, 1 | 0, 0 |

so that, by admissibility, it must be (3, 2). But by a similar argument, it must be (2, 2).

It seems then that "strategic stability" requires the presence of both (3, 2) and (2, 2), i.e., we are led to a concept of set-valued equilibrium. (We will later see in Sections 2.8 and 3.5 that each one of two additional basic requirements also leads to a set-valued equilibrium.) But even a set-valued concept cannot satisfy our iterated dominance requirement, because—by exactly the same argument as above—that requirement would imply both that (2, 2) and that (3, 2) must be the *only* point of the set.[19]

One sees that, to preserve existence, we can only ask for inclusion; hence the iterated dominance requirement: ***A strategically stable set of equilibria in a game G must contain a strategically stable set of equilibria in any game G′ obtained from G by deletion of a dominated strategy.***

All the known solution concepts that satisfy backwards induction—sequential, perfect, and proper equilibrium—fail to satisfy the iterated dominance requirement: (the singleton) "2, 2" is sequential, perfect and proper in $\Gamma(0)$ but is no longer so when the (strictly) dominated strategy *B* is deleted.

In fact, it might seem from our previous examples that the only reason those backwards induction solutions failed to imply strategic stability was their failure to satisfy iterated dominance. One might therefore conclude that strategic stability could be obtained by first reducing the normal form to some submatrix by iterative eliminations of dominated strategies,[20] and then applying the relevant backwards induction solution (i.e., proper equilibrium). But, while the procedure would be successful in a game like $\Gamma(0)$, where eliminating *B*, then *R*, then *T* reduces the

[19] The same example shows that the difficulty will not disappear if we restrict attention to the elimination of strictly dominated strategies.

[20] Taking care in one way or another of the difficulty that the resulting submatrix could depend on the order of the eliminations (as in the game *Ω*).

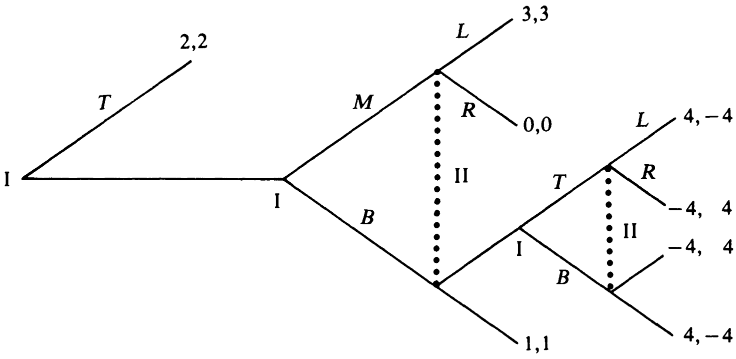


FIGURE 6

normal form to the "good" equilibrium "3, 3," the same procedure fails in the above variant of $\Gamma(0)$ (Figure 6).

In this game, any dominance relationship appears only after the (zero-sum) subgame has been replaced by its value. Thus, all strategies are undominated, and therefore the above-described procedure will simply give all backwards induction solutions, i.e., it will not exclude "2, 2".[21] (In fact, the whole strategy space is persistent.[22] So even if one asked for properness after iterated persistency, the bad equilibrium "2, 2" would persist.)

[21] "2, 2" is proper because $((1, \varepsilon^2, \varepsilon, \varepsilon), (\varepsilon, \varepsilon, 1))$ is $\varepsilon$-proper.

[22] $R = \Pi_{i=1}^n R_i$, where $R_i$ is a closed convex subset of player $i$'s strategy simplex, is ***persistent*** (Kalai and Samet (1984)) if it is minimal with respect to the following property: there is a neighborhood of $R$ such that every $n$-tuple of strategies in that neighborhood has a best reply in $R$. It is easy to see that a persistent set must contain a Nash equilibrium (such an equilibrium is called "persistent").

The normal form of Figure 6 is

| | *LL* | *LR* | *R* |
|---|---|---|---|
| *T* | 2, 2 | 2, 2 | 2, 2 |
| *M* | 3, 3 | 3, 3 | 0, 0 |
| *BT* | 4, −4 | −4, 4 | 1, 1 |
| *BB* | −4, 4 | 4, −4 | 1, 1 |

.

Any persistent set has to contain at least $T$ or $M$, since it contains a Nash equilibrium. Assume it contains $T$: the best reply $LR$ against $(1-\varepsilon)T+\varepsilon BT$ has to be part of the persistent set and similarly $LL$. But $M$ is the only best reply against $\frac{1}{2}LL+\frac{1}{2}LR$, so $M$ is certainly part of any persistent set. Thus the persistent set has to contain $LL$ (best reply against $(1-\varepsilon)M+\varepsilon BB$) and similarly $LR$, and therefore the best replies $BT$ and $BB$ against those strategies. So it has to contain the best reply $R$ against $\frac{1}{2}BT+\frac{1}{2}BB$, and finally the best reply $T$ against $R$: it is the full strategy space.

REMARKS: 1. One might think that the iterated dominance requirement should apply not just for deletion but also for addition of dominated strategies, i.e., that a strategically stable set of equilibria in a game $G$ must be contained in a strategically stable set of equilibria in any game $G'$ obtained from $G$ by addition of a dominated strategy. However, we disagree: consider, for example, the game

| | |
|---|---|
| 3, 2 | 2, 2 |

Every point on the interval from (3, 2) to (2, 2) is a strategically stable equilibrium (II has no incentive to deviate because his payoff is 2 regardless of his choice, whereas I cannot deviate at all). But adding a dominated strategy, we obtain a game

| | |
|---|---|
| 3, 2 | 2, 2 |
| 1, 1 | 0, 0 |

in which (by admissibility) only (3, 2) is strategically stable. F. Forges suggested the following explanation which, while being a bit philosophical, seems to us to be the basic reason for this asymmetry: strategic stability depends on the whole given situation. So, when some implausible alternatives are deleted, the analysis has already taken their unlikeliness into account. However, adding possibilities that were physically not present previously cannot and should not have been anticipated.

2. The admissibility requirement rules out upper-semicontinuity: it implies that $(T, L)$ is the unique strategically stable equilibrium of

| | $L$ | $R$ |
|---|---|---|
| $T$ | 2, 2 | 2, 2 |
| $B$ | 1, 1 | 0, 0 |

even though $(T, R)$ is the unique (strategically stable) equilibrium of

| | $L$ | $R$ |
|---|---|---|
| $T$ | 2, 2 | $2+\varepsilon, 2+\varepsilon$ |
| $B$ | 1, 1 | 0, 0 |

.

### 2.8. *Equivalence of Equilibria*

Our discussion of backwards induction in Section 2.4 concluded with the requirement of sequentiality in any equivalent tree, where two trees were considered equivalent if they had the same reduced normal forms. We start with an example showing that this requirement may be impossible to satisfy.

$\Delta$:

|  | $L$ | $R$ |
|---|---|---|
| $T$ | $1,-1$ | $1,-1$ |
| $M$ | $2,-2$ | $-2,2$ |
| $B$ | $-2,2$ | $2,-2$ |

For any $0 \leqslant \alpha \leqslant 1$, the tree shown in Figure 7 has the same reduced normal form; but in its unique sequential equilibrium, player II goes left with probability $(4-3\alpha)/(8-4\alpha)$.

(In any equilibrium $T$ is at least as good for $I$ as $M$, so giving the hand to nature is also at least as good as $M$. On the other hand, player II has to use both $L$ and $R$ with positive probability in any equilibrium, so that in any sequential equilibrium player I's probability of choosing $B$ has to be strictly between 0 and 1. Thus player I has to expect the same payoff in his second information set from giving the hand to nature as from playing $B$. Solving this equation for player II's probability of $L$ yields $(4-3\alpha)/(8-4\alpha)$.)

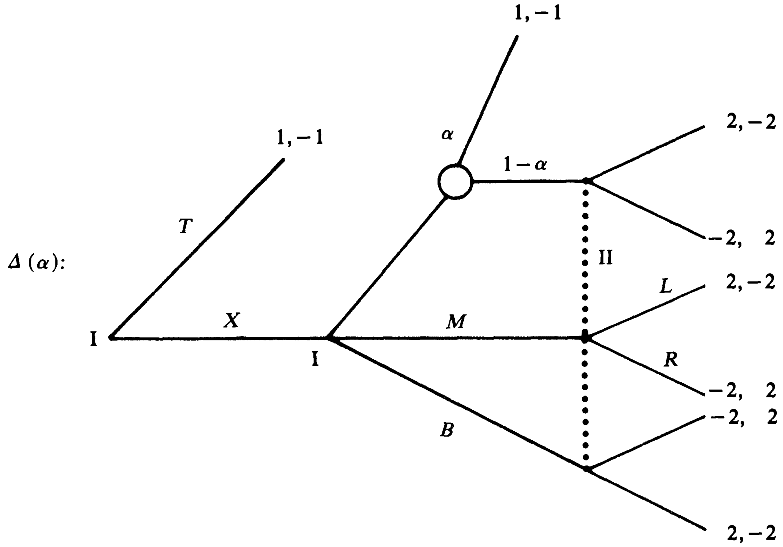


FIGURE 7

(We would like to stress that the argument above, and hence the fact that "sequentiality in every equivalent tree" cannot be satisfied by a single-valued concept, does not depend on the specific definition of sequentiality. Rather, it follows from basic backwards induction, i.e. the requirement that each agent's assessment of the conditional probabilities in his own information set be consistent with the other agents' strategies, and that each agent's strategy maximize his payoff given his assessment. (Sequentiality requires much more, e.g. consistency across assessments of all agents of all players.) Moreover, in the particular example here there can be no ambiguity about what's meant by "backwards induction," because all conditional probabilities are positive and uniquely determined by Bayes' rule from the strategies.)

So it is clear that any concept satisfying "sequentiality in every equivalent tree" will have to consider all equilibria $(1, -1)$ in the game $\Delta$ as equivalent, as being the same.[23]

The example might suggest that equilibria be defined as being equivalent when they give the same payoffs (or the same distribution of payoffs) to all players, i.e., when they differ only off the equilibrium path.[24] However, once we accept such a notion of equivalence, we are forced to also accept as equivalent equilibria which differ along the equilibrium path: Consider first a three-person game tree, $\Delta_3$, which is identical to $\Delta$ except that, after player I's first move, player III, who is uninformed of that move, has to choose $L$ or $R$; this choice has no effect on I and II's payoffs (which remain as in $\Delta$), while the payoff to III is 1 if II has played (i.e., if I has chosen $X$) and matched III's choice, and 0 otherwise. Since the game between I and II is unaffected by III's choice, equilibria that were equivalent in $\Delta$ must still have to be considered equivalent in $\Delta_3$. But, as we have seen, the unique equivalence class of $\Delta$ is its full set of Nash equilibria, where player II's probability of going right, say $s$, varies from $\frac{1}{4}$ to $\frac{3}{4}$. Player III, since his move matters only if $X$, will play as if the superfluous move (i.e. his choice after $T$) has been deleted. Thus he will choose right with probability $t$, where $t = 0$ if $s < \frac{1}{2}$, $t$ varies from 0 to 1 for $s = \frac{1}{2}$, and $t = 1$ for $s > \frac{1}{2}$.

Now consider a five-person game tree, $\Delta_5$ which is identical to $\Delta_3$ except that there are two additional players, IV and V, who are uninformed of the play of $\Delta_3$ and who play one of two zero-sum games (having different values) $\Gamma_L$ or $\Gamma_R$, depending on the move of player III. Since the game between I, II, and III is over before players IV and V enter the picture, equilibria that were equivalent in $\Delta_3$ still have to be considered equivalent in $\Delta_5$. But as $t$ varies from 0 to 1, the payoffs of players IV and V vary from the value of $\Gamma_L$ to that of $\Gamma_R$. (Choosing $\Gamma_L$ and $\Gamma_R$ to be of different types, e.g., one completely mixed and the other with

[23] The set of Nash equilibria in $\Delta$ is $\{((1, 0, 0), (y, 1-y)) : \frac{1}{4} \leq y \leq \frac{3}{4}\}$. We have shown that all equilibria with $\frac{1}{4} \leq y \leq \frac{1}{2}$ (i.e., $y = (4-3\alpha)/(8-4\alpha)$ and $0 \leq \alpha \leq 1$) should be considered equivalent. A variant of $\Delta(\alpha)$ in which the move of nature selects $B$ (rather than $M$) with probability $1-\alpha$ shows that also all equilibria with $\frac{1}{2} \leq y \leq \frac{3}{4}$ should be considered equivalent.

[24] We don't know whether identification of equilibria with the same payoff is sufficient to guarantee the existence of an equilibrium which is sequential in any equivalent tree; i.e., the following is an open problem: Does every game have a payoff vector such that all game trees with the same reduced normal form have a sequential equilibrium with this payoff?

a pure-strategy saddle point, we can even obtain an equivalence class in which the strategies of players IV and V vary with $t$ through all the (differentiable) submanifolds of the set of equilibria.)

We see then that in nongeneric trees, we may sometimes be forced to recognize as equivalent equilibria that differ even along the equilibrium path.

In the examples above, all equilibria within the same connected component of the set of Nash equilibria had to be identified. However, this is not the case in general.

For instance, in Example 8 all equilibria lie within the same connected component (and they even commute); however one—$(T, L)$—is dominant and all the others are dominated, so we certainly cannot consider them equivalent.

In summary, beyond the fact that equivalence classes seem connected, we cannot point to any property (equal payoffs, commutation, being in the same differentiable submanifold, etc.) which is either necessary or sufficient for equivalence.

We wish however to stress that any identification of equilibria within a connected set would not constitute a major departure from the usual notion of single-valued equilibrium. Indeed, *for any generic tree, all equilibria in the same connected component give rise to identical probability distributions over endpoints*—i.e., they differ only off the equilibrium path.

To see this, recall that for a generic tree, the set of probability distributions on the endpoints induced by Nash equilibria is finite (Kreps and Wilson (1982, Theorem 2) and the remarks following its statement; a simple proof is given in Appendix C). Therefore, when the players' strategies vary over a connected set of equilibria, the distributions over endpoints—which are continuous functions of the strategies—must remain constant.

## 2.9. *Main Requirements*

We now rephrase our requirements for strategic stability in term of a set-valued solution concept:

*Existence*: Every game has at least one solution.

*Connectedness*: Every solution is connected.

*Backwards Induction*: A solution of a tree contains a backwards induction (e.g. sequential or perfect) equilibrium of the tree.

*Invariance*: A solution of a game is also a solution of any equivalent game (i.e., having the same reduced normal form).

*Admissibility*: The players' strategies are undominated at any point in a solution.

*Iterated-Dominance*: A solution of a game $G$ contains a solution of any game $G'$ obtained from $G$ by deletion of a dominated strategy.

We wish to emphasize that this list of properties is by no means complete. Moreover, it should not even be viewed as part of an axiom system for "strategically stable equilibrium" because some of the requirements are phrased in terms which are outside decision theory. This list should rather be viewed as a (partial) benchmark against which proposed definitions may be tested (cf. Appendix E for more details).

## 3. STABLE EQUILIBRIUM

### 3.1. *Overview of the Results*

We first give a basic result on the structure of the Nash correspondence. Motivated by this result, we define two preliminary concepts, "hyperstable" and "fully stable" equilibrium, and show that each satisfies existence, backwards induction, invariance, iterated dominance, and a version of connectedness. However, they fail to satisfy admissibility. We view this failure as fundamental, and therefore discard both concepts.

We then give an incomplete definition of what seems to us the "right" concept, which we call "stable equilibrium." In order to not further complicate our terminology, we will refer in this paper to equilibria satisfying the incomplete definition as "stable." We show that stable equilibrium satisfies existence, invariance, admissibility, and iterated dominance. We hope that in the future some appropriately modified definition of stability will, in addition, imply connectedness and backwards induction.

### 3.2. *The Structure of Nash Equilibria*

The theorem below says that the graph of the Nash equilibrium correspondence (when compactified by adding the point $\infty$) is like a deformation of a rubber sphere around the sphere of normal form games (similarly compactified). (We recommend that a reader without an independent interest in the geometry of Nash equilibria skip the precise statement of this theorem and its proof in a first reading of this paper.)

Formally, fix a finite player set $N$ and finite (pure) strategy sets $S_n$. Let $S=\prod_{n\in N} S_n$. Denote by $\Sigma_n$ the space of probabilities on $S_n$; let $\Sigma=\prod_{n\in N}\Sigma_n$. Denote by $\Gamma_n$ the space of payoff functions of player $n$, i.e. $\Gamma_n=R^S$, and let $\Gamma=\prod_n \Gamma_n$. Denote by $E$ the graph of the set of equilibria, i.e. $E=\{(G,\sigma)\in\Gamma\times\Sigma \mid \sigma$ is a Nash equilibrium for $G\}$. For any locally compact space $L$, denote by $\bar{L}$ its one-point compactification. Denote by $p$ the projection mapping $p:E\rightarrow\Gamma$ and denote by $\bar{p}$ its extension by continuity from $\bar{E}$ to $\bar{\Gamma}$, defined by $\bar{p}(\infty)=(\infty)$.

THEOREM 1: *$\bar{p}$ is homotopic to a homeomorphism. More precisely, there exists a homeomorphism $\phi$ from $\Gamma$ to $E$ such that $p\circ\phi$ is homotopic to the identity on $\Gamma$ under a homotopy that extends to $\bar{\Gamma}$.*

PROOF: Let $T_n=\prod_{i\neq n} S_i$; $\Gamma_n$ is the set of all $S_n\times T_n$ payoff matrices $G^n_{s,t}$, but it will be more convenient to use the following reparameterization of $\Gamma_n$: let

$G^n_{s,t}=\tilde{G}^n_{s,t}+g^n_s$, where $\sum_{t\in T_n}\tilde{G}^n_{s,t}=0$, i.e., $g^n_s$ is the average over $t$ of $G^n_{s,t}$. Thus $\Gamma_n$ will be considered as the set of all pairs $(\tilde{G}^n, g^n)$, with $\sum_{t\in T_n}\tilde{G}^n_{s,t}=0$.

Let $z^n_s=\sigma^n_s+\sum_{t\in T_n}G^n_{s,t}\prod_{i\neq n}\sigma^i_{t_i}$.

The $z^n_s$ are continuous functions on $E$. Conversely, given $\tilde{G}$ and any vector $z$, one can recompute the corresponding point of $E$ in a unique and continuous way, as follows:

First, $v^n=\min\{\alpha\mid\sum_{s\in S_n}(z^n_s-\alpha)^+\leqslant 1\}$ (player $n$'s equilibrium payoff); next, $\sigma^n_s=(z^n_s-v^n)^+$; finally

$$(*)\qquad g^n_s=z^n_s-\sigma^n_s-\sum_{t\in T_n}\tilde{G}^n_{s,t}\prod_{i\neq n}\sigma^i_{t_i}\left(=\sum_{t\in T_n}g^n_s\prod_{i\neq n}\sigma^i_{t_i}\right).$$

This homeomorphism, from the set of pairs $(\tilde{G}, z)$ to $E$, is the homeomorphism of the statement; and $p\circ\phi$ maps $(\tilde{G}, z)$ to $(\tilde{G}, g)$.

There only remains to construct the homotopy.

Let, for $t\in[0,1]$, $q_t(\tilde{G}, z)=(\tilde{G}, tz+(1-t)g)$ (and $q_t(\infty)=\infty$). Since $q_0=p\circ\phi$ and $q_1$ (which is the identity) are both continuous, we already know the continuity of $q$ on $[0,1]\times E$; so there only remains to show the continuity of $q$ at all points of $[0,1]\times\{\infty\}$, or equivalently that $\forall M$, $\exists K$ such that $\|(\tilde{G}, z)\|\geqslant K\Rightarrow\forall t$, $\|q_t(\tilde{G}, z)\|\geqslant M$.

Note that $(*)$ implies $|z^n_s-g^n_s|\leqslant|\sigma^n_s|+|\sum_{t\in T_n}\tilde{G}^n_{s,t}\prod_{i\neq n}\sigma^i_{t_i}|$; thus, using the maximum norm throughout,

$$(**)\qquad \|z-g\|\leqslant\|\tilde{G}\|+1.$$

So choosing $K=2M+1$, if $\|\tilde{G}, z\|\geqslant K$ then either $\|\tilde{G}\|\geqslant M$, in which case $\|q_t(\tilde{G}, z)\|\geqslant M$, or $\|\tilde{G}\|<M$ and $\|z\|\geqslant 2M+1$, in which case, by $(**)$, $\|tz+(1-t)g\|\geqslant M$ so again $\|q_t(\tilde{G}, z)\|\geqslant M$. *Q.E.D.*

### 3.3. *Hyperstable Equilibria*

The structure of the Nash equilibrium correspondence implies the following corollary (see Appendix B for a derivation):

PROPOSITION 1: *The set of Nash equilibria of any game has finitely many connected components. At least one of them is such that for any equivalent game (i.e., having the same reduced normal form), and for any perturbation of the normal form of that game, there is a Nash equilibrium close to this component.*

Motivated by this result, we will say that $S$ is a *hyperstable* set of equilibria in a game $G$ if it is minimal with respect to the following property:

PROPERTY (H): $S$ is a closed set of Nash equilibria of $G$ such that, for any equivalent game, and for any perturbation of the normal form of that game, there is a Nash equilibrium close to $S$.

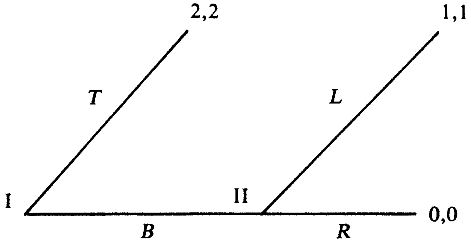
Every game has a hyperstable set of equilibria contained in a single connected component of the set of Nash equilibria. To see this, let $F$ denote the family of subsets of a single connected component satisfying condition (H), ordered by set inclusion. $F$ is nonempty by Proposition 1; every decreasing chain of elements in $F$ has a lower bound (because by compactness the intersection is in $F$); therefore, by Zorn's lemma, $F$ has a minimal element.

It follows (see Section 2.8) that *every generic tree has a hyperstable payoff*, i.e. a (vector) payoff which is implied by all the equilibria in some hyperstable set.

In the next section we define fully stable sets, replacing (H) by a less restrictive condition. So every hyperstable set contains a fully stable set. Since every fully stable set contains a sequential equilibrium (Proposition 3), the same is true for hyperstable sets.

That hyperstable sets are invariant follows from their definition.[25]As to iterated dominance, this follows from an analog of Proposition 2 in which "fully stable" is replaced by "hyperstable."

So "hyperstable equilibrium" satisfies existence, a version of connectedness, backwards induction, invariance, and iterated dominance. However, it does not satisfy admissibility, as the following example shows (payoffs could be perturbed so as to make the tree generic):



i.e.

|   | $L$ | $R$ |
|---|---|---|
| $T$ | 2, 2 | 2, 2 |
| $B$ | 1, 1 | 0, 0 |

FIGURE 8

[25] A concept closely related to hyperstable equilibrium is that of "essential equilibrium": an equilibrium $e$ of a normal form $G$ is *essential* if for any perturbation of $G$ there is a Nash equilibrium close to $e$. Note, however, that essential equilibrium is not an invariant concept. For example, in the game

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

the set $\{(1,0) \cup (3,0)\}$ is essential, but it is no longer so in the equivalent game

| | | |
|---|---|---|
| 1, 0 | 2, 0 | 3, 0 |

.

(In contrast, the only hyperstable set in either game is the full interval, (1, 0) to (3, 0).)

Our result that generic trees have hyperstable payoffs implies in particular that generic normal forms have hyperstable (and therefore essential) equilibria. This result, which is well known (Wu Wen-Tsun and Jiang Jia-He (1962)), is of limited interest—because any normal form arising from a nontrivial tree is nongeneric.

The unique hyperstable set is the full interval from $(T, L)$ to $(T, R)$ but only $(T, L)$ is admissible. (To see that every hyperstable set must include any given mixture of player II, add this mixture as an additional pure strategy and perturb the game by increasing slightly player II's payoff in that column.)

### 3.4. *Fully Stable Equilibrium*

The last example shows that the definition of hyperstability is unsatisfactory, i.e. that requiring stability under all perturbations of payoffs might lead to sets which are too large. A natural idea, then, is to restrict the perturbations of the payoffs to those which arise from perturbations of strategies. That is, we would like to perturb the game by requiring that, whenever a player chooses some pure strategy, it is in fact some (close by) mixed strategy that is played.

Thus, in the simplex of strategies, we get a certain number of interior points, close to the vertices. But we want the definition to remain unchanged when a finite number of additional pure strategies are introduced, convex combinations of the old ones. These may lie anywhere in the simplex, but are also perturbed so as to become interior points. Thus in fact we are in the situation where each player's pure strategy set is replaced by an arbitrary finite subset of completely mixed strategies, containing strategies close to the vertices—i.e., his mixed strategy simplex is replaced by any closed convex polyhedron in the interior of the simplex, and that approximates the simplex in the Hausdorff topology.

We will say that $S$ is a *fully stable* set of equilibria of a game $G$ if it is minimal with respect to the following property:

PROPERTY (F): $S$ is a closed set of Nash equilibria of $G$ satisfying: for any $\varepsilon > 0$ there exists a $\delta > 0$ such that, whenever each player's strategy set is restricted to some compact convex polyhedron contained in the interior of the simplex and at (Hausdorff) distance less than $\delta$ from the simplex, then the resulting game has an equilibrium point $\varepsilon$-close to $S$.

As noted earlier, every hyperstable set includes a fully stable set. Therefore, every game has a fully stable set which is contained in a single connected component of the set of Nash equilibria, and *every generic tree has a fully stable payoff.* (However, fully stable sets may be disconnected, for instance $\{(1,0) \cup (3,0)\}$ in the game

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

.)

Invariance is obvious from the definition, while iterated dominance and backwards induction follow from the two propositions below.

PROPOSITION 2: *A fully stable set contains a fully stable set of any game obtained by deletion of a dominated strategy.*

PROOF: We will show that the set of those equilibria in the fully stable set that assign zero weight to the dominated strategy satisfies $(F)$ in the smaller game: Given a perturbation $G(\varepsilon)$, of the game without the dominated strategy, construct a close-by perturbation, $G(\varepsilon, z)$, by first adding the deleted strategy as an additional extreme point in the relevant player's strategy set, then perturbing it just like the corresponding (mixed) dominating strategy was perturbed in $G(\varepsilon)$, and finally perturbing all the strategies in the amount $z$ towards the new extreme point. Since $G(\varepsilon, z)$ is a perturbation of the original game, it has an equilibrium close to our fully stable set. Such an equilibrium will clearly give zero weight to the dominated strategy. So taking a limit of such equilibria (as $z \rightarrow 0$) will give an equilibrium of $G(\varepsilon)$ close to the fully stable set. *Q.E.D.*

PROPOSITION 3: *A fully stable set of equilibria of a game tree contains a sequential (in fact, even a perfect and proper) equilibrium of the tree.*

The proof of this proposition is an immediate consequence of the two propositions below (recall that a perfect equilibrium of a tree—which is defined as a perfect equilibrium of the agent normal form—is sequential).

PROPOSITION 4: *Given a game tree, a fully stable set of equilibria of its normal form contains a fully stable set of equilibria of its agent normal form.*

PROOF: Given a perturbation of the agent normal form, i.e., a restriction of each agent to a polyhedron of completely mixed strategies, define a polyhedron of completely mixed strategies for any player by taking the convex hull of all points obtained by selecting some extremal strategy for each one of this player's agents. Clearly, if the strategy of any one of his agents is in the agent's polyhedron, then the player's strategy will be in his own polyhedron. Conversely, when the player mixes several behavioral strategies $\sigma_\alpha$ using a lottery over $\alpha$, agent $n$'s component in the behavioral strategy induced by this mixture is the average of the $\sigma_\alpha^n$ weighted by agent $n$'s posterior probability over $\alpha$, and therefore satisfies the agent's restrictions.

Thus, any polyhedral restrictions on the agents' strategies can be obtained by some polyhedral restriction on the players' strategies. *Q.E.D.*

PROPOSITION 5: *A fully stable set of equilibria of a normal form game contains a proper (hence perfect) equilibrium of that normal form.*

PROOF: Restrict each player's strategies to the convex hull of the $k!$ vectors (where $k$ denotes the number of pure strategies) obtained by permuting the coordinates of the vector

$$\frac{1-\varepsilon}{1-\varepsilon^k}(1, \varepsilon, \ldots, \varepsilon^{k-1}).$$

Pick an equilibrium point of this perturbed game in the neighborhood of our set of equilibria. It is an $\varepsilon$-proper equilibrium, because if strategy 1 yields a better payoff than strategy 2, best replies involve only those permutations that give strategy 1 greater weight than strategy 2—so the total probability of strategy 2 will be smaller than $\varepsilon$ times the probability of strategy 1. *Q.E.D.*

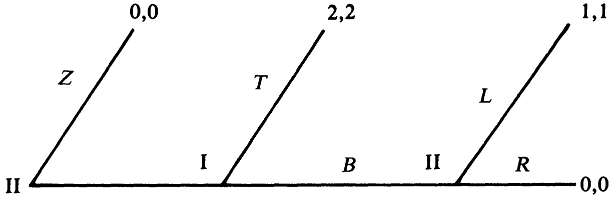
REMARKS: 1. Proposition 4 shows, in fact, that "fully stable equilibrium" satisfies appropriate versions of the backwards induction properties which we labelled BI1 and BI3 (in Section 2.6): *The projection of a fully stable set on a subgame contains a fully stable set of that subgame*[26] and *a fully stable set contains a fully stable set of any truncated game obtained by replacing a subgame with its unique equilibrium.*[27]

2. There is no reason to expect the converse of Proposition 4: a strategically stable equilibrium can certainly become unstable when two different agents of the same player become one. For example, the equilibrium (2, 2) seems strategically stable in the agent-normal form of $\Gamma(0)$ (Figure 3) but, of course, it is not stable in $\Gamma(0)$ itself.

3. The examples in Section 2.8 show that, in order to get backwards induction in every tree, a strategically stable set of the normal form may have to be a full equivalence class, while when focusing through the agent normal form on a given tree, we may get a singleton. Therefore no equality should be hoped for in analogs of Proposition 4.

4. Proposition 4 is false for perfect as well as for proper equilibria: in the example of Appendix A, "1, B" is perfect and proper in the normal form but not in the agent normal form.

So "fully stable equilibrium" satisfies existence, a version of connectedness, backwards induction, invariance, and iterated dominance. But it, too, fails to satisfy admissibility: consider the following variant of Figure 8 (payoffs could be perturbed so as to make the tree generic):



| | $L$ | $R$ | $Z$ |
|---|---|---|---|
| $T$ | 2, 2 | 2, 2 | 0, 0 |
| $B$ | 1, 1 | 0, 0 | 0, 0 |

FIGURE 9

[26] Indeed, given any polyhedral restrictions on the strategies of agents in the subgame, consider in addition some (polyhedral) restrictions on the strategies of agents outside the subgame, and pick an equilibrium of the resulting perturbed game close to our fully stable set. This equilibrium projects to an equilibrium of the (perturbed) subgame, because the subgame is reached with positive probability.

[27] Given any polyhedral restrictions on the strategies of agents in the truncated game, consider in addition some sequence of (polyhedral) restrictions on the strategies of agents in the subgame,

Although only $(T, L)$ is admissible, the unique fully-stable set is the interval from $(T, L)$ to $(T, R)$. (For instance, $(T, R)$ must be included because it is the unique equilibrium when $L$ is perturbed more than $R$ toward $Z$.)

### 3.5. *Stable Equilibria*

Looking at our last example, we see that the reason "full stability" led to a set which was too large, was that a player's choice was allowed to be affected by the perturbation of his own strategies. If we do not want this effect, we have to make the perturbations in a player's payoffs independent of his strategies. A natural way to achieve this is to perturb every pure strategy in the same amount towards the same completely mixed strategy. Clearly, an arbitrary convex combination of pure strategies will then also be perturbed in the same way. Thus the following definition is invariant under addition or deletion of mixed strategies as additional pure strategies (more heuristics can be found in Appendix D).

We will say that a set of equilibria is *stable* in a game $G$ if it is minimal with respect to the following property:

PROPERTY (S): $S$ is a closed set of Nash equilibria of $G$ satisfying: for any $\varepsilon > 0$ there exists some $\delta_0 > 0$ such that for any completely mixed strategy vector $\sigma_1, \ldots, \sigma_n$ ($n$ players) and for any $\delta_1, \ldots, \delta_n$, $(0 < \delta_i < \delta_0)$, the perturbed game where every strategy $s$ of player $i$ is replaced by $(1 - \delta_i)s + \delta_i \sigma_i$ has an equilibrium $\varepsilon$-close to $S$.

REMARKS: 1. This is the same as the definition of full stability, except that instead of being restricted to general polyhedral sets, the players' strategies are restricted to simplices with faces parallel to the faces of the original simplex.

2. If we asked for "some" instead of "any" ($\sigma_1, \ldots, \sigma_n$ and $\delta_1, \ldots, \delta_n$) we would simply get perfect equilibria (cf. footnote 5).

The same arguments used for fully stable sets show that stable sets exist and satisfy the following version of connectedness: *There exists a stable set which is contained in a single connected component of the set of Nash equilibria* and *every generic tree has a stable payoff.* In addition, stable sets are invariant (obvious from the definition).

But stable sets might not satisfy the backwards induction requirement. In the game $\Delta$ (Section 2.8) $\{(T, (\frac{1}{4}, \frac{3}{4})) \cup (T, (\frac{3}{4}, \frac{1}{4}))\}$ is stable; but in its presentation below (Figure 10), the unique sequential equilibrium is $(T, (\frac{1}{2}, \frac{1}{2}))$.

While in the example of Figure 10 the stable set still gives the same payoffs as the sequential equilibrium, in the example shown in Figure 11 (for which we are grateful to Faruk Gul) there is a stable set which gives different payoffs than

converging to the full strategy simplices. This gives a sequence of perturbations of the original game, with equilibria close to our fully stable set. Since each one of those equilibria reaches the subgame with positive probability, any limit point must induce on the subgame its (unique) equilibrium. So any such limit point is an equilibrium of the given perturbation of the truncated game, close to our fully stable set.

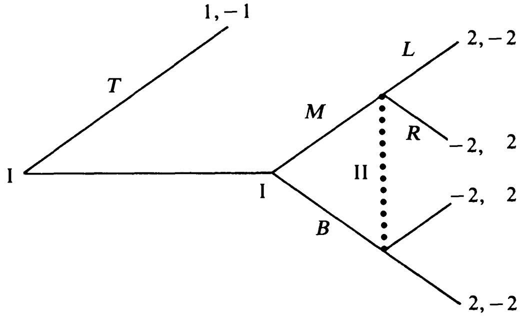


FIGURE 10

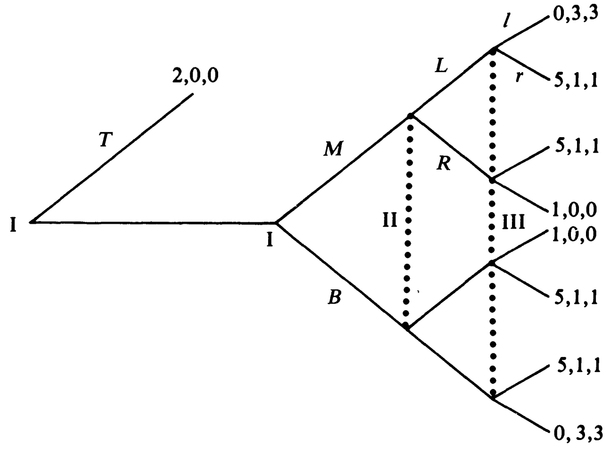


FIGURE 11

the unique sequential equilibrium: It is easy to verify that $\{(T, L, l) \cup (T, R, r)\}$ is a stable set (giving the payoffs (2, 0, 0)), whereas the unique sequential equilibrium is $((0, \frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}))$ with payoffs $(\frac{11}{4}, \frac{5}{4}, \frac{5}{4})$.[28]

But even with our present definition *stable sets are admissible*, because (by minimality) every equilibrium in a stable set is a limit of equilibria of perturbed games, in each of which any dominated strategy becomes strictly dominated.

Stable equilibria also satisfy the following:

[28] Recall (Section 2.8) that we certainly do not want to identify equilibria in different connected components of the set of Nash equilibria. Yet this example shows that with our present definition, stable sets may include points from different connected components:

The set of Nash equilibria consists of three connected components: the sequential equilibrium, and two sets in which player I chooses $T$, and players II and III randomize in such a way that $I$'s expected payoff following either $M$ or $B$ is less than or equal to 2 but in one of those sets the probability of $Ll$ is greater than or equal to that of $Rr$ and in the other it is less than or equal to that of $Rr$. (The sets are disjoint because, if the probability of $Ll$ equals that of $Rr$, then the probabilities of $Lr$ and of $Rl$ must add up to at least $\frac{1}{2}$, and therefore the expected payoff to $I$ must be at least 2.5.) Clearly, $(T, L, l)$ lies in the second component while $(T, R, r)$ lies in the third.

PROPOSITION 6: A. (*Iterated Dominance*) *A stable set contains a stable set of any game obtained by deletion of a dominated strategy.* B. (*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.*

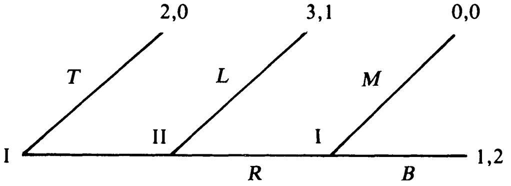
PROOF: Given a perturbation, $G(\varepsilon)$, of the game without the deleted strategy $s$, construct a close-by perturbation $G(\varepsilon, z)$ by first adding $s$ as an additional extreme point in the relevant player's strategy set, then perturbing it like all the other strategies were perturbed in $G(\varepsilon)$, and finally perturbing all of that player's strategies in the amount $z$ towards $s$. Since $G(\varepsilon, z)$ is a perturbation of the original game, it has an equilibrium close to our stable set. Such an equilibrium will give zero weight to $s$ (in case A, because $s$ is strictly dominated in $G(\varepsilon, z)$; in case B, because $s$ is an inferior response in any point close to the stable set). So taking a limit of such equilibria (as $z \to 0$) will give an equilibrium of $G(\varepsilon)$ close to the stable set. *Q.E.D.*

Proposition 6B captures the "forward-induction" logic (Section 2.6) of our basic example $\Gamma(x)$ (Figure 2). Kreps (1984) has used a particular case of it as an "intuitive criterion" to justify the stable equilibria of signalling games, and to show that, in Spence-type signalling games, there is only one stable equilibrium, namely Spence's original separating equilibrium.

REMARK: It seems natural to expect, based on first principles, that *a strategically stable equilibrium must remain so after deletion of a strategy which is an inferior response* (*at that equilibrium*). However, such a requirement cannot be satisfied by a single-valued concept. To see this, consider the perfect information game (payoffs could be perturbed so as to make the tree generic) shown in Figure 12.

The unique backwards induction equilibrium is $(T, R)$, with payoffs "2, 0." So any single-valued concept of "strategically stable equilibrium" must accept $(T, R)$ as the solution of this game. Yet when the inferior response $B$ is deleted, we obtain the game shown in Figure 13, whose unique backwards induction equilibrium (alternatively, the unique admissible equilibrium) is $(M, L)$, with payoffs "3, 1."

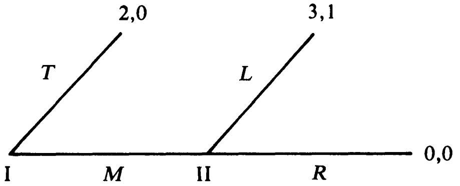
So Proposition 6B can be satisfied only by a set-valued concept. (Note that the unique stable set (of Figure 12) includes $(T, (\frac{1}{2}, \frac{1}{2}))$; and the strategy $B$ is not an inferior response at this equilibrium.)



i.e.,

| | $L$ | $R$ |
|---|---|---|
| $T$ | 2, 0 | 2, 0 |
| $M$ | 3, 1 | 0, 0 |
| $B$ | 3, 1 | 1, 2 |

FIGURE 12



i.e.,

| | $L$ | $R$ |
|---|---|---|
| $T$ | 2, 0 | 2, 0 |
| $M$ | 3, 1 | 0, 0 |

FIGURE 13

Notice that, unlike our previous arguments in favor of a set-valued equilibrium concept, the argument above does not rely on having to satisfy existence.

In summary, then, "stable equilibrium" satisfies existence, a version of connectedness, invariance, admissibility, and iterated dominance. As we have mentioned before, we hope an appropriate modification will, in addition, satisfy connectedness and backwards induction.

### 3.6. *Application of Stable Equilibrium*

We end with a few applications of our results. In each of these applications, we rule out all but one of the candidates for stability by showing that they violate some requirement; by existence, the remaining candidate must then be a stable set.

#### A. *The Previous Examples*

In the game $\Gamma(x)$ (Figure 2), the set of Nash equilibria consists of two connected components: the singleton $(M, L)$, with payoffs "3, 3," and the interval from $(T, R)$ to $(T, (2-x)/(3-x),\ 1/(3-x))$, with payoffs "2, 2." Since the second component disappears after iterated elimination of dominated strategies ($B$, then $R$), it cannot contain a stable set. So $(M, L)$, i.e. "3, 3," is the unique stable set.

The same analysis applies for the variant of $\Gamma(0)$ in Figure 6.[29]

In the games $\Omega$ and $\Delta$ (Sections 2.7 and 2.8, respectively) as well as in Figures 8 and 9 and the example of Appendix B, the set of Nash equilibria consists of a single connected component. But while in the first two examples we cannot rule out any point within the component, in the latter three, all equilibria but one are eliminated because of inadmissibility.

#### B. *An Example from Information Economics*

We conclude with an example, due to Kreps, which captures the central point of many recent contributions to information economics (see Kreps (1984) for a

[29] We rely here on the following fact: a stable set contains a stable set of any truncated game obtained by replacing any zero-sum subgame by its value (cf. Remark 1 in Section 3.4).

discussion). It is the basic ingredient (i.e., a single stage) of the "chain store paradox", and can be described as follows:

First, Nature chooses one of two players, Weak or Strong, with probabilities .1 and .9, respectively.

Next, the chosen player sends either a "strong" or a "weak" signal. A true signal is costless, whereas a false signal costs 1 unit.

Finally, a third player (the Entrant), who is only informed of the signal, must decide whether to fight or to retreat. If he fights, he will win or lose 1 unit depending on whether his opponent was Weak or Strong. The opponent, on the other hand, will lose 2 whenever there is a fight.

The extensive form is given in Figure 14 (where 0 denotes Nature, the arrows indicate the order of play, and payoffs are indicated in the following sequence: Weak, Strong, Entrant).

It is easily verified that the set of Nash equilibria consists of two connected components (within each of which the equilibria differ only off the equilibrium path):

(1) The chosen player sends a strong signal; the Entrant retreats if the signal is strong, and fights with probability greater than or equal to $\frac{1}{2}$ if the signal is weak.

(2) The chosen player sends a weak signal; the Entrant retreats if the signal is weak, and fights with probability greater than or equal to $\frac{1}{2}$ if the signal is strong.

Whereas equilibria in the first component appear sensible, those in the second component do not. (The Entrant's prior probability of his opponent being strong is 90 per cent, but after hearing a strong signal he acts as if his posterior dropped to at most 50 per cent.) Yet none of the previously known solution concepts could distinguish between these two components (i.e., all the Nash equilibria in this example are sequential, perfect, proper, proper after iterated elimination of

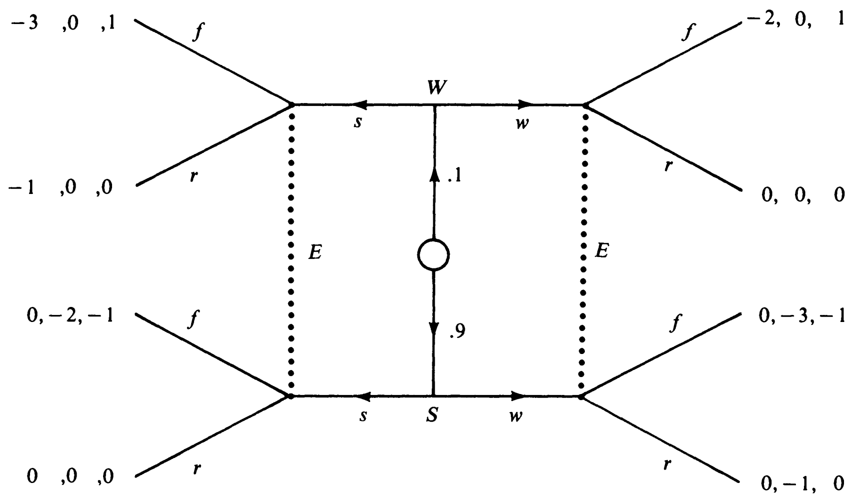


FIGURE 14

dominated strategies, etc....). On the other hand, "stability" does distinguish between them: (1) is stable, whereas (2) is not.

To see this, note that sending a strong signal is an inferior response for Weak in any equilibrium of component (2) (because the resulting payoff will be between $-3$ and $-2$, which is less than the equilibrium payoff, 0). But deleting the possibility for Weak to send a strong signal will make component (2) unstable (because then the Entrant's strategies in (2) will be inadmissible, his choices after a strong signal being dominated by the choice $r$). Thus, by Proposition 6B, (2) does not contain a stable set; by existence, (1) does.

*Graduate School of Business Administration, Harvard University, Boston, MA 02163*

*CORE, and Université Catholique de Louvain, Louvain-la-Neuve, Belgium*

*Manuscript received March, 1983; final revision received December, 1985.*

## APPENDIX A

PROPOSITION 0: *For any tree, and for any proper equilibrium of its normal form, there exist equivalent*[30] *behavioral strategies which form a sequential equilibrium.*

PROOF: Let $x = \lim x_\varepsilon$ be a proper equilibrium where the $x_\varepsilon$ are $\varepsilon$-proper equilibria. Given a tree, let $\sigma_\varepsilon$ be the($n$-tuple of) behavioral strategies equivalent to $x_\varepsilon$, and let $\mu_\varepsilon$ be the vector of conditional probabilities that they imply on information sets. Extract a subsequence along which all those objects converge. We have to show that $\sigma = \lim \sigma_\varepsilon$ is such that each agent maximizes his payoff given $\mu$ and given the strategies of the other agents.

Assume the contrary. Then there is some player, say 1, and a last information set for him, say $J$, such that $\sigma^1$ assigns positive probability to a move in $J$, say $L$, whose expected payoff (given $\mu$ and $\sigma$) is less than that of another move, say $R$. Since player 1's agents in information sets after $J$ are assumed to be maximizing 1's payoff, it follows that the expected payoff to player 1 (starting in $J$ and given $\mu$ and $\sigma^2, \dots, \sigma^n$) of choosing $R$ and then continuing as in $\sigma^1$, is larger than that of choosing $L$, regardless of the continuation. Clearly, the same is true given $\mu_\varepsilon$ and $\sigma_\varepsilon^2, \dots, \sigma_\varepsilon^n$, provided $\varepsilon > 0$ is sufficiently small.

It follows that every normal form strategy of 1 that does not avoid $J$ and chooses $L$ in $J$ has smaller expected payoff, given $x_\varepsilon^2, \dots, x_\varepsilon^n$ than a modification of that strategy that chooses $R$ and then continue as in $\sigma^1$. Since $x_\varepsilon$ is $\varepsilon$-proper, $x_\varepsilon^1$ assigns the first strategy probability less than or equal to $\varepsilon$ times the probability of the second strategy. It follows that $\sigma_\varepsilon^1$ assigns to $L$ probability of at most $k\varepsilon$, where $k$ is the number of (normal-form) strategies of 1. Letting $\varepsilon \to 0$ we see that $\sigma^1$ assigns to $L$ zero probability, a contradiction. *Q.E.D.*

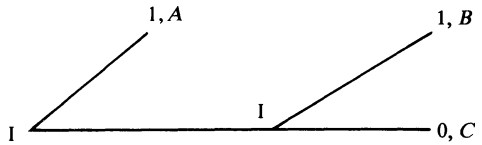


FIGURE 15

[30] "Equivalent" in the sense of Kuhn's theorem. Remember that we identify such equivalent strategies (Section 2.5).

Note that the proposition above is no longer true when "sequential" is replaced by "perfect." For example, "1, $B$" is proper in the normal form

| |
|---|
| 1, $A$ |
| 1, $B$ |
| 0, $C$ |

but is not perfect in the tree shown in Figure 15.

## APPENDIX B

Proposition 1: *The set of Nash equilibria of every game has finitely many connected components. At least one of them is such that for any equivalent game* (*i.e. having the same reduced normal form*) *and for any perturbation of the normal form of that game, there is a Nash equilibrium close to this component.*

(Mas Collel (private communication) has suggested an alternative proof of this proposition, which uses a similar argument as the one below but applied to the mapping used by Nash in his original existence theorem rather than to the mapping $\bar{p}$ of Theorem 1.)

Remark: In the statement above, we think of a connected component of equilibria in one game as also being a connected component of equilibria in any equivalent game. (There is a natural mapping from the equilibria of a game to the equilibria of its reduced normal form: if one pure strategy is equivalent to a convex combination of other pure strategies, simply replace its weight by the appropriate weights on those other strategies. This mapping between equilibria induces a one-to-one mapping between connected components of equilibria.)

Proof: (For notation see Section 3.2.) The picture of the proof is as follows: If a connected component of the part of the rubber sphere $\bar{E}$ that lies above some game $\gamma_0$ had a neighborhood that did not project onto some neighborhood of $\gamma_0$, then that component could be pulled away from the vertical above $\gamma_0$ by a small deformation. So if all connected components were such, we would have a deformation of the rubber sphere with a hole above $\gamma_0$, which is clearly impossible.

Formally, we first note that the equilibrium set of any game consists of a finite number of connected components. This follows from a theorem of van der Waerden (1939, Satz 1, p. 123), that a compact set consisting of the solutions to a finite system of algebraic inequalities has a finite triangulation (i.e., is homeomorphic to a finite union of compact polyhedra).

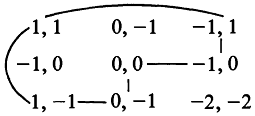
Let then $\gamma_0 \in \Gamma$ be a given normal form, and let $C_1, \ldots, C_n$ denote the different connected components of the Nash equilibria of $\gamma_0$. We first want to show that $\exists i$ such that any game in a neighborhood of $\gamma_0$ has equilibria close to $C_i$.

If not let $0_i \subseteq \Sigma$ be open neighborhoods of the $C_i$ with disjoint closures such that, $\forall i$, $\forall \varepsilon > 0$, $\exists \gamma_{i,\varepsilon}$ such that $\|\gamma_0 - \gamma_{i,\varepsilon}\| < \varepsilon$ and $Eq(\gamma_{i,\varepsilon}) \cap 0_i = \varnothing$. By upper-semi-continuity of the equilibrium correspondence, choose $\varepsilon$ such that, if $\|\gamma - \gamma_0\| \leqslant \varepsilon$, then $Eq(\gamma) \subseteq \bigcup_i 0_i$; letting $\gamma_i = \gamma_{i,\varepsilon}$ we have $Eq(\gamma_i) \subseteq \bigcup_{j \neq i} 0_j$.

Use now Tietze's extension theorem to get a continuous function $h : E \to \Gamma$ such that $\forall i$, $h = \gamma_i$ on $E \cap (\Gamma \times \bar{0}_i)$ and $\|h(e) - \gamma_0\| \leqslant \varepsilon$ everywhere. Then we have $h(e) \neq p(e) \forall e$; thus $\overline{p-h} : \bar{E} \to \bar{\Gamma}$ is not onto because it does not cover zero (its value at $\infty$ being $\infty$ because $h$ is bounded). But $\overline{p-h}$ is homotopic to $\bar{p}$ ($\overline{p-th}$ is a homotopy on $\bar{E}$—the boundedness of $h$ insures the continuity at $\{\infty\}$), so that, by Theorem 1, it is homotopic to the identity (modulo the homeomorphism), and therefore, by Brouwer's theorem, it is onto: we have a contradiction.

Consider now what happens if one adds new pure strategies to the game, equivalent to mixed strategies in the original game. The new game still has such a connected component, and any such component in the new game is mapped to such a component in the old game (as noted before, the mapping from equilibria in the new game to equilibria in the old game induces a one to one mapping between connected components). As one keeps adding strategies, the (finite) set of such components keeps shrinking, but stays nonempty. This implies that the original game has a connected component of equilibria that is stable in the required sense in any game obtained by adding (and/or deleting) equivalent pure strategies. *Q.E.D.*

Remark: The statement of the theorem cannot be strengthened to say that there exists a convex (or even a contractible) set of equilibria with the above stability properties: in the following example, the only set with those properties is homeomorphic to a circle (both in strategy space and in payoff space).



| | | |
|---|---|---|
| 1, 1 | 0, −1 | −1, 1 |
| −1, 0 | 0, 0 | −1, 0 |
| 1, −1 | 0, −1 | −2, −2 |

The set of Nash equilibrium is exactly the circle depicted. We claim that the whole circle is needed if we want to allow for any perturbation of the normal form payoffs, i.e. that no closed proper subset of the circle has the required stability property: for all Nash equilibria, except those where one player mixes—strictly—between his first and last strategies, there exists a perturbation of this normal form with a unique Nash equilibrium in the vicinity of that point. For example, the following perturbed game has a single equilibrium, $((\varepsilon/1+\varepsilon,\ 1/1+\varepsilon,\ 0),\ (0, \frac{1}{2}, \frac{1}{2}))$:

| | | |
|---|---|---|
| $1, 1-\varepsilon$ | $\varepsilon, -1$ | $-1-\varepsilon, 1$ |
| $-1, -\varepsilon$ | $-\varepsilon, \varepsilon$ | $-1+\varepsilon, -\varepsilon.$ |
| $1-\varepsilon, -1$ | $0, -1$ | $-2, -2$ |

For the remaining points, one has to add say a column equivalent to the mixture $(\alpha, 1-\alpha)$ on the first and last columns and to perturb the resulting normal form.

## APPENDIX C

Proposition (Kreps and Wilson): *For generic extensive games, the set of equilibrium probability distributions on endpoints is finite. More precisely, for any tree there is a nontrivial polynomial with integer coefficients, such that any vector of endpoint payoffs for which the tree has infinitely many equilibrium distributions, is a zero of that polynomial.*

Proof: By Kuhn's theorem, it is sufficient to consider behavioral strategies. Erasing the part of the tree that is reached with zero probability, it is sufficient to prove the finiteness of the number of distributions on endpoints arising from completely mixed behavioral equilibria.

In what follows, all the games we consider will have the same tree (but different endpoint values). So we can identify each game with its vector of endpoint payoffs.

Given such a game $G$, define a normalized game $\tilde{G}$ in the following way:

First, consider the set $\Psi_1^n$, of all those information sets of player $n$, in which all his moves are last moves for him (at least one such information set exists in a finite game of perfect recall). For every $I \in \Psi_1^n$, and for every move $k > 1$, add a constant to player $n$'s payoff at all the endpoints that are successors of the information set $I$ and the move $k$, in such a way that the sum of player $n$'s payoffs over all those endpoints will be zero.

Next, consider the set $\Psi_2^n$, of all those information sets of player $n$, in which all his moves are either last moves for him or else are last moves for him before an information set in $\Psi_1^n$. For every $I \in \Psi_2^n$, and for every move $k > 1$, add a constant to player $n$'s payoff at all the endpoints that are successors of the information set $I$ and the move $k$, in such a way that the sum of player $n$'s payoffs over all those endpoints will be zero. Notice that, since the game has perfect recall, this second step will add the same constant to player $n$'s payoff at all those endpoints that follow the same information set $I \in \Psi_1^n$.

Next, define $\Psi_3^n$, $\Psi_4^n$, etc..... $\tilde{G}$ is obtained from $G$ by carrying out this procedure for each player $n$.

Now, given such a normalized game $\tilde{G}$ and a completely mixed equilibrium of $G$, one can reconstruct $G$ in a unique way: for any information set, the expected payoff in $G$ must be the same for all moves in that information set; so for $I \in \Psi_1^n$, the corresponding constant in the above-described procedure can be recovered by subtracting player $n$'s expected payoff in $\tilde{G}$ following move 2 from his expected payoff following move 1, etc.....

We thus have a rational $C^1$ mapping, $\phi$, from pairs $(H, x)$—where $H$ is a normalized game and $x$ is a completely mixed $n$-tuple of behavioral strategies—to games, such that $\phi(\tilde{G}, x) = G$ whenever $x$ is a completely-mixed equilibrium of $G$. So the set of completely-mixed equilibria of $G$ is finite when $\phi^{-1}(G)$ is so.

Clearly, the dimensions of $\{(H, x)\}$ and of $\{G\}$ are the same. It follows that $\phi^{-1}(G)$ consists of finitely many points except when $G$ is a critical value of $\phi$. (For regular $G$, the set is discrete since $\phi$ is $C^1$; it has a finite triangulation by algebraicity (Van der Waerden, loc. cit.); hence it is finite.) But the set of critical values of a rational $C^1$ map from $R^n$ to $R^n$ is contained in the set of zeros of a nontrivial polynomial (e.g., using Sard's theorem and the previously cited result of van der Waerden). Since the rational map has integer coefficients, it follows that the polynomial can also be chosen so.

*Q.E.D.*

## APPENDIX D

### *Some Heuristics behind the Definition of Stability*

*In a two-person game, any admissible equilibrium point* $(\sigma, \tau)$ *is normal-form perfect*: since $\sigma$ is undominated, there exists a completely mixed strategy $y$ of player II such that $\sigma$ is a best reply against $y$. Similarly $\tau$ is a best reply against some completely mixed $x$. Thus, for any $\varepsilon > 0$, $\sigma$ (resp. $\tau$) is a best reply against $\tau_\varepsilon = (1-\varepsilon)\tau + \varepsilon y$ (resp. $\sigma_\varepsilon = (1-\varepsilon)\sigma + \varepsilon x$), and thus $(\sigma_\varepsilon, \tau_\varepsilon)$ forms an $\varepsilon$-perfect equilibrium. This proves our claim. An $n$-person analog of this statement is contained in Selten's criterion of substitute-perfectness (see Selten (1975)).

Since perfectness in the tree is just perfectness in the agent normal form, this suggests that in some sense perfectness is just sequentiality plus admissibility (in the tree).

Thus we see that this same idea of using perturbed games yields our two requirements of backwards-induction and of admissibility (the first step towards iterated dominance), and that the admissibility requirement alone already leads to this idea. (This strongly suggests phrasing a definition of stability in terms of perturbations of the agent-normal form. "Full stability" is a definition of this type (see the proof of Proposition 4). However, as explained in Section 3, one should replace it by a definition that considers only normal-form perturbations, which are the ones that arise directly from the admissibility idea.)

We now wish to show that our definition of stable equilibrium arises naturally from the requirements of perfection and invariance: in the special case of an equilibrium with positive weight on every best reply, we claim that perfectness in every equivalent tree implies stability. (Appendix E shows—at least with sequentiality—that such a restriction may be necessary.)

Indeed, if such an equilibrium $\sigma^*$ is perfect in every equivalent tree, and if $\sigma$ is any vector of completely mixed strategies, first define $\tilde{\sigma}$ to be some convex combination of $\sigma^*$ and $\sigma$ so close to the equilibrium $\sigma^*$ as to yield a higher expected payoff than any pure strategy which is not a best reply. By adding $\tilde{\sigma}$ as an additional pure strategy, and drawing an appropriate tree, like the one shown in Figure 16, the corresponding agent 2 will thus have to use, in any $\varepsilon$-perfect equilibrium close to $\sigma^*$, the additional pure strategy $\tilde{\sigma}$ rather than the inferior strategies, so that the relative weights on all inferior strategies will be essentially (when $\varepsilon \to 0$) determined by the vector $\sigma$: we thus have, for arbitrary $\sigma$, an equilibrium of the $\sigma$-perturbed game close to $\sigma^*$.

## APPENDIX E

### *On the Inadequacy of our Requirements as Axioms*

Let us show that even asking for equilibria which are sequential in every tree having the same reduced normal form would not guarantee strategic stability. Consider the game shown in Figure 17, whose unique backwards induction equilibrium is "3, 3". Let us now add six dummy players, one per box of the matrix, where each dummy player gets 1 in that box and 0 otherwise. Clearly, the addition of the dummies ought not to impact strategic stability, so "3, 3" should still be the only strategically stable equilibrium of the new game, $G$. But, because of the dummies, any game tree with the same reduced normal form as $G$ is essentially one in which both players move simultaneously, i.e., at no relevant information set is one player informed of previous moves by the other. In such a game, every admissible Nash equilibrium is sequential, in particular "2, 2". So the "bad" equilibrium "2, 2" is sequential in every game tree having the same reduced normal form as $G$.

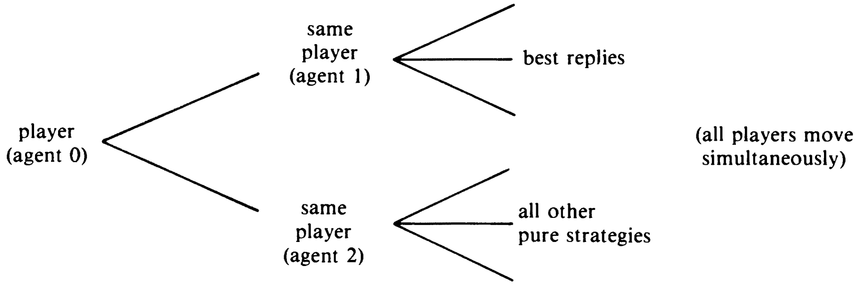
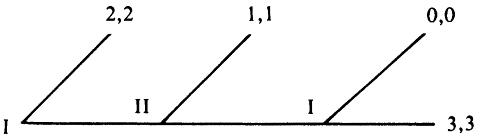


FIGURE 16



| | |
|---|---|
| 2, 2 | 2, 2 |
| 1, 1 | 3, 3 |
| 1, 1 | 0, 0 |

FIGURE 17

We see then that even asking for sequentiality in any tree obtained by application of the six inessential transformations is insufficient for strategic stability. One may be tempted to conclude that the source of the difficulty is that our list of transformations is incomplete. One could then try to define an additional transformation which would eliminate dummy players from a game. (However, while in examples like the one above it is obvious how to identify dummies, in more complicated games the identification of dummies may be difficult.[31])

Our feeling, however, is that the source of the difficulty is in the use of a concept like sequential equilibrium. While sequentiality, invariance, dummy properties, etc., are reasonable properties against which a proposed solution concept may be checked, they cannot serve as a definition or an axiom: one would always find further requirements that are violated in some cases. Presumably, a correct definition or axiom system should involve only rationality criteria (like admissibility) about the game itself as opposed to criteria (like invariance) about the solution correspondence. In addition, the criteria should be phrased in purely decision theoretic terms—e.g., depend only on the best reply correspondence—instead of depending on the tree like sequential equilibrium. Such a definition would automatically yield the required properties of the correspondence (like invariance, dummy properties, backwards induction properties like BI1, correct behavior under deletion of dominated strategies, etc).

## REFERENCES

ARROW, K. J. (1951): "Alternative Approaches to the Theory of Choice in Risk-Taking Situations," *Econometrica*, 19, 404-437.

DALKEY, N. (1953): "Equivalence of Information Patterns and Essentially Determinate Games," in *Contributions to the Theory of Games*, Vol. 2. Princeton: Princeton University Press, pp. 217-245.

[31] Dummies are hard to define—say two sets of players $M$ and $N$ could be defined as being mutually dummies if they form a partition of the player set and if each player's payoff is the sum of a function of the strategies in $M$ and a function of the strategies in $N$. But they should continue to be considered so even if there were higher degree dummies whose payoffs depended nonseparably on $M$ and $N$, or if this game were only a subgame of some bigger game where $M$ and $N$ were not separable, etc.

FORGES, F. (1984): "An Approach to Communication Equilibrium," Discussion Paper 8435, CORE.

KALAI, E., AND D. SAMET (1984): "Persistent Equilibria," *International Journal of Game Theory*, 13, 129-144.

KOHLBERG, E., AND J. F. MERTENS (1982): "On the Strategic Stability of Equilibria," Discussion Paper 8248, CORE.

KREPS, D. (1984): "Signalling Games and Stable Equilibria," mimeo, Stanford University.

KREPS, D., AND R. WILSON (1982): "Sequential Equilibria," *Econometrica*, 50, 863-894.

KUHN, H. (1953): "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games*. Vol. 2. Princeton: Princeton University Press, pp. 193-216.

LUCE, R. D., AND H. RAIFFA (1957): *Games and Decisions*. New York: John Wiley and Sons.

MYERSON, R. B. (1978): "Refinement of the Nash Equilibrium Concept," *International Journal of Game Theory*, 7, 73-80.

NASH, J. (1951): "Non-Cooperative Games," *Annals of Mathematics*, 54, 286-295.

SELTEN, R. (1965): "Spieltheoretische Behandlung eines Oligopolmodel mit Nachfragetragheit," *Zeitschrift für die Gesamte Staatswissenschaft* 12, 301-324 and 667-689.

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

SMITH, J. MAYNARD, AND G. A. PARKER (1976): "The Logic of Asymmetric Contests," *Animal Behavior*, 24, 159-175.

THOMPSON, F. (1952a): "Equivalence of Games in Extensive Form," RM 759, The Rand Corporation.

——— (1952b): "Behavior Strategies in Finite Games," RM 769, The Rand Corporation.

TSUN, W. W., AND J. JIA-HE (1962): "Essential Equilibrium Points of $n$-Person Non-Cooperative Games," *Scientia Sinica*, 10, 1307-1322.

VAN DAMME, E. (1984): "A Relation Between Perfect Equilibria in Extensive Form Games and Proper Equilibria in Normal Form Games," *International Journal of Game Theory*, 13, 1-13.

VAN DER WAERDEN, B. L. (1939): *Einführung in die Algebraische Geometrie* (Second Edition 1973). Berlin-Heidelberg-New York: Springer Verlag.

## References

**Alternative Approaches to the Theory of Choice in Risk-Taking Situations**
Kenneth J. Arrow
*Econometrica*, Vol. 19, No. 4. (Oct., 1951), pp. 404-437.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28195110%2919%3A4%3C404%3AAATTTO%3E2.0.CO%3B2-F

**Sequential Equilibria**
David M. Kreps; Robert Wilson
*Econometrica*, Vol. 50, No. 4. (Jul., 1982), pp. 863-894.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198207%2950%3A4%3C863%3ASE%3E2.0.CO%3B2-4

**Non-Cooperative Games**
John Nash
*The Annals of Mathematics*, 2nd Ser., Vol. 54, No. 2. (Sep., 1951), pp. 286-295.
Stable URL:
http://links.jstor.org/sici?sici=0003-486X%28195109%292%3A54%3A2%3C286%3ANG%3E2.0.CO%3B2-G