GAMES AND ECONOMIC BEHAVIOR 8, 20–55 (1995)

Learning in Extensive-Form Games I. Self-Confirming Equilibria

DREW FUDENBERG

Department of Economics, Harvard University

AND

DAVID M. KREPS

Graduate School of Business, Stanford University; and Berglas School of Economics, Tel Aviv University

Received August 24, 1993

A group of individuals repeatedly plays a fixed extensive-form game, using past play to forecast future actions. Each (asymptotically) maximizes his own immediate expected payoff, believing that others' play corresponds to the historical frequencies of past play. Because players observe only the path of play in each round, they may not learn how others act in parts of the game tree that are not reached infinitely often. Hence, differences and correlations in beliefs about out-of-equilibrium actions may persist indefinitely. The stable points of these learning processes are self-confirming equilibria, a weaker solution concept than Nash equilibria. *Journal of Economic Literature* Classification Numbers: C72, D83.

1. INTRODUCTION

In a Nash equilibrium, each player's beliefs about the strategies of his opponents are exactly correct. This paper investigates the idea that players may come to have correct, or at least approximately correct, beliefs as the result of a process of learning from experience. The idea, as in the literature on fictitious play, is that players play the game repeatedly (or observe repeated play prior to their own turn to play) and expect that the current play of their opponents will resemble the way the opponents have

0899-8256/95 $6.00

played in the past. It is easy to prove that if players observe the strategies chosen by their opponents and their beliefs come to resemble the empirical distribution, then if behavior converges to a steady state (at least, in pure strategies), the steady state will be a Nash equilibrium. Thus the focus of interest in the literature related to fictitious play has been on questions of whether behavior will converge and, to a lesser extent, on the prospects (and modes) of convergence to mixed strategy equilibria.[1]

This paper studies learning processes in the general style of fictitious play under the assumption that players observe only the actions that are actually played in a given extensive-form game, and not the actions that their opponents would have chosen at information sets that were not reached in the course of play.[2] Thus repeated observations of opponents' play need not lead to correct beliefs about their full strategies, which prescribe actions at all information sets. All that can be expected if play converges is that players come to have correct beliefs about behavior at information sets that lie along the path of play. Two players might persistently maintain different beliefs about how a third player would respond to a deviation from the path of play, and one player might persist in correlated beliefs concerning the actions of other players whose information sets lie off the path of play.

Since both of these phenomena can support non-Nash outcomes, learning processes need not lead to Nash equilibrium absent some reason (such as experimentation with off-path actions or restrictions on the prior beliefs) for players to have correct beliefs about off-path play. Rather, the set of possible stable points is the set of self-confirming equilibria.[3]

Throughout we work in the style of the literature on bounded rationality. That is, we exogenously specify behavior rules for the players, rather

[1] There are many papers in this literature; see Fudenberg and Kreps (1993) for a partial bibliography.

[2] In this paper we restrict attention to extensive-form stage games and the problems raised by off-the-path information sets, but similar issues arise whenever players observe something less than the full (pure) strategies their opponents have chosen. For example, players might observe only their own actions and payoffs.

[3] The basic idea of a self-confirming equilibrium—that players need have correct beliefs only about those elements of play that they observe—appears in the literature as early as Hahn's (1977) notion of conjectural equilibrium. Recent formalizations and analyses in a game-theoretic context include Battigalli (1987), Battigalli and Guaitoli (1988), Rubinstein and Wolinsky (1990), Fudenberg and Levine (1993a,b), and Kalai and Lehrer (1993a,b,c). Fudenberg and Levine (1993b) and Kalai and Lehrer (1993a) concern explicit learning models; their results present an interesting contrast with the results given here.

Our specific definition, and the name we use, is taken from Fudenberg and Levine (1993a), with one simplification: They study a learning model with a large number of players 1, players 2, etc., which leads them to a definition of self-confirming equilibrium that allows the (off-the-path) beliefs of different players 1 to differ. The definition we use corresponds to what they call unitary beliefs.

than deriving behavior from utility maximization. Moreover, although players are trying to use past observations to learn their opponents' play, their learning rules need not be objectively valid. In particular, they act as if the process is converging to a steady state even though this need not be the case. Finally, as in the traditional literature on fictitious play, players treat their opponents' future play as exogenous, even though in the processes we consider this is false; a patient player might be able to do better by trying to "teach" his opponents that he will play in a particular way.

The context for the analysis is given in Section 2. Sections 3, 4, and 5 adapt asymptotic empiricism, asymptotic myopia, and stability from the context of Fudenberg and Kreps (1993)—where stage-game (pure) strategies are fully observable—to the current context. In particular, because players can affect what they observe about the actions of others, we allow conscious experimentation.

Section 6 reviews the concept of self-confirming equilibrium, providing illustrative examples and a characterization for two-player games.

Section 7 contains our basic results: In any model where beliefs and behavior satisfy our assumptions, asymptotic steady states must be self-confirming equilibria, and any self-confirming equilibrium is an asymptotic steady state for some model that meets our assumptions.

In Section 7 we discuss asymptotic stability of strategy profiles. We extend these results slightly in Section 8, to asymptotic stability of outcomes, or probability distributions induced by strategy profiles over terminal nodes: Only outcomes that correspond to self-confirming equilibria can be asymptotic steady states.

Our analysis raises the question of when players might be expected to learn enough about the off-path play of their opponents to preclude convergence to a non-Nash outcome. Section 9 presents one way that this could happen, viz., if players consistently "tremble" in the sense of Selten (1975). In a companion paper, Fudenberg and Kreps (1994), we investigate whether and when conscious experimentation by the players rules out non-Nash steady states.

2. PRELIMINARIES

A finite I-player extensive-form game with perfect recall, called the *stage game*, is played repeatedly by the same I players, at dates $t = 1, 2, 3, \ldots$.

The game tree for the stage game is denoted by V: $\prec$ denotes precedence, $Z \subseteq V$ is the subset of terminal nodes, and $X = V \backslash Z$ is the set of action nodes. The information sets $h \in H$ partition X: $h(x)$ is the information set containing x, $i(h)$ is the player who moves at h, H^i is the set of

player i's information sets, and $H^{-i} = H \backslash H^i$ denotes the information sets of i's opponents. The feasible actions at h are denoted by $A(h)$. Actions are labeled so that $A(h) \cap A(h') = \varnothing$ for $h \neq h'$; $h(a)$ denotes the information set at which a is feasible. The set of feasible actions for player i, or $\cup_{h \in H^i} A(h)$, is denoted by A^i; A^{-i} denotes the set of feasible actions for i's opponents. All of Nature's moves (if any) are placed at the start of the tree, so that each move by Nature corresponds to an initial node of the tree. The set of initial nodes is denoted by W; we suppose that the objective distribution ϕ over these initial nodes is strictly positive and is known to all the players.[4] Player i's payoff if terminal node z is reached is $u^i(z)$. Player i knows his own payoff function u^i; we are agnostic whether players know the payoff functions of their opponents, but some of our assumptions and examples will seem more natural if we suppose that they do not.

The set of pure strategies for player i in the stage game is denoted by S^i; $s^i \in S^i$, if s^i: $H^i \to A^i$ such that $s^i(h) \in A(h)$. The space of mixed (behavior) strategies for i is denoted Π^i; $\pi^i \in \Pi^i$ if π^i: $H^i \to \Delta(A(h^i))$, the set of probability distributions over $A(h^i)$.[5] Pure and behaviorally mixed strategy profiles are denoted by s and π and are elements of $S = \prod_i S^i$ and $\Pi = \prod_i \Pi^i$, respectively. Pure and behaviorally mixed strategy profiles for all players except i are denoted by s^{-i} and π^{-i}, coming from the sets $S^{-i} = \prod_{j \neq i} S^j$ and $\Pi^{-i} = \prod_{j \neq i} \Pi^j$.

Each strategy profile π induces a probability distribution $P(\cdot|\pi)$ over the terminal nodes, computed under the assumption that each player's behavior is independent of the behavior of others. The support of $P(\cdot|\pi)$ is denoted by $\overline{Z}(\pi)$; $\overline{X}(\pi)$ and $\overline{H}(\pi)$ will denote the set of all non-terminal nodes and information sets that have positive probability under π, respectively.

Each play of the game at a given date results in a particular terminal node $z \in Z$ being reached, so the *history* at the beginning of date t is an element $\zeta_t = (z_1, \ldots, z_{t-1})$. (For $t = 1$, ζ_1 is used conventionally to denote the initial [informationless] history.) We assume that all players observe the outcome at the end of each round, so that all players known ζ_t at the start of round t.[6] We use ζ to denote an infinite history of play $(z_1, z_2, \ldots)$, $\mathfrak{Z}$ to denote the space of all infinite histories (so that $\mathfrak{Z} = (Z)^\infty$), and

[4] As long as ϕ is known to players, putting all of nature's moves at the start of the tree is without loss of generality. If we had players learning nature's probabilities, complications arise; see footnote 6 following.

[5] We reserve Σ^i for the space of mixed strategies for player i, i.e., $\Sigma^i = \Delta(S^i)$.

[6] This is why the placement of nature's moves matters when players are learning ϕ. Placing nature's moves at the start implies that players will see all of nature's moves in each round; if some of nature's moves are placed in an unreached portion of the tree, they will not be observed.

$\mathcal{Z}_t$ to denote the space of all histories up to time t (so that $\mathcal{Z}_t = (Z)^{t-1}$).[7]

We implicitly assume that all players know from the outset the extensive-form structure of the stage game. Each player assumes that his own behavior π^i and the behaviors π^j of each of his rivals are independent, so that if i plays according to π^i and is certain that (each) rival j plays according to π^j, then the outcome of the game will be the terminal node z with probability $P(z|\pi)$.

3. Beliefs, Assessments, and Asymptotic Empiricism

The behavior of each player at any date t will depend on the history of play up to that date and, more particularly, on what each player *believes* to be the joint strategies being chosen by his rivals. While a very general formulation would specify each player's probability assessment over strategy selection rules of his opponents for each and every future date, we will make do keeping track of each player's beliefs about the joint strategies of his rivals *for the current round of play*, as a function of past play.

3.1. Beliefs and Assessments

Player i's *beliefs* about his opponents' play will be represented by a probability measure γ^i over the set Π^{-i} of behavior strategies for player i's opponents. That is, for $\Lambda \subseteq \Pi^{-i}$, $\gamma^i(\Lambda)$ is i's probabilistic beliefs that his rivals' strategy profile will be some (independently mixed) strategy profile π^{-i} contained in the set Λ. As γ^i is not necessarily a product measure over Π^{-i}, player i's beliefs can reflect correlation in his opponent's strategy selection.

Given beliefs γ^i, player i's *assessment* μ^i about what will happen if he plays strategy π^i is obtained by integrating $P(\cdot|\pi^i, \pi^{-i})$ with respect to player i's beliefs:

$$\mu^i(z|\pi^i, \gamma^i) = \int_{\Pi^{-i}} P(z|\pi^i, \pi^{-i})\gamma^i[d\pi^{-i}]. \tag{3.1}$$

That is, the assessment μ^i is the marginal distribution over terminal nodes induced by player i's beliefs and player i's intentions.

Suppose that i has a single rival, j, who must choose between two pure strategies, and suppose that i *assesses* that it is equally likely that j will choose either pure strategy. Having a formalism for i's *beliefs* allows us to distinguish between the case where i *believes* with certainty that j is

[7] In general, subscripts will denote time and superscripts will denote players. The exceptional case of Z to the power $t-1$ is indicated by $(Z)^{t-1}$.

playing the corresponding mixed strategy and the case where i believes that there is probability 1/2 that j will play one or the other pure strategy. These two situations are equivalent in a static setting, but may have very different implications about what i will learn from observing j.[8]

3.2. *Belief Rules*

We imagine that for each player i, date t, and partial history ζ_t, i holds *beliefs* $\hat{\gamma}_t^i(\zeta_t)$ about the strategy profile that his rivals are about to play. The term *beliefs rule* will be used to refer to a full specification of i's beliefs, as a function of time and history, denoted by $\hat{\gamma}^i$. (The hat is used to distinguish between i's full array of beliefs, for all dates and histories, and a particular probability distribution on Π^{-i}, that is, a single $\gamma^i = \hat{\gamma}_t^i(\zeta_t)$.)

A special case is where i uses Bayesian inference: player i views the successive selections of his rivals as i.i.d. draws from some fixed but (to player i) unknown strategy profile,[9] and, relative to this prior assessment, player i uses Bayes' rule to update his beliefs. In symbols, if $\hat{\gamma}_t^i(\zeta_t)$ gives i's beliefs at the start of round t, player i uses (pure) strategy s^i in this round, and the resulting outcome is z, then i's posterior beliefs are given by

$$\hat{\gamma}_{t+1}^i(\zeta_t, z)(\Lambda) = \frac{\int_\Lambda P(z|s^i, \pi^{-i})\hat{\gamma}_t^i(\zeta_t)(d\pi^{-i})}{\int_{\Pi^{-i}} P(z|s^i, \pi^{-i})\hat{\gamma}_t^i(\zeta_t)(d\pi^{-i})}, \tag{3.2}$$

for $\Lambda \subseteq \Pi^{-i}$ (assuming the denominator is positive.)

3.3. *Asymptotically Empirical Beliefs*

We confine attention to belief rules that are *asymptotically empirical*, in the rough sense that for any player i and information set $h \in H^j$, $j \neq i$, if h is hit fairly often as time passes, then i becomes more and more certain that j's choice of strategy entails a choice of action at h that (asymptotically) equals the empirical frequency distribution of j's choices at h.

The following notational definitions are required to make this precise:

(1) For all ζ_t, $h \in H$, and $a \in A$, let $\kappa(h; \zeta_t)$ be the number of times that information set h has been reached and let $\kappa(a; \zeta_t)$ be the number of

[8] Compare with the analysis in Fudenberg and Kreps (1993), where we formalized (only) i's joint probability assessment over the pure strategy profile of his rivals, a concept closest to assessments as defined here.

[9] That is, player i assesses the sequence of selections by his rivals as exchangeable.

times a is played in the $t-1$ plays of the game recorded by ζ_t. Note that $\Sigma_{a \in A(h)} \kappa(a; \zeta_t) = \kappa(h; \zeta_t)$.

(2) For all ζ_t and $h \in H$ such that $\kappa(h; \zeta_t) > 0$, define a probability distribution $\tilde{\pi}(h; \zeta_t)$ on $A(h)$ by

$$\tilde{\pi}(h; \zeta_t)(a) = \frac{\kappa(a; \zeta_t)}{\kappa(h; \zeta_t)} \qquad \text{for all } a \in A(h).$$

(3) For all $\zeta \in \mathcal{Z}$, let $H_{\text{p.f.}}(\zeta)$ be those information sets that are reached a strictly positive fraction of the time along the history ζ, using a limit infimum test; i.e., $h \in H_{\text{p.f.}}(\zeta)$ if $\liminf_{t \to \infty} \kappa(h; \zeta_t)/t > 0$.

DEFINITION. Player i's belief rule $\hat{\gamma}^i$ is *asymptotically empirical* if for every $\varepsilon > 0$, infinite history ζ, $j \neq i$, and information set $h^j \in H_{\text{p.f.}}(\zeta) \cap H^j$,

$$\lim_{t \to \infty} \hat{\gamma}^i_t(\zeta_t)(\{\pi^{-i}: \|\pi^j(h^j) - \tilde{\pi}(h^j; \zeta_t)\| < \varepsilon\}) = 1. \tag{3.3}$$

That is, player i assigns probability tending to zero to strategy profiles π^{-i} in which, at the information set $h^j \in H^j$, player j plays something ε or more different from the empirical distribution over j's actions at h^j.

This definition compounds two basic features: (1) i's strategic uncertainty about what j does at h^j vanishes if evidence about j's play at h^j accumulates sufficiently quickly; (2) all past evidence is (asymptotically) equally weighted. For example, if i believes his rival's actions are exchangeable and i computes beliefs using Bayesian inference, i.e., (3.2), and if his initial prior beliefs γ^i_1 are *non-doctrinaire* in the sense of assigning positive probability to every open neighborhood of Π^{-i}, then his beliefs are asymptotically empirical. More generally, asymptotic empiricism holds if i believes that the behavior strategies of his rivals will converge to the play of some single fixed profile, he uses Bayesian inference, and no finite set of observations causes him to attach zero probability to the limit strategy profile lying in some open set.

In contrast, suppose i believes that rival j chooses some (unknown) behavior strategy repeatedly, except that at (random and unobserved) dates, j changes that behavior strategy to some other. Then i's strategic uncertainty will (for most specifications) never vanish; if there is a constant nonzero probability that at any date t j redraws her behavior strategy, then no matter how sure i becomes that he knows what j has been playing "lately," his prediction about j's behavior in the next round must include the (nonvanishing) chance that j has shifted to some new strategy. Moreover, i will naturally put somewhat more weight on past observations of

what j has been doing than on observations in the far distant past. So, for this case, neither part of asymptotic empiricism is valid.

3.4. *Asymptotic Independence*

Because asymptotic empiricism entails vanishing strategic uncertainty at information sets reached a nonvanishing fraction of the time, it implicitly entails *asymptotic independence*. An example illustrates the problem: Imagine a three-player game in which players 1 and 2 each have two actions between which they must choose simultaneously in each period; up or down for player 1, and left or right for player 2. Consider a history ζ along which the limiting frequencies of up–left, up–right, down–left, and down–right are 0.4, 0.1, 0.1, and 0.4, respectively. If player 3's beliefs are asymptotically empirical, then in the limit player 3 will assess probability 0.25 ($= 0.5 \times 0.5$) and not 0.4 that his rivals' joint actions will be up–left.

To understand how this can happen, return to the example of player i forming beliefs according to Bayes' rule as in (3.2), beginning with a nondoctrinaire prior on the space Π^{-i}. Player i, observing a sequence of actions by two rivals as in the previous paragraph, will come to the conclusion that they are about to play up–left with probability 0.25, because i's prior assessment puts probability one on the event that his rivals choose their strategies simultaneously and independently, and in Bayesian inference, no amount of evidence (that has positive prior likelihood) can cause a prior-probability-zero event to assume positive probability. Player i may well believe in correlation at any finite time t, but the extent of correlation must vanish as time passes and his strategic uncertainty (by assumption) vanishes.

The reasonableness of this property of beliefs is tied up with our contention that players know the informational conditions under which the stage game is played, together with the implicit assumption that all mechanisms by which players could objectively coordinate or correlate their play are recorded in the extensive form of the game. Correlated assessments of rivals' play can reflect one's own strategic uncertainty, but as that strategic uncertainty vanishes, so must any correlation. If we assume that (for our players) strategic uncertainty vanishes at information sets reached a nonvanishing fraction of the time, asymptotic independence is forced.

We are not entirely happy with asymptotic independence, which of course reflects unhappiness with asymptotic empiricism as defined above. In the face of strong evidence to the contrary, players ought to question whether they understand the informational structure of the stage game. But to "fix" this problem constitutes a substantial diversion. For now, we proceed with asymptotic empiricism (and asymptotic independence),

noting this unpalatable implication; we return to it (briefly) in closing remarks.

4. Behavior Rules, Experimentation, and Asymptotic Myopia

A *behavior rule* for player i specifies how i will act at each date for each history. Formally, this is given by a sequence of functions $\hat{\pi}^i = (\hat{\pi}^i_1, \hat{\pi}^i_2, \ldots)$, where $\hat{\pi}^i_t$ has domain $\mathfrak{Z}_t$ and range Π^i.

Given beliefs γ^i, we denote player i's expected current payoff to strategy π^i by $u^i(\pi^i, \gamma^i)$, which is

$$u^i(\pi^i, \gamma^i) = \sum_{z \in Z} \mu^i(z|\pi^i, \gamma^i) u^i(z).$$

We also write $u^i(\pi^i, \pi^{-i})$ for i's expected payoff if he plays π^i and his rivals play according to π^{-i}.

In our earlier work on learning in strategic-form games, we assumed that behavior rules were *asymptotically myopic* with respect to the player's beliefs in the following sense: There exists a sequence of nonnegative numbers $\{\varepsilon_t\}$ such that $\lim_{t\to\infty} \varepsilon_t = 0$ and, for each t and ζ_t,

$$u^i(\hat{\pi}^i_t(\zeta_t), \hat{\gamma}^i_t(\zeta_t)) + \varepsilon_t \geq \max_{s^i \in S^i} u^i(s^i, \hat{\gamma}^i_t(\zeta t)). \tag{4.1}$$

(This is not quite our original formulation, as it must be adapted here to deal with beliefs.) We offered some rationales for this assumption and discussed its merits and drawbacks. Very briefly, if players discount the future very heavily, and/or they believe that, at least asymptotically, their actions will not affect the strategies of their rivals, then asymptotic myopia is reasonable. Moreover, one can cobble together stories of various forms of random matching among large populations of players to justify either a large discount rate and/or the hypothesis that one's future rivals' actions are asymptotically unaffected by one's own current actions.[10]

In this paper, we wish to proceed in a similar spirit and assume that players' behavior rules are asymptotically myopic with respect to their belief rules. But the extension to our current context of learning to play an extensive-form game is not straightforward for at least two reasons. We discuss the two complications and then reformulate asymptotic myopia.

[10] Ellison (1993) provides conditions under which a patient rational player can improve on myopic behavior for a fixed population size.

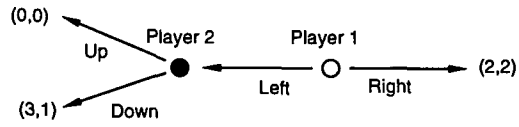

FIG. 1. An extensive-form game.

4.1. *Ex ante or ex post Expectations?*

The first complication concerns the stage game depicted in Fig. 1. Imagine player 2 entertains beliefs that player 1 will play Right with probability p close to one. Then the strategy Up is not at all costly to player 2 *ex ante*: Choosing Up is suboptimal by the *ex ante* expected amount $1 - p$. Thus if player 1 plays Right increasingly often, player 2 (with asymptotically empirical beliefs) sees the choice of Up as vanishingly suboptimal. And if player 2 persists with Up, then player 1 will (probably) be more and more inclined to choose Right.

But if player 2 is ever called upon to move, it is clear *ex post* that the choice of Up will cost her one unit of payoff. We are inclined to say that "suboptimality cost calculations" should be formulated in terms of *ex post* expected payoffs, so that player 2 may not persist with Up whenever given the chance, even if player 1 chooses Right with a frequency that approaches one.

Notwithstanding this inclination, in this paper we formulate asymptotic myopia using *ex ante* expected payoff calculations. By so doing we are using a weaker form of asymptotic myopia, which permits more behavior rules to qualify. After seeing the consequences of this weak assumption on behavior rules, we might wish later to explore what happens if asymptotic myopia is formulated on the basis of *ex post* expected payoffs. But that must await another paper (by ourselves or others).

4.2. *Experimentation*

The second complication can also be posed in terms of the stage game in Fig. 1, with the emphasis now on the behavior of player 1. Imagine that player 1 believes that player 2 will choose Up in each round independently with some probability p and player 1's initial beliefs are that p is uniformly distributed over $[0, 1]$. Then in the first round player 1's marginal assessment is that player 2 will choose Up with probability 1/2, and player 1's immediate expectations favor the choice of Right. If player 1 chooses Right, he does not receive any information about the value of p, because player 2 is not given the opportunity to move. So in the second round, 1's beliefs remain his prior, and again short-run considerations lead to a choice of Right. If player 1 acts myopically in the sense of maximizing

his expected payoff in each round, given his beliefs, he will choose Right in each round. But if player 1 does not discount the future very heavily, he may choose to play Left for some period of time to learn the value of p; if the data lead to the conclusion that $p < 1/3$ (which has prior probability 1/3 according to player 1), Left becomes short-run optimal.

In the context of learning to play a strategic-form game, in which each player learns his rivals' pure strategies after each round of play,a player who believes that his own actions do not affect the subsequent choices of his rivals will wish to play whatever strategy maximizes his immediate expected payoff. If a player believes that his own actions will *asymptotically* have no impact on the actions of his rivals, then asymptotic myopia is mandated.[11] But in the current context of learning to play an extensive game, this argument fails because the player's immediate actions can affect what he learns about his rivals' behavior.

4.3. Experiments Taken at Random and Asymptotic Myopia

The challenge, then, is to formulate a general property in the spirit of asymptotic myopia that allows players to experiment with suboptimal strategies, to learn more about how their rivals act in otherwise unreached parts of the game tree. The first thing to note in this regard is that experimentation of a particular kind is *possible* within the confines of condition (4.1); viz., condition (4.1) does not preclude *experiments taken at random*: Suppose that in round t each player chooses each available (pure) strategy for the extensive-form game with probability at least α/t for some constant α. With the remaining probability, each player chooses any strategy that is optimal with respect to his beliefs. Then the degree of suboptimality of the overall mixed strategy used by the player (computed, of course, on an *ex ante* basis) vanishes as time passes. And it is an easy consequence of the Borel–Cantelli lemma that each player will (almost surely) use each pure strategy infinitely often over the course of play.[12]

4.4. Asymptotic Myopia with Conscious Experiments

We wish to weaken the condition for asymptotic myopia, (4.1), to permit players to adopt behavior rules that experiment with suboptimal actions

[11] We are not being precise about what is meant by "asymptotically has no impact," so this is somewhat loose.

[12] Note, however, that players need not take every action infinitely often, not even at information sets that are reached infinitely often. To take a simple example, in the game in Fig. 1, imagine that player 1 chooses Left in round t with probability $1/t$, and player 2 chooses Up in round t with probability $1/t$, independently of what player 1 has done. Then player 2 is almost surely given infinitely many chances to act, but the combination Left–Up is (jointly) chosen only finitely often.

with probability one in specific circumstances, as long as these experiments are not taken too often in a time-average sense. We wish to permit these sorts of nonrandom experiments on two grounds: First, our intent is to allow as broad a class of behavior rules as we can, and while some decision makers may be content with experiments taken at random, other decision makers may choose their experiments at specific times and in specific circumstances, albeit at vanishing frequency. For example, a player who envisions himself as facing a multi-armed bandit, with a discounted payoff criterion, will not optimally conduct randomized experiments. Second, and related to this, a player's decision whether to experiment may well depend on the outcome of his past experiments. For example, if a player chooses a suboptimal action in period t with probability α/t, there is positive probability that, at any given time T, the player has yet actually to try the experiment. We do not wish to preclude a player from deciding, in this circumstance, to run the experiment once and for all (finally) to get some information.

We make the following definition.

DEFINITION. Fix i and i's beliefs rule $\hat{\gamma}^i$. The behavior rule $\hat{\pi}^i$ for i is *asymptotically myopic with calendar-time limitations on experimentation* if there exist: (1) a sequence of strictly positive numbers $\{\varepsilon_t\}$ with $\lim_{t\to\infty} \varepsilon_t = 0$, (2) a nondecreasing sequence of nonnegative integers $\{\eta_t;\ t = 1, 2, \ldots\}$ with $\lim_{t\to\infty} \eta_t/t = 0$, (3) behavior rules $\check{\pi}^i$ and $\bar{\pi}^i$ for i, and (4) for each t, ζ_t, and $h \in H^i$, a number $\check{\alpha}_t^i(\zeta_t)(h) \in [0, 1]$, such that:

(a) For all t, ζ_t, and $h \in H^i$, $\hat{\pi}_t^i(\zeta_t)(h) = \check{\alpha}_t^i(\zeta_t)(h) \times \check{\pi}_t^i(\zeta_t)(h) + (1 - \check{\alpha}_t^i(\zeta_t)(h)) \times \bar{\pi}_t^i(\zeta_t)(h)$.

(b) For all t, ζ_t, and $h \in H^i$, $u^i(\check{\pi}_t^i(\zeta_t), \hat{\gamma}^i(\zeta_t)) + \varepsilon_t \geq \max_{s^i \in s^i} u^i(s^i, \hat{\gamma}^i(\zeta_t))$.

(c) If $\check{\alpha}_t^i(\zeta_t)(h) < 1$, then $\kappa(\alpha'; \zeta_t) \leq \eta_t$ for some $a' \in A(h)$, and $\bar{\pi}_t^i(\zeta_t)(h)$ gives positive probability only to actions $a \in A(h)$ such that $\kappa(a; \zeta_t) \leq \eta_t$.

To explain: Condition (a) says that at date t with history ζ_t, $\hat{\pi}_t^i(\zeta_t)$ prescribes behavior at information set h that is a convex combination of two pieces: $\check{\pi}_t^i(\zeta_t)(h)$ and $\bar{\pi}_t^i(\zeta_t)(h)$. We imagine that player i decides, information set by information set, whether to conduct a conscious experiment. We interpret the $\check{\pi}^i$ part as player i's nonexperimental behavior and $\bar{\pi}^i$ as player i's experimentation, so if $\check{\alpha}_t^i(\zeta_t)(h) = 1$, player i has decided not to experiment at h. Condition (b) says that i's nonexperimental behavior is vanishingly suboptimal. Condition (c) says that if player i chooses to experiment at h at all, that is, if $\check{\alpha}_t^i(\zeta_t)(h) < 1$, then he must be experimenting with actions $a \in A(h)$ that have been taken fewer than η_t times in the past. Since $\eta_t/t \to 0$ as $t \to \infty$, this means that experiments are taken a vanishing fraction of the time.

We call this asymptotic myopia with *calendar-time* limitations on experimentation because the player is allowed to experiment freely with actions that have been taken "infrequently" relative to calendar time. If information set h is visited a vanishing frequency of time—more precisely if $\kappa(h; \zeta_t) < \eta_t$ for all t along the history ζ—then i is free to take any actions that he wishes at the information set h. Thus i's behavior at information sets that are reached rarely enough is unrestrained by asymptotic myopia.

Note that although we refer to the $\check{\pi}^i$ part of i's behavior strategy as *nonexperimental*, $\check{\pi}^i$ can incorporate experiments taken at random, as per our discussion above.

Is it reasonable to suppose that a boundedly rational player will behave asymptotically myopically with experiments that vanish with calendar time? To answer this question, two issues must be addressed. First, holding aside the question of experiments, is asymptotic myopia at all sensible? Again, we will not attempt to defend this behavioral postulate here; Fudenberg and Kreps (1993) gave our rationales for it. Second, in this setting where actions can influence what information is received, so that experiments may be reasonable forms of behavior, is it sensible to insist that the frequency of experimentation vanish with calendar time in the sense of the definition, or might it be reasonable for someone to experiment more frequently than this?

There are certainly scenarios in which this definition is too restrictive. For example, if player i supposed that his opponents played the same profile of behavior strategies repeatedly but, at random times, shifted to some other profile, then experimentation at a nonvanishing rate is entirely reasonable. But if asymptotic empiricism is justified—if i believes that his rivals will asymptotically settle into repeated play of a given strategy profile—"the value of information" to be obtained in any experiment presumably diminishes to zero the more often the experiment is taken, and thus the frequency of experimentation should vanish.

4.4. *Rationalizing Calendar-Time Limitations: Players Who Compare the Sum of Immediate and Future Payoffs*

To explain this last sentence, and to shed light on our definition of asymptotic myopia, it helps to delve a bit deeper into the calculations that *might* guide the behavior of player i. Note well, we do not mean to limit ourselves to players who reason in this fashion; this is only an example. Also, we will not be precise. In particular we sluff over details having to do with players who move more than once along paths through the game tree.

Imagine player i chooses behavior at date t as follows. For each pure strategy $s^i \in S^i$, i has immediate expected payoff $u_t^i(s^i, \hat{\gamma}_t^i(\zeta_t))$. Suppressing the dependence on history and calendar time, abbreviate this as $u^i(s^i)$. In

addition, i has some sense of his "current-value expected future payoffs" if he chooses s^i this period, which we abbreviate $f^i(s^i)$. We suppose that i chooses from among those pure strategies s^i that satisfy $u^i(s^i) + f^i(s^i) + \varepsilon_t' \geq \max_{\tilde{s}^i} u^i(\tilde{s}^i) + f^i(\tilde{s}^i)$, for some $\varepsilon_t' \downarrow 0$.

Suppose (withough justification just yet) that

$$\text{for some sequence } \delta_k \to 0, f^i(s^i) - f^i(\tilde{s}^i) \leq \delta_{\kappa(s^i)}, \text{ for all } s^i \text{ and } \tilde{s}^i, \quad (4.2)$$

where $\kappa(s^i)$ is the number of times s^i has been attempted. Let η_t be any nondecreasing sequence of positive integers with $\eta_t \to \infty$ and $\eta_t/t \to 0$, and let $\varepsilon_t = \varepsilon_t' + \delta_{\eta_t}$. Then i (choosing as we have imagined) will satisfy asymptotic myopia for η_t and ε_t: If s^i has been tried η_t times or more by time t, then it can be better in terms of future value than any other strategy by at most δ_{η_t}. Thus if it is worse than some other strategy in terms of current value by more than ε_t, then it must be worse than this other strategy in terms of current plus future value by at least ε_t', and s^i will not (therefore) be chosen.

Can we justify the uniform bound in (4.2)? Suppose player i believes at the outset that his rivals are playing according to some fixed strategy profile. If i discounts his payoffs using some discount rate less than 1, $f^i(s^i)$ is the future value function of a problem very much like the classic multi-armed bandit problem, where each pure strategy $s \in S^i$ that i might choose corresponds to one arm of the bandit. The problem differs from the standard bandit model in that the returns to the various arms may be correlated, but the solution to this "extensive-form bandit" has many of the features of the solution to the case of independent arms.[13] In this setting, (4.2) has appeal along the following intuitive lines: The more "arm" s^i is tried, the more is learned about the consequences of trying this strategy, and the less there is to learn.

This intuition (and the uniform bound in (4.2)) holds for standard multi-armed bandit problems, as long as prior beliefs are non-doctrinaire. But it fails in general for extensive-form bandit problems, as the following example indicates. Consider the game depicted in Fig. 2. (Only player 1's payoffs matter, so only they are given.) Imagine that player 1 believes at the outset that players 2, 3, and 4 will repeatedly play mixed strategies, with p the probability with which player 2 chooses Left, q the probability that player 3 chooses left, and r the probability that player 4 chooses gauche. Player 1 initially believes that (p, q, r) has uniform distribution on the unit cube, which makes Out the short-run optimal strategy. But if pqr is low enough, In would be better for player 1, and so with small

[13] Fudenberg and Levine (1993b) analyze optimal solutions to extensive-form bandit problems.

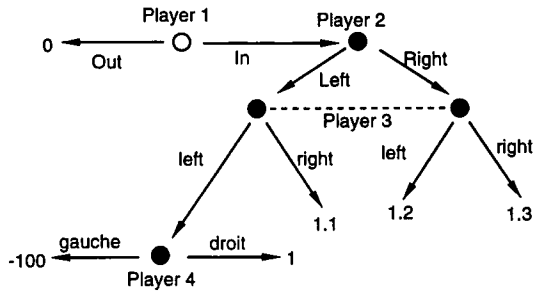

FIG. 2. A troublesome example. Only player 1's payoffs are given.

enough discount rate, player 1 would optimally choose In, to learn about the values of p, q, and r. Now imagine that whenever player 1 chooses In, either Left–right or Right–left is observed, each with limiting frequency 1/2. Assuming that player 1's beliefs are strongly asymptotically empirical, player 1 comes to conclude that $p = q = 1/2$. By asymptotic independence, player 1 believes that there is 1/4 chance that, if he chooses In, he will (finally) learn something about the value of r. Until something is learned about r, In remains short-run suboptimal by an amount bounded away from zero. But as long as 1's discount rate is very small, the expected value of information obtained from In more than makes up for this short-run suboptimality. Along the path where 2 and 3 alternate between Right–left and Left–right, player 1 never abandons In, despite the fact that this strategy remains distinctly suboptimal. N.B., the strategy employed by player 1, which is an optimal strategy according to dynamic programming given 1's initial beliefs, will fail to meet our definition of asymtotic myopia; hence the definition unduly limits the amount of experimentation that player 1 may undertake.

Comparing this example with the result we claim for standard, independent-arms bandit problem, it is clear where the difficulty arises, viz., from the players' doctrinaire belief that their opponents' play is uncorrelated (despite their non-doctrinaire beliefs over the strategy of each individual opponent), which they maintain no matter how strongly the data suggest otherwise. This suggests that abandonment of asymptotic independence will solve this problem. Alternatively, we can argue that if players 2 and 3 are using a fixed stragegy, then the sort of correlated history that underlies this example is unlikely to occur. Either of these can provide a basis for the bound (4.2) and thus justify calendar-time bounds on experimentation along the lines sketched above; see the concluding remarks. But, taking note of this example, we are forced to conclude that calendar-time limitations on experimentation can be sharper than we would like, at least in some (exceptional) circumstances.

And while our rationalization is a bit flawed, it is also based on a limited view of player i's deliberations. In this respect, we reiterate that this rationalization is not equivalent to the formal assumption. For one thing, if we move outside the context of extensive-form bandit problems, to cases in which a player believes his actions can affect the play of his rivals but that any effect vanishes with calendar time, the bound (4.2) still has intuitive (and, excepting the problem raised by the example, formal) appeal. For another, asymptotic myopia permits more experimentation than would be called for in the story developed above, as it permits continued experimentation, albeit at a vanishing rate, with strategies that are decidedly suboptimal. This is permitted to encompass behavior of individuals who do not reason as described here, but who, for example, act in a fashion to maximize the long-run (undiscounted) average payoff they receive.[14]

4.6. *A Comment on Asymptotic Empiricism*

As a final comment, we return to the definition of asymptotic empiricism and, in particular, to the reason why (3.3) is required only for information sets that are reached a nonvanishing fraction of the time. The question is, What credence do players give to evidence generated at information sets visited infinitely often but a vanishing fraction of time? If a player believes that his rivals are playing the same strategy profile repeatedly, he ought to put a lot of credence in this evidence. But our formulation of asymptotic myopia suggests two reasons that such evidence might be considered to be of lesser quality than data generated at an information set visited a nonvanishing fraction of the time.

First, we assume players are asymptotically myopic using *ex ante* evaluation of expected payoffs. Insofar as players assess vanishingly small probability of reaching an information set that has been visited a vanishing fraction of the time,[15] their behavior at those information sets is relatively unconstrained by asymptotic myopia. Thus a player may believe that the actions of his rivals at information sets visited a vanishing frequency of time could be capricious and hence are too irregular to be predicted by the empirical frequencies of previous actions.[16]

[14] In bandit problems, any strategy that picks the short-run optimal action a fraction of the time that approaches one, while picking each action infinitely often, will be average-payoff optimal almost surely. Of course, maximizing average payoffs is a notoriously weak criterion, admitting many optimal strategies.

[15] This *insofar* has a purpose; this is not an implication of asymptotic empiricism. As the example in the previous subsection shows, asymptotic independence may cause a player to assess nonvanishing probability for reaching an information set that is *never* reached in the course of play.

[16] Having introduced the notion that behavior might be capricious or (more to the point) irregular when it does not have much effect on expected payoffs, we should note that this

Second, as we have noted, calendar-time limitations on experimentation do little to restrict behavior at information sets reached a vanishing fraction of time. Choose $h \in H^i$, and suppose that $\kappa(h; \zeta_t) < \eta_t$ for all t. Then i can choose any action, and in particular can act fairly capriciously, at information set h. Especially when a second player j suspects that i will experiment in a way that is limited by this sort of calendar-time test but does not know the sequence $\{\eta_t\}$ that limits the experiments of player i, j may be relatively more wary of data generated at h if h is visited a vanishing frequency of time.

5. Unstable and Locally Stable Strategy Profiles

We next provide formal "convergence criteria" that we will use. Throughout this discussion, an extensive-form stage game is fixed.

Definition. A *learning model* for the extensive-form stage game is an array of behavior and beliefs rules, one each for each player i. A learning model is said to be *conforming* if each beliefs rule is asymptotically empirical and each behavior rule is asymptotically myopic with calendar-time limitations on experiments, relative to the corresponding beliefs rule.

Given any learning model (or, more simply, an array of behavior rules), we define in the usual fashion the induced probability measure over the space of complete histories $\mathfrak{L}$. As long as there is no ambiguity about the fixed learning model, $\mathbf{P}$ will denote this probability measure, and $\mathbf{E}$ will denote expectation taken with respect to $\mathbf{P}$.

Whenever a conforming learning model is fixed, $\hat{\pi}$ will refer to the array of behavior rules, $\check{\pi}$ will refer to the array of "nonexperimental parts" of the behavior rules (as given by the definition of asymptotic myopia), and so on.

Definition. A strategy profile $\pi_* \in \Pi$ is *unstable* if there exists some $\varepsilon > 0$ such that, for all conforming learning models, $\mathbf{P}(\|\check{\pi}_t(\zeta_t) - \pi_*\| < \varepsilon \text{ for all } t) = 0$.

Definition. A strategy profile $\pi_* \in \Pi$ is *locally stable* if there exists some conforming learning model such that $\mathbf{P}(\lim_{t\to\infty} \check{\pi}_t(\zeta_t) = \pi_*) > 0$.

It should be clear that these definitions are mutually exclusive. It is not *a priori* obvious that they are exhaustive, but Propositions 7.1 and 7.2 will show that every strategy profile is either unstable or locally stable.

---

poses problems as well for actions taken where players are close to indifferent, e.g., in situations where they are meant to be randomizing. Noisy payoffs, in the sense of Harsanyi's (1973) work on purification, can be a device for avoiding this sort of problem; see, for example, Section 7 of Fudenberg and Kreps (1993).

Note that in both definitions, the "target" profile π_* is compared with the nonexperimental parts of each player's behavior rules and not with the behavior rules themselves. For these definitions to have some empirical content, and in particular for the definition of local stability to have content, we will want to show that the strategies actually played (given by the behavior rules) resemble to some extent the target strategy.

Compared to the corresponding definitions from Fudenberg and Kreps (1993), three things are noteworthy.

(1) In the definition of unstable profiles given here, ε must work uniformly for all conforming models. In Fudenberg and Kreps (1993), ε is permitted to vary with the model of behavior and beliefs. But (as in fact noted in Fudenberg and Kreps (1993)) all the results in the earlier paper go through for the stronger definition here.[17]

(2) On the other hand, here we require only that the probability of staying in the ε-neighborhood of π_* have prior probability zero; previously we required that this be true conditional on any partial history of previous play. But it is easy to see, given the uniformity of ε over all conforming models, that this seemingly weaker requirement is equivalent: The dynamics beginning at any partial history of play in a conforming model are precisely the same as the dynamics beginning at date 1 in a different conforming model.

(3) In the definition of local stability given here, there must be positive probability of the nonexperimental part of behavior converging to the target profile *ex ante*, in some conforming model. In Fudenberg and Kreps (1993), we required that for a fixed conforming model, for every $\varepsilon > 0$ we could find a partial history such that convergence to the target strategy profile had conditional probability at least $1 - \varepsilon$, conditional on the partial history. These are in fact equivalent; cf. Lemma A.1 of Fudenberg and Kreps (1993).

6. Self-Confirming Equilibria

Definition. The strategy profile π_* is a *self-confirming equilibrium* if for each player i there are beliefs γ_*^i such that

(a) π_*^i maximizes $u^i(\pi^i, \gamma_*^i)$, and
(b) $\gamma_*^i(\{\pi^{-i}: \pi^j(h^j) = \pi_*^j(h^j)$ for all $j \neq i$ and $h^j \in \overline{H}(\pi_*)\}) = 1$.

In other words, self-confirming equilibrium requires that each player's strategy be a best response to his beliefs and that each player's beliefs are correct along the equilibrium path of play.

[17] In fact, the proofs given in the earlier paper are entirely adequate as given.

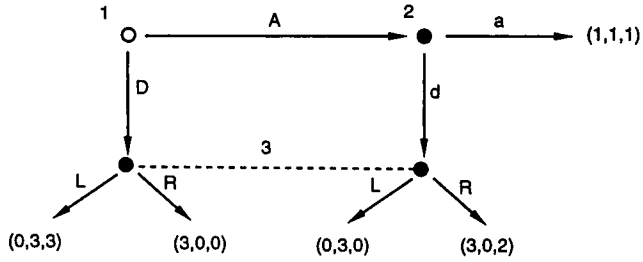

FIG. 3. A three-player horse-shaped game.

To illustrate how self-confirming equilibria can differ from Nash equilibria (in terms of the strategy profiles), consider the following two examples.

EXAMPLE 6.1. *Inconsistent beliefs about the behavior of a third player.* Consider the game depicted in Fig. 3. If players 1 and 2 have the same beliefs about the strategy of player 3, one of them (or both) will strictly prefer giving 3 the move over moving across. Thus there is no Nash equilibrium of this game in which the outcome is $A - a$. But if player 1 believes player 3 will choose L with probability exceeding 2/3 and player 2 will choose a with positive probability, then player 1 prefers A to D. If player 2 believes player 3 will choose R with probability exceeding 2/3 and player 1 will choose A with positive probability, then player 2 prefers a to d. Thus there is a (non-Nash) self-confirming equilibrium in which player 1 chooses A and player 2 chooses a, based on diverse beliefs by the two of them about the off-the-path strategy of player 3.

EXAMPLE 6.2. *Persistent correlation in one player's beliefs about the strategies of others.* In the game in Fig. 4, player 1 can play U, which ends the game, or play L, M, or R, all of which lead to a simultaneous-move game between players 2 and 3, neither of whom observes player 1's action. The game between players 2 and 3 is a simple game of coordination; the payoffs to them for a, a' are (6, 8), while d, d' gives payoffs (10, 5), and a, d' and d, a' both give zero payoff to both 2 and 3.

If player 1 assesses that player 2 chooses a with probability p^2 and player 3 chooses a' with probability p^3, *independent* of the actions of player 2, it is straightforward to show that U is never 1's best response. Thus in no Nash equilibrium of this game, where player 1 *knows* the strategies of players 2 and 3, will U be chosen. But suppose player 1 chooses U. This choice puts the information sets of players 2 and 3 off the path of play, and so player 1 can entertain correlated conjectures about the strategic choices of 2 and 3. In particular, if player 1 believes that players 2 and 3 choose $a - a'$ with probability close to 1/2 and $d -$

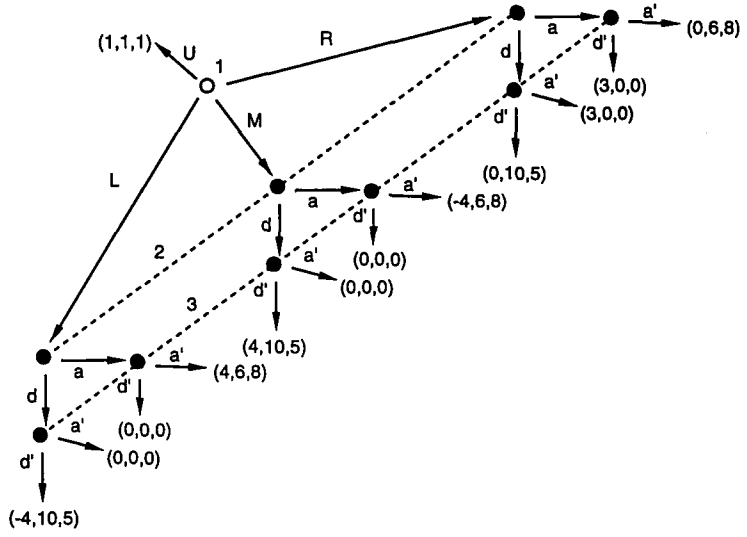

FIG. 4. Illustrating correlation in off-the-path assessments.

d' with probability close to 1/2 (which requires correlated conjectures), U is 1's best response. Thus the choice of U (with certainty) is part of a self-confirming equilibrium.

The phenomena that underpin these two examples—two players holding different beliefs about the actions of a third, and one player ascribing correlation to the off-the-path actions of two rivals—necessarily entail at least three players. For two-player extensive-form games, we have the following result, which is proved in the Appendix.

PROPOSITION 6.1. *In a two-player extensive-form game, every self-confirming equilibrium is equivalent to a Nash equilibrium in the sense that, if π_* is a self-confirming equilibrium profile, then there is a Nash equilibrium profile $\bar{\pi}$ which gives the same distribution on outcomes as does π_*.*

7. BASIC RESULTS

PROPOSITION 7.1. *Every strategy profile π_* that is not a self-confirming equilibrium is unstable.*

PROPOSITION 7.2. *Every strategy profile π_* that is a self-confirming equilibrium is locally stable.*

Discussion and details of the proofs will be given in later subsections. In view of the widespread acceptance of the concept of a Nash equilibrium as *the* preeminent solution concept for extensive-form games, it is Proposition 7.2 that is immediately more interesting. To illustrate this, we briefly indicate how the non-Nash self-confirming equilibria in Examples 6.1 and 6.2 could be locally stable.

EXAMPLE 6.1 (continued). Consider the game depicted in Fig. 3 and any behavior profile in which player 1 chooses A and player 2 chooses a. Suppose player 1 begins believing that players 2 and 3 are playing fixed mixed strategies over and over, although he does not know which pair of strategies they are playing. His initial beliefs are given by the following probability distribution γ_1^1 on $\Pi^2 \times \Pi^3$:

$$\gamma_1^1(\{\pi^2(a) \leq p, \pi^3(L) \leq q\}) = p^{100}q^{100},$$

for $0 \leq p, q, \leq 1$. Note that γ_1^1 is a product measure on $\Pi^2 \times \Pi^3$; player 1's strategic uncertainty about the strategies used by 2 and 3 exhibits independence. Player 1 updates his beliefs in the light of new evidence by using Bayes' rule in the manner of Eq. (3.2).

Player 2's initial and subsequent beliefs are similar in structure to player 1's, beginning with the prior

$$\gamma_1^2(\{\pi^1(A) \leq p, \pi^3(R) \leq q\}) = p^{100}q^{100}.$$

Player 3's beliefs are unimportant.

Because players 1 and 2 have non-doctrinaire prior beliefs and they update beliefs using Bayes' rule, their belief rules are asymptotically empirical.

As for behavior rules, suppose that all behavior is precisely myopic.

Given his initial beliefs, player 1 assesses probability

$$\int_0^1 100q^{100}\mathrm{d}q = 100/101$$

that player 3 will choose L given the opportunity, and that player 2 will choose a with probability 100/101. So myopic optimization leads 1 to choose A. Given his initial beliefs, player 2 assesses probability 100/101 that 1 will choose A and 100/101 that 3 will choose R, so player 2 chooses a. The initial outcome is (A, a).

When players update their beliefs given this initial outcome, player 1 increases the mass on strategies in which 2 is likely to pick a, and player 2 increases the mass on strategies in which 1 is likely to pick A. The exact calculations are both easy and unimportant. The important point is that,

in the second round of play, *neither 1 nor 2 changes beliefs about 3, hence neither changes her assessment of what 3 would do if given the chance to move*. Because no evidence was produced about the play of player 3, and 1s and 2s believe that the strategies used by their rivals are drawn independently, there is no information in 2's play of a, for example, about what 3 might do.

Hence in round 2, 1 chooses A and 2 chooses a. And so on forever. The outcome in each round is (A, a). Supposing that 3's behavior is fixed, behavior profiles have converged (trivially) to a non-Nash self-confirming equilibrium profile. The point is very simple. Players 1 and 2 begin with disparate beliefs on what strategy 3 is likely to use. This leads them to behavior that keeps 3 from moving. And if 3 never moves, then 1 and 2 have no opportunity to learn what 3 would in fact do, so that their disparate beliefs can persist.

When confronted with this example, colleagues often have asked the following question. Suppose that player 2 knows player 1's payoff function and knows that player 1 knows his own payoffs. Then when player 2 sees player 1 play A, she can infer that player 1 expects player 3 to play L with substantial probability. Should this not lead player 2 to revise her beliefs about player 3 in the direction of increasing the probability that player 3 plays L? In the spirit of the literature on the impossibility of players "agreeing to disagree," should players 1 and 2 not end up with the same beliefs about player 3? While we do not preclude this sort of indirect learning in our model, it need not take place. First, the indirect learning supposes that players know (or have strong beliefs about) one another's payoffs, which is consistent with our model but is not necessary for it. If player 2 is unaware of player 1's payoffs (and vice versa), then 2 would not find it particularly surprising that 1 chooses A. Second, even if player 2 knows player 1's payoffs (and knows that player 1 knows them), and hence is able to infer that player 1 believes player 3 is likely to play L, it is not clear that this will lead player 2 to revise her own beliefs. It is true that player 2 will revise her beliefs if she views the discrepancy between her own beliefs and player 1's as due to information that player 1 has received but player 2 has not. But player 2 might also believe that 1 has no objective reason for her beliefs and has simply made a mistake. The "agreeing to disagree" literature ensures that all differences in beliefs are attributable to differences in information by supposing that the players' beliefs are consistent with Bayesian updating from a common prior distribution. But assuming a common prior assumes away the key question of learning outside of equilibrium. Indeed, the question of whether learning leads to Nash equilibrium would seem to be a special case of the question of whether (and when) learning leads to common *posterior* beliefs starting from arbitrary priors. To emphasize this point, recall that assuming players have a common prior distribution over one another's strategies is equiva-

lent to assuming that their beliefs correspond to a correlated equilibrium (Aumann, 1987), and assuming an independent common prior is equivalent to Nash equilibrium (Brandenburger and Dekel, 1987).

EXAMPLE 6.2 (continued). Suppose, in the game in Fig. 4, that player 1 begins with beliefs γ_1^1 that take the following form: γ_1^1 has an atom of mass 0.495 on the pure strategy profile (a, a'), an atom of mass 0.495 on the pure strategy profile (d, d'), and the remaining 0.01 probability is uniformly distributed over all the possible (mixed) strategy profiles that 2 and 3 could employ. If player 1 uses Bayes' rule to update his beliefs, he has a beliefs rule that is asymptotically empirical (because his prior beliefs are non-doctrinaire). And if his behavior is myopic, with these initial beliefs, he will choose the action A.

But if he chooses the action U in the first round, then players 2 and 3 do not get a chance to move, and 1's beliefs going into the second round are identical with his beliefs in the first round. Again he chooses U, again he learns nothing, and play is "trapped." It is straightforward to flesh this out into a fully specified example in which there is a locally stable strategy profile that has 1 playing U with probability one, despite the fact that in all Nash equilibrium profiles, U occurs with probability zero.

We stress that this does not rely on player 1 believing that players 2 and 3 actually correlate their play. To the contrary, player 1 is certain that they do not do so, and that he *could* learn which (uncorrelated) strategy profile they are using by giving them a chance to play enough times. (Moreover, a lot of information would probably be communicated in the first observation.) However if player 1 is impatient or for any other reason behaves myopically, he never finds out how they would behave. His persistent strategic uncertainty means that he persists in correlated assessments of their play, which (given he behaves myopically) results in his persistent strategic uncertainty.[18]

7.2. *Proving Proposition 7.1*

We will not give all the details of the proof of Proposition 7.1, relying instead on a detailed sketch. This sketch, while long and cumbersome,

[18] Where might the initial correlated beliefs of player 1 come from? Suppose that before the first play of the three-player game described above, players 2 and 3 have repeatedly played a 2×2, two-player coordination game whose payoffs are exactly as in Fig. 4. That is, initially players 2 and 3 play a game without a player 1, and then later on player 1 is added. Suppose further that player 1 does not observe play in the initial two-player game. It seems natural to suppose that players 2 and 3 will view their part of the game in Fig. 4 as the same as the two-player game that preceeded it, and hence they will use their previous experience to guide their play in the current game. Player 1 (and we) might assess high probability that play in the initial two-player game has converged to one of the pure-strategy equilibrium, without being able to predict which of those two equilibria has emerged.

will probably be incomprehensible to readers who are not familiar with the proofs of Lemma 6.2 and Proposition 6.1 from Fudenberg and Kreps (1993).

Fix some strategy profile π_* that is not a self-confirming equilibrium.

LEMMA 7.1. *If π_* is not a self-confirming equilibrium, there exists an $\varepsilon' > 0$ and a player i' such that for all beliefs $\gamma^{i'}$ such that*

$$\gamma^{i'}(\{\bar{\pi}^{-i'}: \max_{i \neq i', h \in H^i \cap \bar{H}(\pi_*)} \|\bar{\pi}^i(h^i) - \pi_*^i(h^i)\| < \varepsilon'\}) > 1 - \varepsilon',$$

there exists an $s^{i'}$ such that $u^{i'}(s^{i'}, \gamma^{i'}) \geq u^{i'}(\pi^{i'}, \gamma^{i'}) + \varepsilon'$ for all $\pi^{i'}$ such that $\|\pi^{i'} - \pi_*^{i'}\| < \varepsilon'$.

Proof of the Lemma. This is standard once we note that the set of profiles such that

$$\max_{i \neq i', h \in H^i \cap \bar{H}(\pi_*)} \|\bar{\pi}^i(h^i) - \pi_*^i(h^i)\| < \varepsilon$$

is compact, and $u^{i'}(\cdot;\cdot)$ is continuous. ■

Suppose that the profile π_* is not a self-confirming equilibrium. Fix a player i' and an ε' satisfying the conclusions of Lemma 1, and let

$$\varepsilon = \min \{\varepsilon'/2, \pi_*(a)/2; a \in A, \pi_*(a) > 0\}.$$

Suppose that for some conforming model,

$$\mathbf{P}(\{\zeta: \|\check{\pi}_t(\zeta_t) - \pi_*\| < \varepsilon \text{ for all } t\}) > 0.$$

Denote the event $\{\zeta: \|\check{\pi}_t(\zeta_t) - \pi_*\| < \varepsilon \text{ for all } t\}$ by Λ.

The idea will be to show that with probability one on Λ, every information set $h \in \bar{H}(\pi_*)$ is hit a nonvanishing frequency of time and that the empirical distribution of actions taken there lies within ε of π_*. Thus by weak asymptotic empiricism, player i' will come to hold beliefs that force him (under the force of asymptotic myopia) to abandon anything within ε of $\pi_*^{i'}$ as the nonexperimental portion of his behavior rule (all of this, almost surely on Λ). This will then contradict the definition of Λ, giving a contradiction that proves the proposition. The reader familiar with the proof of Proposition 6.1 in Fudenberg and Kreps (1993) should be able to see all the steps in this proof except for the first step. So in what follows, we indicate how the first step is proved.

We "simulate" the process using an independent family of uniform random variates. Specifically, let $\{\chi_k^i(h): i = 1,2,3; k = 1,2, \ldots; h \in H\}$

be an independent family of uniform random variates. (If the game has more than one initial node, a further sequence of uniform random variates is used to simulate the starting position. In the sketch to follow, we ignore this complication, to keep the exposition relatively simple.) At the initial information set h_0, use the triple $\{\chi_1^i(h_0) : i = 1,2,3\}$ to simulate the action chosen by the player to whom h_0 belongs. Specifically, let i_0 be this player, and let $\hat{\pi}\,{}^{i_0}_{\dagger}(\zeta_1)$ be the prescribed behavior of this player in the first round of play. Write $\hat{\pi}\,{}^{i_0}_{\dagger}(\zeta_1)$ at h_0 as $\check{\alpha}(h_0, \zeta_1)\check{\pi}\,{}^{i_0}_{\dagger}(\zeta_1) + (1 - \check{\alpha}(h_0, \zeta_1))\bar{\pi}\,{}^{i_0}_{\dagger}(\zeta_1)$, where $\check{\alpha}(h_0, \zeta_1)$ is the probability that i_0 initiates play without a permitted experiment, $\check{\pi}\,{}^{i_0}_{\dagger}(\zeta_1)$ is the nonexperimental portion of i_0's strategy, and $\bar{\pi}\,{}^{i_0}_{\dagger}$ represents the experimental portion (if any). Use $\chi_1^1(h_0)$ to simulate a choice of action at h_0 according to $\check{\pi}_1$ at h_0, use $\chi_1^2(h_0)$ to simulate a choice of action at h_0 according to $\bar{\pi}_1$ at h_0, and use $\chi_1^3(h_0)$ to simulate the bivariate random event whether to experiment or not. Depending on the results of this first stage of simulation, a second information set h_1 will be reached; use the triple $\{\chi_1^i(h_1) : i = 1,2,3\}$ to simulate what action is taken at that information set, and so on.

The key is this: As each information set h is reached in turn, we use the triple $\{\chi_k^i(h) : i = 1,2,3\}$, *where k is one plus the number of times that this information set has been reached so far in the simulation.* That is, we do not use the "next" triple of uniform random variates for information set h until, in the simulation, h is reached. And each time we simulate what goes on at a particular information set, we use the first variate in the triple to simulate the nonexperimental portion of the strategy, the second to simulate the experimental portion, and the third to decide "whether to experiment," if (according to the player's decision rule) there is a chance that he will conduct an experiment in that round.

As long as we stay within the event Λ in our simulation, whenever we simulate the nonexperimental portion of a player's strategy, we use a distribution over actions that is within ε of π_* (for that action). Hence by an adaptation of the strong law of large numbers, with probability one on Λ, for every information set h hit infinitely often, the empirical frequencies with which action $a \in A(h)$ is taken *in the nonexperimental portion* of the strategy has limit superior no larger than $\pi_*(a) + \varepsilon$ and limit inferior no smaller than $\pi_*(a) - \varepsilon$. (See Lemma 6.2 from Fudenberg and Kreps (1993).)

Of course, this does not imply that the frequencies of actions actually taken at information sets that are reached infinitely often have lim sups and lim infs within these ranges, because the action actually taken depends on whether an experimental action is taken and, if so, what action that experiment is. But we claim that for information sets $h \in \overline{H}(\pi_*)$, the frequency of visits to h (almost surely on Λ) has strictly positive lim inf. Thus as calendar time goes to infinity, the number of times in which experimentation is permitted has vanishing frequency, and so experimen-

tation will have no impact on the lim inf and lim sup of the empirical frequencies.

To show that for information sets $h \in \overline{H}(\pi_*)$ the frequency of visits to h has strictly positive lim inf (almost surely on Λ) involves an induction on the length of the shortest path of positive probability (under π_*) from the initial node (or, an initial node) to h. The initial information set is certainly reached with nonvanishing frequency. Take any one-action path of positive probability, starting at the initial node. (Let a be this action, and let h be the information set reached.) The lim inf of the occurrence of a in the nonexperimental portion of the behavior rule is at least $\pi_*(a) - \varepsilon$ which is strictly positive, and since the initial information set is reached a nonvanishing frequency of the time, the lim inf of the occurrence of a in the actual strategy (with experiments) is the same as the lim inf of the occurrence of a in the nonexperimental portions. Thus the lim inf frequency with which h is reached is at least $\pi_*(a) - \varepsilon$, which is strictly positive. The induction step should now be apparent. ■

7.3. *Proof of Proposition 7.2*

Once again we give only a sketch. The reader is presumed to be familiar with the proof of Proposition 6.3 from Fudenberg and Kreps (1993).

We mimic this proof almost verbatim. That is, we construct a conforming model in which behavior is precisely myopic, i.e., no experimentation takes place, or $\hat{\pi} = \check{\pi}$, and in each round each player chooses a strategy that is precisely short-run optimal. Players begin with the beliefs γ_*^i that support (in the fashion of a self-confirming equilibrium) play of π_*, and they persist in playing π_* and believing γ_* unless and until data build up that make continued belief in γ_*^i impossible. As long as players continue to play according to π_*, information sets in the complement of $\overline{H}(\pi_*)$ are unreached, so there is no need to change beliefs about what will transpire there. And then, as in the proof of Proposition 6.3 (Fudenberg and Kreps, 1993) one can set the force of asymptotic empiricism so that, with positive probability, what transpires is insufficient to have players abandon their belief in γ_*^i. The details are tedious but straightforward.

8. Unstable Outcomes

A strategy profile is unstable if there is zero probability that the nonexperimental portion of behavior remains forever within an arbitrarily small neighborhood of the target strategy. This definition does not distinguish between (nonexperimental) behavior at on-the-path and off-the-path information sets.

Close examination of the proof of Proposition 7.1 indicates that a weaker

definition and thus a stronger result is available. When a strategy profile is not a self-confirming equilibrium, the proof shows that a defection must occur along the path of play. This suggests that Proposition 7.1 can be extended to show that *outcomes* that do not arise from any self-confirming equilibrium are unstable in an appropriate sense. This extension requires some notation and a definition.

An *outcome* is a probability distribution over the endpoints of the game tree; we use ρ to denote a typical outcome. For a strategy profile π, we write $\rho(\pi)$ to denote the outcome engendered by π. It is trivial that $\rho(\cdot)$ is a continuous function of π.

An outcome ρ is a *self-confirming equilibrium outcome* if there is some self-confirming equilibrium strategy profile π such that $\rho = \rho(\pi)$. An outcome is *not* a self-confirming equilibrium outcome if there is no self-confirming equilibrium that gives this outcome.

DEFINITION. The outcome ρ_* is *unstable* if there exists $\varepsilon > 0$ such that for every conforming model, the probability that $\|\rho(\check{\pi}_t(\zeta_t)) - \rho_*\| < \varepsilon$ for all t is zero.

PROPOSITION 8.1 *If ρ is not a self-confirming equilibrium outcome, then ρ is unstable.*

Here is a sketch of the proof. For any outcome ρ, let $\overline{X}(\rho)$ and $\overline{H}(\rho)$ denote, respectively, the collections of action nodes in the game tree and information sets whose successors (among terminal nodes) have positive probability under ρ. For $x \in \overline{X}(\rho)$ and $a \in A(h(x))$, define

$$\psi(\rho)(x, a) = \frac{\rho(Z(x, a))}{\rho((Z(x))},$$

where $Z(x)$ is the set of all terminal successors of x and $Z(x, a)$ is the set of all terminal successors of (x, a). The following are easily established.

(1) For a general outcome ρ, there may exist nodes x and x' from the same information set h and $a \in A(h)$ such that $\psi(\rho)(x, a) \neq \psi(\rho)(x', a)$.

(2) If $\rho = \rho(\pi)$ for some legitimate strategy π, then $\psi(\rho)(x, a) = \pi(a)$ for all $x \in \overline{X}(\rho)$ and $a \in A(h(x))$. Thus, for a given outcome ρ, if there is a strategy π with $\rho(\pi) = \rho$, then for all nodes x, $x' \in \overline{X}(\rho)$ such that x and x' come from the same information set h, and for all actions a available at that information set, $\psi(\rho)(x, a) = \psi(\rho)(x', a)$.

(3) Conversely, suppose that $\psi(\rho)(x, a) = \psi(\rho)(x', a)$ for all nodes x, $x' \in \overline{X}(\rho)$ such that x and x' are in the same information set and for all $a \in A(h(x))$. Then any strategy π such that $\pi(a) = \psi(\rho)(x, a)$ for $x \in \overline{X}(\rho)$ and $a \in A(h)$ satisfies $\rho = \rho(\pi)$.

(4) $\psi(\rho)$ is continuous in ρ (on its domain of definition).

For each outcome ρ, define

$$\Pi(\rho) = \{\pi \in \Pi : \rho(\pi) = \rho\}.$$

Note that $\Pi(\rho)$ is the empty set for a given ρ when the antecedent of (3) is violated; when the antecedent of (3) is satisfied, then (3) gives an alternate characterization of $\Pi(\rho)$. It is clear from this alternate characterization that $\Pi(\rho)$ is closed. Moreover, if $\pi, \pi' \in \Pi(\rho)$, then π is identical to π' for all $h \in \bar{H}(\rho)$. Finally, $\Pi(\rho)$ has a product structure: If $\pi, \pi' \in \Pi(\rho)$ and we construct a strategy which composes π^i for some players and π'^j for the rest, this third strategy will also lie in $\Pi(\rho)$.

Fix an outcome ρ_*, whose stability (more precisely, whose unstability) is to be investigated. If $\Pi(\rho_*)$ is empty, then there exists some $\varepsilon > 0$ such that $\|\rho(\pi) - \rho_*\| > \varepsilon$ for every strategy π, and thus ρ_* must be unstable by definition.[19] Thus we can assume w.l.o.g. that, for the given ρ_*, $\Pi(\rho_*)$ is nonempty. Let π_* be any (arbitrarily selected) member of $\Pi(\rho_*)$. Note that π_* is completely determined by ρ_* at information sets from $\bar{H}(\rho_*)$.

We are done if we show that ρ_* is unstable under the assumption that $\Pi(\rho_*)$ contains no self-confirming equilibrium strategy profile. First, Lemma 7.1 is extended:

LEMMA 8.1. *If $\Pi(\rho_*)$ contains no self-confirming equilibrium profiles, there exists an $\varepsilon' > 0$ and a player i such that for all beliefs γ^i such that*

$$\gamma^i\left(\left\{\bar{\pi}^{-i} : \max_{j \neq i, h \in H^j \cap \bar{H}(\rho_*)} \|\bar{\pi}^j(h) - \pi^j_*(h)\| < \varepsilon'\right\}\right) > 1 - \varepsilon', \tag{8.1}$$

there exists an s^i such that $u^i(s^i, \gamma^i) \geq u^i(\pi^i, \gamma^i) + \varepsilon'$ for all π^i such that

$$\sup_{h \in \bar{H}(\rho_*) \cap H^i} \|\pi^i(h) - \pi^i_*(h)\| < \varepsilon'. \tag{8.2}$$

In other words, this says that if player i believes that others are likely to play in a manner that would give the outcome ρ_*, then i will prefer some strategy that causes the outcome to differ from ρ_*.

Proof of Lemma 8.1. Suppose to the contrary that for each integer n, for each player i there exist beliefs γ^i_n and a strategy π^i_n such that (8.1) and (8.2) hold for $\varepsilon' = 1/n$ and such that $u^i(s^i, \gamma^i_n) < u^i(\pi^i_n, \gamma^i_n) + 1/n$ for all s^i. Let π_n be the profile where each player plays π^i_n. Since the probabil-

[19] Suppose there exists π_n such that $\|\rho(\pi_n) - \rho_*\| \leq 1/n$ for each n. Take a subsequence along which π_n converges to, say, π_*, and use the continuity of $\rho(\cdot)$ and ψ to derive a contradiction.

ity of any outcome is the product of the probabilities of the actions along the corresponding path through the tree, (8.2) ensures that $\|\rho(\pi_n) - \rho_*\| \leq K/n$, where K is the length of the longest path through the tree.

Now take a subsequence along which π_n and all the γ_n^i converge. Denote the limit strategy profile by π_∞ and the limit beliefs by γ_∞^i. By continuity, $\rho(\pi_\infty) = \rho_*$, and π_∞ is a self-confirming equilibrium, supported by beliefs γ_∞^i. This contradiction proves the lemma. ■

To complete the proof of the proposition, note that as long as players are playing strategies whose nonexperimental parts give an outcome sufficiently close to ρ_*, information sets $h \in \overline{H}(\pi_*)$ will (almost surely) be hit with nonvanishing frequency. (This takes an induction argument as in the proof of Proposition 7.1.) Thus the empirical frequencies of behavior at those information sets will be close to that (uniquely) mandated for all $\pi \in \Pi(\rho_*)$. Applying asymptotic myopia and weak asymptotic empiricism completes the proof. ■

9. Concluding Remarks

9.1. *On Strategically Equivalent Extensive-Form Games*

Since self-confirming equilibrium requires beliefs to be correct along the equilibrium path of play, it is inherently an extensive-form solution concept, in contrast to Nash equilibrium, which can be defined on the strategic form of the game. Two extensive-form games with the same strategic form can have different sets of self-confirming equilibria.

By virtue of Propositions 7.1, 7.2, and 8.1, we have made a case for the appropriateness of self-confirming equilibrium as a solution concept in the learning story we have been telling. This story suggests that two extensive-form games that give rise to the same strategic-form game might be played differently. Put succinctly, when players are learning, how much of their opponents' strategies is revealed matters, and this might depend on the extensive form of the game.

Contrast this with the position taken by Kohlberg and Mertens (1986), that the strategic form encodes all the strategically relevant information, and two extensive-form games with the same reduced strategic form will be played in the same way. This position has been challenged in the past, for example on grounds that extensive-form presentation might affect strategic expectations and thus actual play (e.g., in Kreps, 1990). Here we see a different sort of challenge to this story. In our view, the general problem with the position of Kohlberg and Mertens (except as an assertion about the play of mythical, completely rational beings) is that it does not take into account the process that leads to equilibrium, if indeed an

equilibrium is reached. When these issues are raised, their satisfactory resolution may be misguided by the notion of strategic equivalence.

9.2. *Nonvanishing Trembles and Almost-Nash Equilibria*

While our results indicate that self-confirming equilibria (profiles or outcomes) that are not Nash can be locally stable, there are ways in which our story can be modified that gets us "back" to Nash as the appropriate reduced form solution concept. The story of this sort that we like the best is based on a supposition that players actively experiment with suboptimal strategies and/or actions in a way that generates enough information about off-the-path behavior to preclude non-Nash stable points. This story is quite complex, however, and is the subject of a companion paper (Fudenberg and Kreps, 1994).

A second story is short and can be given here. Suppose players "tremble," in the sense of Selten's trembling-hand perfection. To be precise, suppose that for every action $a \in A$ there is a small probability $\varepsilon_a > 0$ such that for each $a \in A^i$, player i cannot reduce the probability with which a is chosen to less than ε_a. These lower bounds are uniform in time and across histories. (Nothing significant changes if the lower bounds are time and history dependent, as long as they approach some strictly positive limit almost surely.) This ensures that every information set is reached with positive probability, so that for asymptotic empiricism and myopia defined more or less as above (allowing for these trembles), every non-Nash strategy profile of the tremble-constrained game is unstable.[20]

9.3. *Statistical Tests and Odd Histories*

Our formulations of asymptotic empiricism and asymptotic myopia are predicated, at least implicitly, on a belief by each player that his rivals will (asymptotically) play the same strategy profile repeatedly and that the play of different rivals will be independent. But the data provided along particular histories can confound this hypothesis of *asymptotic stability*. We saw instances of this in the subsection on asymptotic independence and in the troublesome example of Section 4 (concerning the value of information in experiments). But other examples could be created, e.g., where the rivals of player i seem to be playing in some cyclical pattern. We have insisted on asymptotic empiricism and asymptotic myopia along *all* histories, including those that present evidence against asymptotic

[20] Moreover, Nash equilibria of the constrained game are exactly the ε-constrained equilibria that Selten uses to define perfection. A trembling-hand perfect equilibrium is the limit point of ε-constrained equilibria as ε converges to zero. Thus for small ε, Nash profiles that are not approximately perfect would also be unstable.

stability. It seems, therefore, that the behavioral assumptions are too strong.

Note that for our results, the sorts of odd histories that would cast doubt on asymptotic stability would have probability zero. That is, our results concern situations in which behavior *is* settling into repeated play of a given strategy profile;[21] and so strong-law type results tell us that odd histories are unlikely to be seen.

This suggests a way in which we can weaken the behavioral postulates that we make, without much change to our fundamental results. We can imagine that each player at each date looks at the history of play and tests statistically whether it seems that play is asymptotically stable. If the data lead to a rejection of this basic hypothesis, the player is not constrained either to hold asymptotically empirical beliefs or to act in an asymptotically myopic manner.

For the results in this paper, it is unnecessary to include these sorts of statistical tests. For Fudenberg and Kreps (1994), these statistical tests become crucial, and hence we leave precise formulations and details to that companion paper. We wish only to signal here that this concept of statistical tests could be used in the current context to give us greater confidence in the assumptions of asymptotic empiricism and myopia.

9.4. *Learning about the Extensive Form*

In our formulation of asymptotic empiricism, we have assumed that players know the informational structure of the extensive-form stage game. (Of course, this was crucial to asymptotic independence.) It is certainly possible to imagine situations in which a player is not sure about the informational structure of the stage game; e.g., he may be unsure whether his opponents observe his action before choosing their own. In this situation, players would infer what they can about the information structure from the history of play. While we do not pursue this idea here, we do wish to indicate that it also can be used, in part, to ameliorate concerns we may have about asymptotic independence or the calendar-time limitations we imposed on conscious experimentation.

APPENDIX: Proof of Proposition 6.1

Fix a self-confirming equilibrium profile π_*. Let γ_*^i (for $i = 1, 2$) be the beliefs that together with π_* satisfy (a) and (b) in the definition of a self-confirming equilibrium. The first step of the proof is to construct a strategy

[21] For the criterion of unstability, play lies within a small neighborhood of such behavior.

profile $\bar{\pi}$ such that (1) for $i = 1, 2$,

$$u^i(\pi^i, \bar{\pi}^{-i}) = u^i(\pi^i, \gamma^i_*), \tag{A.1}$$

where in the expression on the left-hand side, the second argument $\bar{\pi}^{-i}$ is shorthand for beliefs that put a unit mass on $-i$ using the strategy $\bar{\pi}^{-i}$, and (2) $\bar{\pi}$ agrees with π_* at all information sets $h \in \bar{H}(\pi_*)$.

To find $\bar{\pi}$, we use Kuhn's theorem (Kuhn, 1953), which establishes a correspondence between behaviorally mixed strategies and mixed strategies.

Specifically, recalling that Π^i is the set of behavior strategies of player i and letting $\Delta(S^i)$ be the space of mixed strategies for player i, define $\Upsilon^i: \Pi^i \to \Delta(S^i)$ by

$$\Upsilon^i(\pi^i)(s^i) = \prod_{h \in H^i} \pi^i(s^i(h)).$$

For every $\pi^i \in \Pi^i$, $\Upsilon^i(\pi^i)$ is one among many mixed strategies equivalent to π^i in the sense that, whatever $-i$ does, the distribution over endpoints if i uses $\Upsilon^i(\pi^i)$ is identical with the distribution if i uses π^i. (This specific choice of $\Upsilon^i(\pi^i)$ corresponds to independent randomizations by a player at each of his information sets.)

We also define $\Psi^i: \Delta(S^i) \to \Pi^i$ such that for every $\sigma^i \in \Delta(S^i)$, $\Psi^i(\sigma^i)$ is equivalent to σ^i. This takes a bit more work.

For each information set $h \in H^i$, let $H^i_<(h) = \{h' \in H^i : h' < h\}$ and let $H^i_{\not<}(h) = \{h' \in H^i : h' \not< h, h' \neq h\}$. (Because the game has perfect recall, the notion of precedence among information sets of a single player is well defined.) Let $S^i(h)$ be all strategies by i that do not preclude h. That is, $s^i \in S^i(h)$ if, for every $h' \in H^i_<(h)$, $s^i(h')$ specifies the single action in $A(h')$, denoted by $a(h', h)$, that allows play to continue to h. Otherwise, s^i is unrestricted. That is, if we define $\bar{S}^i(h) = \prod_{h' \in H^i_{\not<}(h)} A(h')$, then there is a obvious one-to-one correspondence between $S^i(h)$ and $A(h) \times \bar{S}^i(h)$.

For $a \in A(h)$ for $h \in H^i$, define

$$\Psi^i(\sigma^i)(a) = \frac{\sum_{\{s^i : s^i \in S^i(h), s^i(h) = a\}} \sigma^i(s^i)}{\sum_{\{s^i : s^i \in S^i(h)\}} \sigma^i(s^i)}.$$

In cases where the denominator is zero, any definition will do.

Now define $\bar{\pi}$ by

$$\bar{\pi}^i = \Psi^i \left[\int_{\Pi^i} \Upsilon^i(\pi^i) \gamma^i_*[d\pi^i] \right], \tag{A.2}$$

for $i = 1, 2$. That is, we construct player i's strategy out of $-i$'s beliefs; for each π^i in the support of γ^{-i}, we pass to the corresponding mixed strategy, average over γ^{-i}, and then reconvert to a behaviorally mixed strategy.

We claim that $\bar{\pi}^i$ agrees with π_* at information sets $h \in \overline{H}(\pi_*)$. To this end, fix some information set $h \in \overline{H}(\pi_*)$ and assume that player i moves at h.

Let $\overline{\Pi}^i(h) = \prod_{h' \in H^i_{\xi}(h)} \Delta(A(h'))$; that is, $\overline{\pi}^i$ specifies behavior by i at information sets in $H^i_{\xi}(h)$. Since $h \in \overline{H}(\pi_*)$, so is h' for every $h' \in H^i_{>}(h)$. Since beliefs γ^i_* are not disconfirmed (see part (b) of the definition of a self-confirming equilibrium), every π^i in the support of γ_*^{-i} agrees with π^i_* on h and on $h' \in H^i_{>}(h)$. We can therefore think of γ^i_* as the product of a probability distribution $\bar{\gamma}^{-i}$ on $\overline{\Pi}^i(h)$ and a degenerate measure (at π_*) on the other components of a full behavior strategy. With this definition, for any $s^i \in S^i$ we can write

$$\int_{\Pi^i} \Upsilon^i(\pi^i)(s^i)\gamma_*^{-i}[d\pi^i] = \int_{\Pi^i} \prod_{h' \in H^i} \pi^i(s^i(h'))\gamma_*^{-i}[d\pi^i]$$

as

$$\left[\prod_{h' \in H^i_{<}(h)} \pi^i_*(s^i(h'))\right][\pi^i_*(s^i(h))]\left[\int_{\overline{\Pi}^i(h)} \prod_{h' \in H^i_{\xi}(h)} \overline{\pi}^i(s^i(h'))\bar{\gamma}^{-i}[d\overline{\pi}^i]\right].$$

Moreover, if $s^i \in S^i(h)$, we know that $s^i(h') = a(h', h)$ for $h' \in H^i_{<}(h)$, so we can simplify this term further to

$$\left[\prod_{h' \in H^i_{<}(h)} \pi^i_*(a(h', h))\right][\pi^i_*(s^i(h))]\left[\int_{\overline{\Pi}^i(h)} \prod_{h' \in H^i_{\xi}(h)} \overline{\pi}^i(s^i(h'))\bar{\gamma}^{-i}[d\overline{\pi}^i]\right],$$

which, letting K be the constant $\prod_{h' \in H^i_{<}(h)} \pi^i_*(a(h', h))$, is

$$K[\pi^i_*(s^i(h))]\left[\int_{\overline{\Pi}^i(h)} \prod_{h' \in H^i_{\xi}(h)} \overline{\pi}^i(s^i(h'))\bar{\gamma}^{-i}[d\overline{\pi}^i]\right].$$

Use the definitions of Ψ^i and Υ^i to write out $\bar{\pi}^i(a)$ (from (A.2)) in full detail:

$$\bar{\pi}^i(a) = \frac{\Sigma_{\{s^i \in S^i(h), s^i(h) = a\}} \int_{\Pi^i} \prod_{h' \in H^i} \pi^i(s^i(h'))\gamma_*^{-i}[d\pi^i]}{\Sigma_{\{s^i \in S^i(h)\}} \int_{\Pi^i} \prod_{h' \in H^i} \pi^i(s^i(h'))\gamma_*^{-i}[d\pi^i]}.$$

(Since γ_*^{-i} is correct along the path of play and we are looking at an information set along the path, the denominator of this expression is nonzero, and so the definition applies.) From the previous steps, the numerator immediately simplifies to

$$K\pi_*^i(a)\left[\sum_{s^i\in\bar{S}^i(h)}\int_{\bar{\Pi}^i(h)}\prod_{h'\in H_{\xi}^i(h)}\bar{\pi}^i(\bar{s}^i(h'))\bar{\gamma}^{-i}[d\bar{\pi}^i]\right].$$

Define $K' = \Sigma_{\bar{s}^i\in\bar{S}^i(h)}\int_{\bar{\Pi}^i(h)}\prod_{h'\in H_{\xi}^i(h)}\bar{\pi}^i(\bar{s}^i(h'))\bar{\gamma}^{-i}[d\bar{\pi}^i]$, and this means that the numerator is $K\cdot K'\cdot\pi_*^i(a)$. The denominator is slightly more complex:

$$K\left[\sum_{(a',\bar{s}^i)\in A(h)\times\bar{S}^i(h)}\pi_*^i(a')\int_{\bar{\Pi}^i(h)}\prod_{h'\in H_{\xi}^i(h)}\bar{\pi}^i(\bar{s}^i(h'))\bar{\gamma}^{-i}[d\bar{\pi}^i]\right]$$

$$= K\sum_{a'\in A(h)}\pi_*^i(a')\left\{\sum_{\bar{s}^i\in\bar{S}^i(h)}\int_{\bar{\Pi}^i(h)}\prod_{h'\in H_{\xi}^i(h)}\bar{\pi}^i(\bar{s}^i(h'))\bar{\gamma}^{-i}[d\bar{\pi}^i]\right\}$$

or

$$K\sum_{a'\in A(h)}\pi_*^i(a')\{K'\} = K\cdot K'\cdot\sum_{a'\in A(h)}\pi_*^i(a') = K\cdot K'.$$

Dividing the numerator by the denominator cancels the $K\cdot K'$ terms, and we are left with $\bar{\pi}^i(a) = \pi_*^i(a)$.

The rest is easy. Because π_* is a self-confirming equilibrium relative to the γ_*^i,

$$u^i(\pi_*^i, \gamma_*^i) \geq u^i(\pi^i, \gamma_*^i) \qquad \text{for all } \pi^i\in\Pi^i.$$

Thus by (A.1),

$$u^i(\pi_*^i, \bar{\pi}^{-i}) \geq u^i(\pi^i, \bar{\pi}^{-i}) \qquad \text{for all } \pi^i\in\Pi^i.$$

Since π_*^i is identical to $\bar{\pi}^i$ at all information sets that are hit with positive probability (under π_*, hence under $\bar{\pi}$), we know that

$$u^i(\pi_*^i, \bar{\pi}^{-i}) = u^i(\bar{\pi}^i, \bar{\pi}^{-i}).$$

Thus we know that

$$u^i(\bar{\pi}^i, \bar{\pi}^{-i}) \geq u^i(\pi^i, \bar{\pi}^{-i}) \qquad \text{for all } \pi^i\in\Pi^i$$

(and for i = 1, 2), which means that $\bar{\pi}$ is a Nash equilibrium strategy profile. ■

ACKNOWLEDGMENTS

We are grateful to Robert Anderson, Robert Aumann, Ehud Kalai, David Levine, two referees, and the associate editor for helpful comments. We thank IDEI, Toulouse, and the Institute for Advanced Studies, Tel-Aviv University, for their hospitality while this research was being conducted. The financial assistance of the National Science Foundation (Grants SES 88-08204, SES 90-08770, SES 89-08402, and SES 92-08954) and the John Simon Guggenheim Foundation is gratefully acknowledged.

REFERENCES

AUMANN, R. J. (1987). "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica* **55,** 1–18.

BATTIGALLI, P. (1987). "Comportamento razionale ed equilibrio nei giochi e nelle situazioni sociali," unpublished dissertation, Universitá Commerciale L. Bocconi, Milano.

BATTIGALLI, P., AND GUAITOLI, P. (1988). "Conjectural Equilibrium," mimeo, Bocconi University.

BRANDENBURGER, A., AND DEKEL, E. (1987). "Rationalizability and Correlated Equilibrium," *Econometrica* **55,** 1391–1402.

ELLISON, G. (1993). "A Little Rationality and Learning from Personal Experience," mimeo, Harvard University.

FUDENBERG, D., AND KREPS, D. (1993). "Learning Mixed Equilibria," *Games Econ. Behav.* **5,** 320–367.

FUDENBERG, D., AND KREPS, D. (1994). "Learning in Extensive-Form Games. II. Experimentation and Nash Equilibrium," mimeo, Stanford University.

FUDENBERG, D., AND LEVINE, D. (1993a). "Self-Confirming Equilibrium," *Econometrica* **61,** 523–546.

FUDENBERG, D., AND LEVINE, D. (1993b). "Steady State Learning and Nash Equilibrium," *Econometrica* **61,** 547–574.

HAHN, F. (1977). "Exercises in Conjectural Equilibrium Analysis," *Scand. J. Econ.* **79,** 210–226.

HARSANYI, J. C. (1973). "Games with Randomly Distrubed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points," *Int. J. Game Theory* **2,** 1–23.

KALAI, E., AND LEHRER, E. (1993a). "Rational Learning Leads to Nash Equilibrium," *Econometrica* **61,** 1019–1046.

KALAI, E., AND LEHRER, E. (1993b). "Subjective Equilibria in Repeated Games," *Econometrica* **61,** 1231–1240.

KALAI, E., AND LEHRER, E. (1993c). "Subjective Games and Equilibria," mimeo, 1993.

KOHLBERG, E., AND MERTENS, J.-F. (1986). "On the Strategic Stability of Equilibria," *Econometrica* **54,** 1003–1037.

KREPS, D. (1990). *Game Theory and Economic Modelling*. Oxford: Oxford Univ. Press.

KUHN, H. (1953). "Extensive games and the Problem of information," in *Contributions to the Theory of Games*, (H. Kuhn and A. Tucker, Eds.) Vol. 2, pp. 193–216, Princeton, NJ: Princeton Univ. Press.

RUBINSTEIN, A., AND WOLINSKY, A. (1990). "Rationalizable Conjectural Equilibrium: Between Nash and Rationalizability," mimeo, Northwestern University.

SELTEN, R. (1975). "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *Int. J. Game Theory* **4**, 25–55.