5. Adjustment Models with Persistent Randomness

5.1. Introduction

In the stochastic approximation processes discussed in the previous chapter, the adjustment in the agents' assessments in period t is of order $1/t$, so that the effect of the random terms eventually vanishes, and the long-run behavior of the system corresponds to that of a deterministic system in continuous time. In such models, the sorts of adjustment processes that make sense all have the property that strict equilibria are locally stable, so that if the state converges to a strict Nash equilibrium it stays there.[88] As with the equilibrium refinements literature, these models offer little guidance in predicting which of several strict equilibria is most likely to be observed.[89]

This chapter considers systems in which the step size and effects of random terms both remain constant over time, so that the system is stochastic even in the limit. Recently, Foster and Young [1990] suggested that such processes can be used to select between strict equilibria of a game. Foster and Young studied a continuous-time stochastic system based on the replicator dynamic; we discuss their model in section 5.8, along with the related papers of Cabrales [1993] and Fudenberg and Harris [1992]. Most of this chapter, though, discusses the much larger literature on discrete-time, autonomous, finite-population "stochastic adjustment" models that follow the papers Kandori, Mailath, and Rob [1993] and Young [1993]. In light of the use of the term "mutation" to motivate the study of the ESS concept and the asymptotic stability of the

[88] The foregoing rather loosely identifies Nash equilibria of a game with states of the dynamic process. A more precise formulation would be "the states in which the aggregate (over players) distribution of play corresponds to a strict equilibrium are locally stable steady states."

[89] This is a slight exaggeration, as one might believe that the likelihood of observing various equilibria is correlated with the relative sizes of their basins of attraction. However, this implicitly supposes a more-or-less uniform prior over possible initial positions.

replicator dynamics, it may be worth emphasizing that the classical evolutionary games literature considers only deterministic systems, and that the "mutations" considered there are one-time events.[90] If one thinks of "mutations" as real and recurring, although unlikely, phenomena, it might seem more appropriate to include them explicitly in the model; this is the basic point of the stochastic adjustment literature.

The literature on these stochastic adjustment models is diverse and offers a variety of results on different types of games and adjustment procedures. However, there is one important result on 2x2 games that obtains in many (but not all)[91] such models, and as a result deserves emphasis: this is the selection of the risk-dominant equilibrium as the unique long-run steady state. Of particular importance is the connection (and lack of connection) of risk dominance to Pareto efficiency. In pure coordination games the two concepts are the same. However, in general games, risk dominant equilibria may fail to be Pareto efficient. The conclusion from the study of stochastic adjustment models is that learning procedures tend to select equilibria that are relatively robust to mutations (risk dominant equilibria), and this is a different criterion than Pareto efficiency.

5.2. Overview of Stochastic Adjustment Models

Before turning to the details of the individual papers, it is helpful to have in mind an outline of the sort of procedure the papers generally follow. This procedure, described just below, relies heavily on the idea of the ergodicity of a Markov process, so

[90] Foster and Young [1990] and Fudenberg and Harris [1992] are exceptions that consider stochastic differential equations models of evolution.

[91] For example, the work of Ely [1995] shows that when location is endogenous there is a tendency towards Pareto efficiency. Binmore, Samuelson and Vaughn [1994] also challenge this result, but they use an unusual form for the stage game that makes their results hard to compare with other work in the area.

understanding just what ergodicity entails is crucial; for this reason the Appendix provides a brief review of ergodicity in finite Markov chains.

The procedure has several steps:

Step 0: Specify a "state space." Typically this is either the number of agents of each player population using each action (as in a model of anonymous random matching) or the actions played by each individual agent. The latter case is relevant if different agents of the same player population can behave differently, as in models of local interaction where each agent only interacts with his "neighbors." The state may also include information about the actions played in previous periods. For the time being, we will specialize to the case of a finite state space and discrete time, which is the simplest case mathematically and also the one that has received the most attention in the literature.[92]

Step 1: Specify an "intentional' or "unperturbed" adjustment dynamic, such as the best response dynamic or the replicator dynamic. Most often, this process is deterministic, although the unperturbed model may incorporate randomness from the outcome of the random matching procedure, or because each agent's opportunity to adjust its action arrives randomly. However, the process should be "deterministic enough" that the states corresponding to each of the strict Nash equilibrium are steady states. Typically, the adjustment process also has the "converse" property that in one-shot, simultaneous-move games, only Nash equilibria are steady states.[93] Finally, for the

[92] The approach described here has also been applied to continuous-time, continuous-state processes.
[93] In chapter 7 we consider stochastic evolution when the extensive form is non-trivial.

techniques described below, the unperturbed dynamic should be time-independent, which rules out fictitious play.[94]

As in previous chapters, we will denote the state space by Θ; the Markov transition matrix of the intentional process will be denoted by P. Then if $\theta, \xi \in \Theta$, the element $P_{\theta\xi}$ of this matrix is the probability that the state is θ at date t+1 conditional on the state being ξ at date t. With this convention, probability distributions over states are represented by column vectors φ, and $\varphi_{t+1} = P\varphi_t$.[95]

Step 2: Introduce a "small noise" term; this might correspond to "mistakes", "mutations," or the replacement of old players by new ones. Parameterizing the amount of noise by ε, this gives us a new Markov operator P^ε on the same state space. P^ε should be a continuous function of ε and P^ε should converge to P as $\varepsilon \to 0$; this condition is usually quite natural.

However, the stochastic approximation arguments do not hold for all continuous operators P^ε, since for example the "null" noise operator $P^\varepsilon = P$ is continuous in ε. What is important is that there be "enough" noise in the system. More precisely, the Markov system corresponding to P^ε should be ergodic. In particular, this means that it has a unique invariant distribution; that is, a unique distribution φ^*_ε such that $\varphi^*_\varepsilon = P^\varepsilon \varphi^*_\varepsilon$. With a finite state space there are very simple sufficient conditions for this;, some of which are discussed in the Appendix to this chapter. One important condition is that $\left[P^\varepsilon\right]^n$ is strictly positive for some integer n. Often, the mutation process is defined in such a way that the ergodicity of P^ε is obvious.

---

[94] The basic ideas could be transferred to fictitious play but require different analytical tools. In addition, there are variants on fictitious play that are time-independent.

[95] This follows a standard convention in probability theory. Frequently this literature adopts the opposite convention, that probability distributions over states are row vectors, and the transition probability matrix is the transpose of the matrix considered here.

Step 3: Verify that $\lim_{\varepsilon \to 0} \varphi_\varepsilon^* \equiv \varphi^*$ exists, and determine what it is.[96] Since by definition $\varphi_\varepsilon^* = P^\varepsilon \varphi_\varepsilon^*$, and $P^\varepsilon \to P$, a standard continuity argument shows that $\varphi^* = P\varphi^*$, that is, φ^* is an invariant distribution for the unperturbed process P. Calculating φ^* is ordinarily the hardest step. As we will see, there are various ways of doing this that do not require the explicit calculation of the φ_ε^*.

Step 4: Check whether φ^* is a point mass. If it is, then the corresponding strategy profile is the "stochastically stable equilibrium" in the terminology of Foster and Young. This terminology makes sense when φ^* is a point mass, since the only point masses that are invariant distributions of the unperturbed process P are steady states, and we have supposed that only Nash equilibria are steady states of the unperturbed process. In other words, if φ^* is a point mass, it must correspond to a Nash equilibrium. [97]

Example 5.1: (variant of Canning [1992]): The deterministic unperturbed process corresponds to simultaneous-move Cournot adjustment: There are two populations, player 1's and player 2's, with one agent in each population; each period each player chooses a best response to the action his opponent played in the previous period. The stage game is a symmetric coordination game with payoffs

	A	B

[96] Because the space of distributions over Θ is compact, we know that this sequence has at least one accumulation point. With arbitrary perturbed processes, the sequence might have several accumulation points, so that the limit need not exist, but the sorts of perturbed processes considered in the literature do guarantee the existence of a limit.

[97] However, if φ^* is not a point mass, it need not be a Nash equilibrium . For this reason Canning's [1992] use of the term "equilibrium distribution" to mean "invariant distribution" is unfortunate.

A	2,2	0,0
B	0,0	1,1

so that (A,A) and (B,B) are both equilibria. The Markov matrix may be written as

$$P = \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \text{states} \\ \end{matrix}$$
$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} A,A \\ A,B \\ B,A \\ B,B \end{matrix}$$

and the only steady states are (A,A) and (B,B). However, (0,1/2,1/2,0) is an invariant distribution corresponding to the two-cycle between (A,B) and (B,A).

Now we add noise in the form of a minimum probability ε for each action. That is, when player 1, say, prefers to play A (because player 2 played A last period) player 1 must play B with probability at least ε. Likewise, when player 1 played B last period, there is probability ε that player 2 plays A, so that pr(BA|BA)= ε^2, pr(AA|BA)= $(1-\varepsilon)\varepsilon$, and so on. The perturbed system has the Markov matrix

$$P^{\varepsilon} = \begin{bmatrix} (1-\varepsilon)^2 & (1-\varepsilon)\varepsilon & (1-\varepsilon)\varepsilon & \varepsilon^2 \\ (1-\varepsilon)\varepsilon & \varepsilon^2 & (1-\varepsilon)^2 & (1-\varepsilon)\varepsilon \\ (1-\varepsilon)\varepsilon & (1-\varepsilon)^2 & \varepsilon^2 & (1-\varepsilon)\varepsilon \\ \varepsilon^2 & (1-\varepsilon)\varepsilon & (1-\varepsilon)\varepsilon & (1-\varepsilon)^2 \end{bmatrix}$$

This system is ergodic, with (unique) invariant distribution (1/4, 1/4, 1/4, 1/4). This is not an equilibrium, and is *not* a description of play in any period, although it does correspond to the asymptotic limit of the empirical joint distribution of strategy profiles. (That is, in the long run, each of the four strategy profiles will occur 1/4 of the time.)

This example raises the question of when is it reasonable to hope that limit distribution is a point. Friedlin and Wentzell's [1982] basic insight is that limit distribution is concentrated on a subset of the ω-limit sets of the deterministic process.[98] If the deterministic process has stable cycles then there is no reason to expect that adding noise will take them away. For this reason, the stochastic adjustment approach has been applied to classes of games where the deterministic dynamics have no stable cycles. The leading and simplest such class is a homogeneous population playing a symmetric 2×2 game with two strict equilibria.[99]

5.3. Kandori-Mailath-Rob Model

In Kandori-Mailath-Rob, a single population of N players plays a symmetric 2×2 game. Denote the two actions by A and B. We will focus on the most interesting case, in which there are strict equilibria at (A,A) and (B,B), and a mixed equilibrium in which the probability of strategy A is α^*. We will assume that α^* is less than ½, so that the best response to (1/2 A,1/2B) is to play A; this means that the equilibrium at (A,A) is risk dominant.

Step 0: The State Space of the Process

The state of the system θ_t at date *t* is the number of players using strategy A.

[98] Recall from chapter 1 that the ω-limit sets of a dynamic process are the points that are reached infinitely often from at least one initial condition. In the deterministic finite-state case we consider here, the only ω-limits are steady states and cycles.

[99] Asymmetric populations playing such games can cycle, as in the persistent miscoordination in the example illustrated in Figure 2.3.

Let $u_A(\theta_t)$ and $u_B(\theta_t)$ denote the payoffs from playing strategies A and B, respectively, against the mixed strategy $(\theta_t / N, (N-\theta_t)/N)$ corresponding to a randomly drawn player from a population where θ_t players are playing strategy A.

Step 1: The Deterministic Process P

Kandori-Mailath-Rob take $\theta_{t+1} = P(\theta_t)$, where the only condition on the adjustment dynamics is that $\text{sgn}(P(\theta_t) - \theta_t) = \text{sgn}(u_A(\theta_t) - u_B(\theta_t))$ at all states where not all agents are using the strategy with the highest current payoff. They call such a dynamic "Darwinian." In the two-action games they consider (but not more generally) Darwinian dynamics are "aggregate monotonic" in the sense of the Samuelson and Zhang paper discussed in chapter 4; two-action games also have the special property that all Darwinian dynamics on a single population have the same basins of attraction for each steady state. The desired interpretation is that each period, some players observe the state of the system and choose the strategy that is the best response to last period's state. There are two small points that might cause concern about this interpretation at first sight, but neither one turns out to matter:

a) Players include themselves in computing the play of a randomly drawn opponent. (If players only look at their opponents, then for a given θ_t, the distribution of opponents' play depends on which strategy the player is currently using.) However, for reasonably large *N*, this should not matter, and at the cost of a bit more notation it is easy to extend the conclusions below to this more realistic case where agents do not count themselves in the sample.

b) The most obvious model that leads to the purely deterministic process considered here is the one where all players adjust each period. This yields the best-response dynamic

$$\theta_{t+1} = BR(\theta_t) = \begin{cases} N & u_A(\theta_t) > u_B(\theta_t) \\ \theta_t \text{ for } & u_A(\theta_t) = u_B(\theta_t) \\ 0 & u_A(\theta_t) < u_B(\theta_t) \end{cases}$$

This is also the case where the results are easiest. However, it is not the case on which Kandori, Mailath and Rob want us to focus: If all players adjust each period, it is not obvious why players should choose their actions to maximize payoffs given the previous period's state. Kandori, Mailath and Rob note that the myopic response makes more sense if only a few players adjust each period, so that the current state is more or less locked in for a while. Indeed, if only 1 player adjusts each period, and players are sufficiently impatient, then as in the alternating-move Cournot model, myopic response is optimal and indeed consistent with a perfect-foresight equilibrium. On the other hand, if the one player who has the opportunity to adjust his play is chosen at random from the population, then the adjustment process will be stochastic: Whether the state changes depends on whether the player who gets to adjust is currently playing the best response. However, the analysis of the model will show that this does not matter, since only the speed of this modified process is stochastic, and not its direction.

The more important interpretational issues are the same as in the alternating-move Cournot process, namely the requirement of myopic responses, which requires a combination of impatience and lock-in that may not be plausible in the desired applications of the model. We also question the authors' view that the model describes an

interesting process of learning: since the players who adjust are perfectly informed of the current state, and this is all they care about, it is not clear what they might be learning about. At best this model must be viewed as a rough approximation to a model in which players are less perfectly informed. As should be clear, our preference is for models that have a stronger learning theoretic foundation.

In our opinion, myopic best responses can be best viewed as the limit as the "memory" shrinks of systems where players best respond to the empirical distribution of play over the last few periods, and since for reasonably long memories such systems do seem like learning systems, the distinction here is not hard and fast but rather turns on how long a memory length is reasonable. As we will see in discussing Young [1993] such bounded-memory systems can be analyzed with the same techniques. Moreover, although the memory length can alter the stochastically stable set in some games, it does not do so in the 2x2 games considered by Kandori, Mailath and Rob.

Step 2: Add Noise to Get a Process P^{ε}, and verify that P^{ε} is ergodic.

Suppose that each period, after players have computed their intended adjustments, and before the game is actually played, each player each period "mutates" or is replaced with probability 2ε; mutants are equally likely to initially adopt either strategy, and henceforth follow the deterministic adjustment process. Note that all players have a chance of mutating, and not just those who are "consciously adjusting" their play. For example, even if only one player adjusts at a time there is a positive probability that the entire population mutates at once.

Ergodicity follows from the fact that every state has a positive probability of mutating into any other state.. Note that if only some players adjust each period, and only these can mutate (so that mutations look like trembles) then the system is no longer

strictly positive, but is still ergodic. (See the Appendix for sufficient conditions for ergodicity.)

Step 3 Computing the Limiting Distribution

Let N^* be the least integer greater than $N\alpha^*$; if $\theta_t \geq N^*$ the best response is to play action A. Recall that $\alpha^* < 1/2$.

Proposition 5.1: Suppose that N is large enough that $N^* < N/2$, then the limit φ^* of the invariant distributions is a point mass on the state $\theta_t = N$ corresponding to all agents using action A.

This result is easiest to prove for the case of the best response dynamic, for then the long-run behavior of the system can be determined by analyzing a two-state process: this has also been noted by Canning [1992]. The key idea is that each steady state has a basin of attraction, and intentional play depends only on which of these two basins the state is in, and not its location within the basin. The only way to move from one basin to the other is through simultaneous mutation by a number of players. Moreover, it takes more players to mutate to move from the basin of the risk-dominant equilibrium A to the basin of B than vice versa. As the probability of mutation ε gets small, the probability of M simultaneous mutations or more is of order ε^M. Since it takes fewer mutations to get from A to B than vice versa, this means that the odds of moving from A to B become infinitely greater than the odds of moving from B to A. This in turn means that the process must spend much more time in the A basin than the B basin, and that the invariant distribution places far more weight on A than on B.

Proof of Proposition 5.1: Let $D_A = \{\theta_0 \geq N^*\}$ be the basin of attraction of state N (all agents use action A) under the deterministic process P, and let D_B be the basin of state *0* (all agents use action B). All states θ_0 in D_A have the same value of $BR(\theta_t)$ and hence lead to the same probability distribution over states next period. The same is true for all states in D_B. Hence to compute the invariant distribution, it suffices to know the distribution over the two basins. This distribution in turn is determined by the relative probabilities of transitions from one basin to the other. We define $q_{BA} = prob(\theta_{t+1} \in B_B | \theta_t \in B_A)$ and $q_{AB} = prob(\theta_{t+1} \in B_A | \theta_t \in B_B)$. Then we solve

$$\begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} = \begin{bmatrix} 1-q_{BA} & q_{BA} \\ q_{AB} & 1-q_{AB} \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix}$$

to find that $\frac{\varphi_2}{\varphi_1} = \frac{q_{BA}}{q_{AB}}$.

The last step then is to compute the limit of this ratio as $\varepsilon \to 0$. If $\theta_t \in B_A$, the intended state is *N*. In order for θ_{t+1} to be in D_B there will need to be at least $N\text{-}N^*$ mutations into strategy B. Since each of the N players has a chance of mutating, the probability of a transition with exactly $N\text{-}N^*$ transitions is, from the binomial formula, equal to

$$\binom{N}{N^*} \varepsilon^{N-N^*} (1-\varepsilon)^{N^*}.$$

There can also be transitions with more than *N*-*N** simultaneous mutations, but these will be much less likely as the probability of ε of mutation goes to 0. For example transition with $N - N^* + 1$ mutations has probability that is of order $N - N^* + 1$ in ε.

In a similar way, we may compute that any transition from D_B to D_A must involve at least *N** simultaneous mutations, and *N** simultaneous mutations has probability

$$\binom{N}{N^*}\varepsilon^{N^*}(1-\varepsilon)^{N-N^*}.$$

Substituting into the equation for $\frac{\varphi_2}{\varphi_1}$, we conclude that

$\frac{\varphi_2}{\varphi_1} = \frac{\varepsilon^{N-N^*}(1-\varepsilon)^{N^*} + O(\varepsilon^{N-N^*+1})}{\varepsilon^{N^*}(1-\varepsilon)^{N-N^*} + O(\varepsilon^{N^*+1})}$, so that the ratio goes to 0 as ε goes to 0.

☑

It is important to note that the same conclusion would follow if we supposed that mutants are more likely to choose the action B than action A. That is, we could suppose that the probability of mutation is $\varepsilon_A + \varepsilon_B$, with $\varepsilon_B = k\varepsilon_A$ for any positive *k;* this would change the ergodic distribution for any fixed value of ε_A but would not alter the conclusion that the ergodic distribution converge to a point mass on "all A" in the limit as the mutation probability goes to 0. In order to change this conclusion, the ratio $\frac{\varepsilon_A}{\varepsilon_B}$ would need to go to 0 in the limit. If we do not restrict the ratio $\frac{\varepsilon_A}{\varepsilon_B}$ in the limit, and allow the mutation rates to depend more generally on the state, then Bergin and Lipman [1995] show that the limiting distribution may place any weights on the two Nash equilibria.

5.4. Discussion of Other Dynamics

The best response dynamics are easy to analyze because it is transparent that only two states need to be considered to compute the invariant distribution. When the deterministic process *P* evolves more slowly, then computation of the invariant distribution for fixed $\varepsilon > 0$ requires inverting the $N+1 \times N+1$ matrix P^{ε}. This is harder in practice than in theory. Fortunately, the insight of Friedlin and Wentzell

[1982] shows that the two-state calculation is sufficient to compute the limit of the invariant distribution for any deterministic process whose only steady states are 0 and N.

As we remarked earlier, Friedlin and Wentzell's insight is that in the limit of very infrequent perturbations, the stochastic system will spend most of its time in the ω-limit sets of the deterministic process. Consequently, it suffices to consider the much smaller Markov system whose states are the ω-limit sets of the original deterministic process. Intuitively, as the perturbations become rare, the time interval between perturbations becomes very long, so that after each shock the system moves near an ω-limit before the next shock arrives. In 2x2 games with two strict equilibria, the only steady states of any one-dimensional myopic adjustment process are 0 and N,[100] so the general case reduces to that of the best-response dynamic.[101]

Let us develop the general result, since it has proven useful in the more complicated systems arising from other games.[102].

Consider a 1-parameter family $P^\varepsilon \to P$ of ergodic Markov chains on a fixed state space. In order to determine the limit of the corresponding ergodic distributions φ^ε,we need to know the relative sizes of the transitions probabilities $P^\varepsilon_{\theta\xi}$ that are converging to 0. In the example studied above, the elements of P^ε that converge to zero have the form ε^c, where c is the number of mutations required to move from one state to another, so that the number of mutations corresponds to the order (in ε) of the corresponding transition. We will generalize this by defining the cost of a transition to

[100] For simplicity it is supposed that there is no integer that exactly corresponds to the mixed equilibrium.

[101] If there were an integer $z = \alpha^* N$ then there would also be a third steady state corresponding to the mixed equilibrium. It can be shown that this equilibrium would have zero weight under the limiting distribution, but most papers in the literature suppose there is no such integer z for simplicity.

[102] Friedlin and Wentzell developed the basic ideas in what follows, but since their motivation for it was to eventually characterize continuous-time systems they left some gaps concerning discrete-time systems. These gaps have been filled in by Young [1993] and Kandori and Rob [1992].

be its order in ε, so that probabilities proportional to ε have cost 1, probabilities proportional to ε^2 have cost 2, and so on. Formally, we define the *cost* $c(\theta|\xi)$ *of a transition* to state θ given state ξ as

$$c(\theta|\xi) \equiv \lim_{\varepsilon\to 0}\left(\log P^{\varepsilon}_{\theta\xi} / \log\varepsilon\right).$$

Our basic assumption is that this limit exists for every pair θ,ξ. Notice that since $\log\varepsilon$ is negative, the bigger the probability of transition, the smaller the cost. Notice also that if the transition has positive probability in the limit system ($P_{\theta\xi} > 0$), then its cost $c(\theta|\xi)$ is 0.

We now consider moving from an ω- limit set $\omega \subseteq \Theta$ to another set $A \subseteq \Theta$ which need not be an ω- limit set. This move may take place in several steps, so we consider a path $\vec{\theta} = (\theta_0, \theta_1, \theta_2, \ldots, \theta_t)$ where $\theta_0 \in \omega$ and $\theta_t \in A$, and where consecutive states in the path are not required to be distinct. We look for paths that result in the highest probability of transition; this is the same as looking for paths with the least cost. (Because every θ_0 is in the same limit set, it does not matter which one is used to begin the path, since transitions within a limit set have a cost of 0.) Since the probability of a path is the product of the transition probabilities, the cost of the path is the sum of the transition costs $\sum_{\tau=1}^{t} c(\theta_\tau|\theta_{\tau-1})$. This leads us to define

$$\vec{c}(A|\omega) \equiv \min_{\vec{\theta}:\theta_0\in\omega,\theta_t\in A} \sum_{\tau=1}^{t} c(\theta_\tau|\theta_{\tau-1})$$

Our goal is to analyze transition costs between ω-limit sets of the process *P*. Let Ω denote these ω-limit sets. The direct application of Friedlin and Wentzell's technique that we will now present requires that one first determine every member of Ω. Later we will discuss Ellison's less general sufficient condition for stochastic stability that makes do without this step.

Given a finite set Ω and an $\omega \in \Omega$, an $\omega-$tree is a tree on the set Ω in the sense normally used in game theory[103] *except* that the direction of motion is the reverse of the usual one, so that the paths start at many initial nodes and converge at a single "root" which is the unique terminal node of the tree , which is the ω node in an $\omega-$tree. Let H_ω denote the set of all $\omega-$trees, and for any $\omega-$tree h, let $h(\omega)$ denote the successor of ω, and let $D(\omega')$ denote the basin of ω' in the limit dynamic *P*. Note that $\vec{c}(D(h(\omega'))|\omega') = \vec{c}(h(\omega')|\omega')$ because the cost of transitions within a basin of attraction is 0.

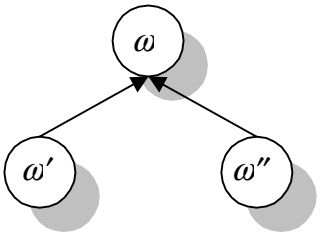
Proposition 5.2 (Freidlin and Wentzell [1982], Young [1993], Kandori, Mailath and Rob [1993]): The limit φ^* of the invariant distributions φ_ε exists, and is concentrated on the limit sets ω that solve $\min_{\omega\in\Omega} \min_{h\in H_\omega} \Sigma_{\omega'\in\Omega/\omega}\vec{c}(D(h(\omega'))|\omega')$.[104]

Note also that the formula sums over all paths for a given $\omega-$tree, as opposed to only counting the cheapest such path. For example, in a simple 3-node $\omega-$tree

---

[103] That is, a directed graph that branches out. See for example Kreps [1990] or Fudenberg and Tirole[1991] for a formal definition

[104]Freidlin and Wentzell give an explicit formula for the ergodic distribution of the perturbed dynamics, which comes from solving the (complicated) linear equation $P^\varepsilon\varphi^*_\varepsilon = \varphi^*_\varepsilon$. Young shows that the limit distribution exists because this formula shows that the relative probabilities of any two states is a ratio of polynomials in the transition probabilities, and these probabilities themselves have been assumed to be polynomial in ε, so that the relative probabilities are polynomial in ε. The formula given in the proposition follows from the observation that the states with positive probability under the limit distribution are those whose probability in the φ^*_ε is of the lowest order in ε. Kandori., Mailath and Rob [1993] give essentially the same argument as Young, but assert that the uniqueness follows from the uniqueness of the φ^*_ε for each ε.



the formula sums $\vec{c}(D(\omega)|\omega')$ and $\vec{c}(D(\omega)|\omega'')$. Appendix 2 to this chapter gives some intuition for this summation and for Proposition 5.2

In complex systems, the cost in Proposition 5.2 may be hard to compute, not least because the computation may require identification of every ω-limit set, but in simple examples it is not too hard. For example, in the simple Kandori-Mailath-Rob case above, the cost of moving from one state to another is just the number of mutations required to get there, so the cost of moving from one steady state to the basin of another is the minimum number of mutations required for the move.[105] Specifically, the limit set $\Omega = \{0, N\}$, consists of the two steady states. The only 0 $\omega-$tree is $N \to 0$, and $\vec{c}(D(N)|0) = N - N^*$; the only N $\omega-$tree is $0 \to N$, and $\vec{c}(D(0)|N) = N^*$, so that the problem is reduced to that of the best response dynamic.

5.5. Local Interaction

In the Kandori-Mailath-Rob model, when the mutation rate is close to 0, although the relative amount of time spent at the risk-dominant equilibrium is much larger than at the other one, However the system is likely to stay at the other equilibrium for a long time if it begins near to it. Indeed, Ellison [1992] argues that for plausible payoff values, and N = 50 or 100 players, the system changes basins so infrequently that for practical

[105] We should mention that we write cost as we write transition probabilities, with the target state first, and the state we condition on second. This is the standard convention in probability theory. Unfortunately, the existing literature in this area uses the opposite convention.

purposes its behavior (for the first 10^5 to 10^{20} periods, depending on the size of the payoff differences) is determined by the initial condition.

An alternative is to consider a model where the players interact only with their neighbors. Here, a few players playing the risk dominant strategies can spread to the entire population quite rapidly, so the ergodic distribution may be a much more interesting description of actual play. This is in fact shown by Ellison [1992], who, like Kandori-Mailath-Rob, considers 2×2 games. Ellison uses this result to argue that "the nature of the matching process is crucial to ... whether historical factors or risk dominance will determine play."

In the Ellison model, the N players are evenly spaced along a circle, and each player is matched against a randomly chosen opponent from his 2 nearest neighbors.[106] Each agent must select a single action to use against both opponents. As in Kandori-Mailath-Rob, players are perfectly informed of last period's state, and hence of the last period's distribution of opponents' play. Players are assumed to choose their actions to maximize their expected payoff against this distribution, so that the unperturbed dynamic is a "local" version of the best-response dynamic. As before, both "all A" and "all B" are steady states; the difference is in how these steady states respond to mutations. As in Kandori-Mailath-Rob, mutations are modeled as the replacement of a player with a newcomer who is equally likely to choose either action.

The key observation is that in the case of local interactions, the steady state "all B" can be upset by a small number of mutations. In this case, it is easy to see that any cluster of two adjacent agents playing A will spread and eventually take over the entire population: each of the two agents in the cluster assigns probability at least 1/2 to his next

[106] Ellison also considers the case of interactions with the $2K$ nearest neighbors. His strongest results in this case are in his [1995] paper.

opponent playing A, and so sticks with A; moreover, each of the two agents on the boundary of a cluster of A's assigns probability 1/2 to his next opponent playing A, and so switches from B to A. This means that random events that shift the process from the state "all B" to the basin of the state "all A" have an arrival rate of order ε^{-2}, independent of the total population size *N*. In contrast, in the "uniform-matching" model of Kandori, Mailath and Rob, the arrival rate is $\varepsilon^{-\alpha^* N}$. This insight is not quite a proof, but it does suggest why the convergence speed will be much faster in the local interaction model.

Before discussing speeds of convergence, we should first check that, as in Kandori-Mailath-Rob, the limit of the ergodic distribution as the mutation rate goes to zero is a point mass on "all A". We again apply the process outlined above.

Step 0: The State Space

Because location of the agents matters, the state space here is the set $\Theta = \{A, B\}^N$ of *N*-vectors whose components specify the actions of the individual agents.

Step 1: The Deterministic Process

We begin by examining the deterministic system in order to compute the ω-limit sets of the intentional adjustment process, and characterize their basins of attraction. Note that under the unperturbed dynamic, the number of players playing A can never decrease, for if j players play A at date t, all of their neighbors play *A* at t+1. Moreover, a cluster of two adjacent A's leads to "All A." Notice that in addition to the steady states "All A" and "All B", when *N* is even there is one other ω-limit set, namely the two-cycle between the states "ABAB...", "BABA...". We can see that the steady states and basins of attraction of this process are:

ω_1) "all B;" the basin of this state is simply the state itself.

ω_2) (which only exists if N is even) is the two-cycle just mentioned. Its basin includes at least the two states in the cycle

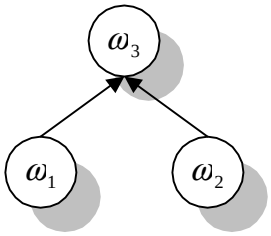
ω_3) "all A; " the basin of this state includes at least all states with 2 adjacent A's; and any state with a string "ABBA"

Step 2: Adding mutations, it is a simple observations that the system is ergodic.

Step 3: Computing the Limiting Distribution

To compute φ^*, compute $\min_{\omega\in\Omega}\min_{h\in H_\omega}\Sigma_{\omega'\in\Omega}\vec{c}(D(h(\omega'))|\omega')$. Ellison shows that for N even, the minimum cost is 2, and that it is given by the $\omega-$tree $\omega_1\to\omega_2\to\omega_3$; for N odd the minimum cost is 1. In the case of N even, the costs of both ω_2 given ω_1 and that of from ω_3 given ω_2 are 1. Given ω_1 (all B) a single mutation leads to "ABB..."; the deterministic process then goes to the two-cycle ω_2. From either point in the two cycle ω_2 a single mutation then leads to a state with 2 adjacent A's; the deterministic process then goes to ω_3.

We must now check that 2 is actually the minimum cost. First note that there are two other $\omega-$trees with root ω_3, $\omega_2\to\omega_1\to\omega_3$ and



In both cases, the cost of ω_3 given ω_1 is greater than one: a single mutation leads to the cycle not the steady state at "all A," so two mutations adding adjacent A's are needed to get into the basin of ω_3. In the first case we must add in the cost of ω_1 given ω_2, which

is *N*/2 (since the number of A's cannot decrease). In the second case we must add in the cost of ω_3 given ω_2, which we already noted is 1. In both cases, the cost exceeds 2.

Finally we must consider $\omega-$tree with roots ω_1 and ω_2. We claim that (if *N*>4) these have cost higher than 2. Any ω_1-tree must have path from ω_3 to ω_1; since the deterministic process never decreases the number of A's, this cost is at least *N*. For the same reason, cost of any ω_2-tree is *N*/2. Thus the stochastically stable set is "all A."

In light of Kandori-Mailath-Rob, this conclusion is not surprising, but one should note that this dynamic and the Kandori-Mailath-Rob dynamic need not have the same stochastically stable sets in 3x3 (or larger) games. The reason for this is simple: In the two-neighbor model, the stochastically stable set is determined by the best response to the 6 possible configurations of neighbors, while the stochastically stable set with uniform matching and a large population depends on how players respond to every mixed strategy that can be generated by some state of the aggregate population. As a consequence, the stochastically stable set with uniform matching may depend on the details of the payoffs that do not matter in the two-neighbor model.[107]

Returning to 2x2 games, the interesting observation is that the convergence times are faster under local interaction. This is most easily seen in simulations: supposing the system starts in state "all *B*", how long does it take to get to state "all *A*"? This depends on the payoffs (which determine α^* and hence how many mutations are needed in Kandori-Mailath-Rob) and also on *N* and on the probability ε of mutation. Note that with only 2 neighbors, all agents want to play A if they have at least one neighbor playing A, so that the speed of convergence is the same for all values of $\alpha*(<1/2)$.With more than

[107] Ellison gives an example of this, based on an example of Young [1993] used to show how the equilibrium selected under uniform matching can depend on the details of the payoff matrix.

two neighbors, this is no longer the case, and convergence can be faster in games where $\alpha*$ is smaller.

We now compare rates of convergence between the global and local interaction cases for $\alpha^* = 1/3$. In the Kandori-Mailath-Rob global interaction case with the best-response dynamic, the expected waiting time until everyone plays A can be computed analytically. They are displayed in the table below, where the entries are the expected waiting time for passage from the state "all B" to the state "all A."

	$\varepsilon = .025$	$\varepsilon = .05$	$\varepsilon = .1$
N=50	10^{14}	10^{9}	10^{5}
N=100	10^{27}	10^{17}	10^{9}

By way of contrast, in the two neighbor local interaction case, the expected waiting times until 75% of the population plays *A* with 2 neighbors can be found numerically from a simulation

	$\varepsilon = .025$	$\varepsilon = .05$	$\varepsilon = .1$
N=50	11	8	6
N=100	11	8	6

With 12 neighbors, rather than 2, we find the expected waiting time for 75% A

	$\varepsilon = .025$	$\varepsilon = .05$	$\varepsilon = .1$
N=50	460	46	11

Ellison confirms these simulations with an analytic result. Finite state Markov processes always converge at an exponential rate given by the second largest eigenvalue of the transition probability matrix (the largest eigenvalue is always 1). Ellison computes this eigenvalue to determine how rapidly the system converges from the initial condition to the ergodic distribution. For the Kandori-Mailath-Rob model Ellison shows that the exponential rate is $\varepsilon^{\alpha^* N}$, which is intuitive, as this is the probability of the required number of mutations. In contrast, with 2 neighbor matching the exponential rate for small ε is approximately independent of N, and is of order ε.

5.6. The Radius and Coradius of Basins of Attraction

So far we have characterized the stochastically stable set using Friedlin and Wentzell's technique of constructing "trees" that link the limit sets of the unperturbed stochastic process, assigning a cost to each $\omega-$tree, and then determining which $\omega-$tree has the lowest cost. This approach has the advantage that it can in principle always be applied, but it has two related drawbacks. First, the method may require that one determine *all* of the limit sets of the unperturbed process. This determination was straightforward in the systems Kandori-Mailath-Rob and Young considered, but can be difficult in systems with a large number of limit sets. Second, determining the least-cost $\omega-$tree can be a complicated graph-theory problem. Recently, Ellison [1995] has provided an alternative and much simpler sufficient condition for a set to be stochastically stable. This sufficient condition is not necessary, so the technique is not useful in all cases, but when it does apply it is has the additional benefit of yielding the rate of convergence as well as the identity of the limit set.

Ellison's condition is based on two concepts, the *radius* and *coradius* of a limit set, which are defined in terms of the costs of various transitions. We continue to use $\omega \in \Omega$ for ω-limit sets of the limit dynamic P, and $D(\omega)$ for the basin of ω. The *radius of* ω is just the cost of leaving $D(\omega)$. Let us also denote $\sim D \equiv \Theta \setminus D$. Formally,

$$R(\omega) \equiv \vec{c}(\sim D(\omega)|\omega)$$

In the Kandori-Mailath-Rob model the least cost path away from a steady state is a direct jump, that is, the path considered has only two elements, the initial state in ω and the subsequent state which is not in $D(\omega)$. This is true more generally if (i) the shocks take the form of i.i.d. mutations, and (ii) in the basin of each limit set ω, $\min_{\theta \notin D(\omega)} c(\theta|\theta_t)$ is non-decreasing under the limit dynamic. The first condition says that we may simply measure cost by the number of mutations required to reach a given point, so that, for example, the cost of a three-state sequence which has two mutations in the first period and one in the second is the same as the cost of a three-mutation transition. Condition (ii) requires that the deterministic dynamic cannot decrease the cost of getting out of the basin.

When the least cost path is a direct jump, the way to show that the radius equals some particular *r* is to first exhibit a direct jump out of the basin with cost *r*, and then argue that any direct jump of lower cost must remain in $D(\omega)$.

Intuitively, the radius measures how easy it is for perturbations to push the system out of $D(\omega)$, and hence captures the expected time the system remains in $D(\omega)$ each time this basin is entered. From this perspective, the other datum we need is a measure of how quickly mutations return the system to $D(\omega)$ from states outside of it.

The simplest such measure is the *coradius* of a limit set. The *coradius* of ω, denoted $CR(\omega)$, is defined by

$$CR(\omega) = \max_{\theta} \vec{c}(\omega|\theta).$$

This expression will be used to give a bound on the wait until the system returns to $D(\omega)$ that is useful in some applications, but the bound is not tight; Section 5.7 discusses the tighter bound given by the "modified coradius."

Proposition 5.3 (Ellison [1995]): If there is a limit set ω such that $R(\omega) > CR(\omega)$, then very stochastically stable state is contained in ω.

proof: This is a consequence of the more general proposition 4.5 below.

As one application of this result, consider extending the sort of "Darwinian dynamics" studied by Kandori-Mailath-Rob to symmetric, 2-player, *M*x*M* games as follows: Say that the deterministic dynamic P is "best-response-respecting" if $[P(\theta)]_i > \theta_i$ whenever the *i*th pure strategy is a best response to the mixed strategy corresponding to θ. This is a very weak form of monotonicity; it reduces to KMR's "Darwinian" condition in 2x2 games.[108] A symmetric equilibrium (A,A) is "p-dominant" (Morris, Rob and Shin [1993]) if A is a strict best response to any mixed strategy that places probability at least p on A.[109] In 2x2 games 1/2 dominance is equivalent to risk dominance; in *N*x*N* games it is more restrictive than the pairwise risk dominance notion[110] proposed by Harsanyi and Selten.

Proposition 5.4: If (A,A) is a 1/2-dominant equilibrium, then for all sufficiently large populations the stochastically stable set obtained by perturbing any best-response-respecting dynamic is a point mass on all agents playing A.

---

[108] Small but obvious modifications of the definition and of proposition 5.4 are required to handle the case where each player responds to the distribution corresponding to the play of the other *N*-1 agents in the population.

[109] Note that p-dominance for a given p implies p' dominance for all $p' \leq p$.

[110] One strategy pairwise risk dominates another one if it risk dominates it in the 2x2 game formed by deleting all other strategies.

Proof: Since A is ½ dominant, the radius of "all A" is at least N/2. . The coradius of "all A" is bounded by the number of mutations required to directly jump to a point in the basin of all A. Because A is 1/2- dominant it suffices that the fraction of A's be slightly less than 1/2, which suggests that the coradius should be less than qN for some $q < 1/2$. Due to the finite population size the least integer greater than qN may actually be greater than (N-1)/2, but this can be avoided by taking N sufficiently large.

☑

This proof is essentially the same as that for the 2x2 case studied in Kandori-Mailath-Rob. Although the hypothesis could be weakened using the notion of modified coradius discussed below, it is sufficient to include several results from the literature. First, Kandori and Rob [1995] show that in pure coordination games the Pareto-optimal equilibrium is selected; in such games the Pareto-optimal equilibrium is 1/2 dominant.

Second, Kandori and Rob [1993] consider symmetric coordination games with the "total bandwagon property" that the best responses to any distribution are in the support of that distribution (so that action A cannot be a best response to a distribution in which all agents use B or C) and the "monotone share property" which says that if S' is a strict subset of S, then the (unique) mixed strategy equilibrium $m^*(S)$ with support S gives each pure strategy in S' a strictly smaller probability than does the (unique) mixed strategy equilibrium with support S'. For generic payoffs, the only ω-limit sets in these games are the states where all agents choose the same action. Kandori and Rob show that their assumptions on the payoff functions imply that the cheapest path from one ω-limit set to another is a direct jump to the corresponding basin, as opposed to a path that first jumps to some third equilibrium., so that the modified coradius and the radius are the same. This allows Kandori and Rob to determine the stochastically stable equilibrium in the case of three actions by explicitly solving for the least-cost $\omega-$tree, but the

minimization is too complicated to be solved in general. Instead, Kandori and Rob show that if, in addition to their assumptions, there is a single equilibrium that pairwise risk dominates all of the others, then that equilibrium is stochastically stable. It is straightforward to check that the equilibrium in question must be 1/2 dominant. Moreover, pairwise risk dominance and the total bandwagon property are enough to imply 1/2 dominance; the monotone share assumption is not needed.

Third, Ellison's analysis extends immediately to *I*-player games played by a single population of players, provided that ½- dominance is extended to mean that action A is ½- dominant if it is a best response to any mixed strategy profile in which each opponent gives probability at least 1/2 to A. (We should point out that we have not seen a definition of 1/2-dominance for *I*-player games in the literature.) Note that this reduces to the original definition in two-player games, because the payoff to action A against a mixed strategy of the opponent is linear in the opponent's randomizing probabilities.) This extension of 1/2dominance reveals the structure behind Kim's [1993] analysis of symmetric *I*-person coordination games in which each player has only two actions. In contrast to two-player, two-action games, *I*-player two-action games need not have a 1/2 dominant action. However, Kim assumes that a player's payoff depends only on his own action and the total number of opponents playing the same action, and that the payoff to using an action is increasing in the number of opponents that use it; this implies that there are only two pure Nash equilibria, the ones in which all players play the same action. Moreover, the action that is the best response when each opponent randomizes 1/2-1/2 is also the best response to any profile in which each opponent gives the action probability at least 1/2; in other words that action is 1/2 dominant. In particular, except in one knife edge case, one of the two pure equilibria in a game of this type must be 1/2 dominant.

This explains why Kim finds that the long-run equilibrium is the one that is 1/2 dominant in the sense defined above.[111]

5.7. The Modified Coradius

The coradius gives an upper bound on the expected time until a return to ω, but a tighter bound called the modified coradius is available, which turns out to be useful in a variety of settings. The insight behind the idea of the modified coradius is that the most probable path from one basin of attraction to another need not involve jumps due to perturbations in consecutive periods, provided that the intermediate points are themselves steady states. In this case the system may simply remain at the intermediate steady state for a while, before moving on to the next steady state, and this is far more likely to happen than two jumps in consecutive periods.

With this motivation, we now define the *modified coradius* of a limit set. Let $\omega_1(\vec{\theta}), \omega_2(\vec{\theta}), \ldots \omega_I(\vec{\theta})$ be the sequence of limit sets through which the path $\vec{\theta}$ passes, with the convention that a limit set can appear on the list several times, but not consecutively. The modified coradius is just

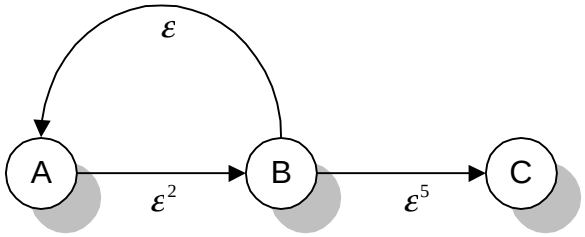
$$CR^*(\omega) = \max_{\theta \notin D(\omega)} \min_{\vec{\theta}|\theta_0=\theta, \theta_t \in \omega} \sum_{\tau=1}^{t} c(\theta_\tau|\theta_{\tau-1}) - \sum_{i=2}^{I-1} R(\omega_i(\vec{\theta})).$$

Note that by construction $CR^*(\omega) \leq CR(\omega)$.

The definition of the modified coradius involves several subtleties, all of them linked to the fact that the modified coradius is only useful in obtaining a bound on the worst-case expected waiting time. First of all, while the modified coradius provides a

[111] Kim also shows that for this class of models, other dynamic adjustment procedures, for example rational expectations with lock-in, leads to a different conclusion about which equilibrium is selected in the very long-run.

correct upper bound on the worst-case waiting time, the bound need not be tight. Moreover, the modified coradius at limit sets ω' that are not maximizers of the expression above does not bound the corresponding waiting time, nor does the modified coradius give a correct bound at every ω where the maximum waiting time is obtained. We will not explain all of these complications here, but the following example should at least indicate why the modified coradius may provide a better bound than the unmodified one in some simple cases:



Here $P_{BA} = \varepsilon^2, P_{AB} = \varepsilon, P_{BC} = \varepsilon^5$, and $c(B|A) = 1, c(A|B) = 2, c(B|C) = 5$. Consequently, the coradius of C is the sum of cost going from A to C, equal to 7. According to this calculation, the amount of time to get from A to C should be on the order of ε^{-7}. The modified coradius subtracts from this the radius of B, namely 1, resulting in a cost of 6, suggesting a waiting time of ε^{-6}. How long does it actually take to get from A to C in this example? The system has probability of ε^2 of moving from A to B. Once at B it is likely to remain for quite some time. However, it is more likely to return to A than move on to C. Indeed, the system must return to A ε^{-4} times on average, before moving on to C. Each trip from A to B back to A takes roughly ε^{-2} periods, so the length of time before hitting C is roughly $\varepsilon^{-4}\varepsilon^{-2} = \varepsilon^{-6}$. This is the calculation made by the modified coradius.

Proposition 5.5 (Ellison [1995]): If there is a limit set ω such that $R(\omega) > CR^*(\omega)$, then

(a) Every stochastically stable state is contained in ω.

(b) For any $\theta \notin \omega$, the expected waiting time until ω is reached, starting at θ, is of order (at most) $\varepsilon^{-CR^*(\omega)}$ as $\varepsilon \to 0$.

Sketch of proof: Let φ_ε denote the unique invariant distribution corresponding to an ε perturbation. For part (a), it is enough to show that $\frac{\varphi_\varepsilon(\theta)}{\varphi_\varepsilon(\omega)} \to 0$ as $\varepsilon \to 0$ for all $\theta \notin \omega$.

As a first step towards this goal, we claim that

(*) $$\frac{\varphi_\varepsilon(\theta)}{\varphi_\varepsilon(\omega)} = \frac{\text{E}\{\#\text{of times } \theta \text{ occurs before reaching } \omega | \theta\}}{\text{E}\{\#\text{of times } \omega \text{ occurs before reaching } \theta | \omega\}}$$

where the expectation in the numerator is with respect to the ergodic distribution. The distribution in the denominator is more complicated, as it requires using an expectation over states in ω conditional on the state having entered the set ω; fortunately all that will matter about this distribution is that it can be bounded below uniformly over all distributions on ω .

To see why (*) is correct, consider the auxiliary two-state (non-Markov) process formed by taking realizations of the original process and mapping state θ to state 1 and every state in ω to state 2, and omitting all periods in which other states occur. For example, if the original process is in state θ in periods 1 and 2, in some state θ' outside of ω in period 3, in ω in periods 4 and 5, and in state θ in period 6, then the first 5 realizations of the auxiliary process are (1,1,2,2,1...). The relative frequency of 1's and 2's in the auxiliary process is the same as the relative frequency of θ and ω in the original one, and moreover in the auxiliary process these relative frequencies are simply the relative sizes of the run lengths.

Examining equation (*), the numerator of the RHS is at most the waiting time to reach ω from θ and the denominator is bounded below by a non-vanishing constant

times the minimum over states in ω of the expected waiting time to reach a state outside of $D(\omega)$. Thus the proof of both parts of the theorem boils down to showing that the waiting time to leave $D(\omega)$, starting in ω, is approximately $\varepsilon^{-R(\omega)}$, while the waiting time to reach ω is of order $\varepsilon^{-CR^*(\omega)}$.

We will not prove these facts here, though the proof of the first is not hard. We recommend instead that the reader check them in simple 2 or 3 state examples.

☑

As a final application, suppose that a game with a 1/2-dominant equilibrium is played in a model of local interaction on a two-dimensional lattice. Specially, consider an $N_1 \times N_2$ lattice on the surface of a torus, and suppose that the limit system P is given by each player choosing the strategy that is a best response to the distribution of strategies used by his four immediate neighbors in the previous period.[112] Unlike Ellison's one-dimensional model, this system has a large number of steady states. Define a *vertical stripe* to be a location in the first dimension such that all players at that location play the same strategy. If there are only 2 actions, A and B, with A being 1/2 dominant, then any state formed of vertical stripes is a steady state provided that there are at least 2 adjoining B-stripes between each A-stripe. We can similarly form equilibria consisting of horizontal stripes. Yet another type of steady state is for all players play *B* except for 2x2 rectangles of players playing A surrounded by opponents playing B.

Now consider perturbing the dynamic with the now-familiar stochastic replacements: each period each player being replaced with i.i.d. probability ε. What is the stochastically stable set? This model, unlike the one-dimensional model of Ellison [1993], does not have the "contagion" property, where a one-time occurrence of a few

[112] As in previous papers on local interaction , for example, the Ellison [1993] paper discussed earlier, and Blume [1993] who studied the play of 2x2 games in an infinite two-dimensional lattice, this assumes that players are constrained to use the same action when paired with each of their neighbors.

mutations (only 2 in the 2-neighbor model) is enough to send the system from "all A" to "all B." Instead, starting from "all B", four simultaneous adjacent mutations sends the system to a steady state with a 2x2 rectangle of A's. Clearly, the program of first determining all of the limit sets of this dynamic, and then finding the least cost $\omega-$trees that connect them, would require a great deal of computation. However, Ellison shows that the limit distribution is a point mass on all players using the 1/2-dominant action and moreover the expected waiting time is of order ε^{-3}, where "3" is the modified coradius of all A.

This shows that the fast convergence times of the one-dimensional model do not require that model's property of contagion. Instead, "fast convergence" obtains because the system can move from all B to all A by a sequence of jumps from one steady state to another, with each jump having waiting time at most ε^{-3}. This suggests that, *ceteris paribus,* convergence times will be quicker in models with many intermediate steady states than in models where direct jumps between pure strategy equilibria are the quickest paths.

5.8. Uniform Random Matching with Heterogeneous Populations

Kandori, Mailath, and Rob [1993] considered a single homogenous population of agents playing a 2x2 symmetric game. As we saw, in that model the stochastically stable outcome is the risk-dominant equilibrium for any "Darwinian" adjustment dynamics. However, KMR acknowledge that that this robustness to the specification of the adjustment dynamics does not extend to the case where there are distinct populations of player 1's and player 2's, a case analyzed by Hahn [1995].

Hahn's model follows Kandori-Mailath-Rob in every detail except that he assumes there are two populations and allows the game played to be asymmetric. Recall in Kandori-Mailath-Rob the two actions are denoted A and B Hahn defines the state space so that the state $\theta_t = (\theta_t^1, \theta_t^2)$ at time t is now the number of agents in each of the two populations who are playing A. . Following Kandori-Mailath-Rob, Hahn assumes that the unperturbed dynamic has the following form:

$$\theta_{t+1}^i = P_i(\theta_t) = \theta_t^i + f^i(\theta_t^{-i}),$$

where the f^i are monotone increasing and satisfy $\text{sgn}(f^i) = \text{sgn}\left(u^i(A, \theta_t^{-i}) - u^i(B, \theta_t^{-i})\right)$.

We refer to a system satisfying this condition as *monotone.*

Continuing to follow Kandori-Mailath-Rob, suppose that the game has two strict equilibria and an equilibrium in mixed strategies, denoted $(\alpha^{*1}, \alpha^{*2})$. Since there are two populations, though, the system has two dimensions, and not one, with two stable steady states corresponding to the pure strategy equilibria and an unstable saddle at the mixed equilibrium. One expects that this since this equilibrium is unstable it should have probability 0 in the ergodic distribution; to simplify Hahn suppose that the equilibrium mixing ratios cannot be attained in any state.[113]

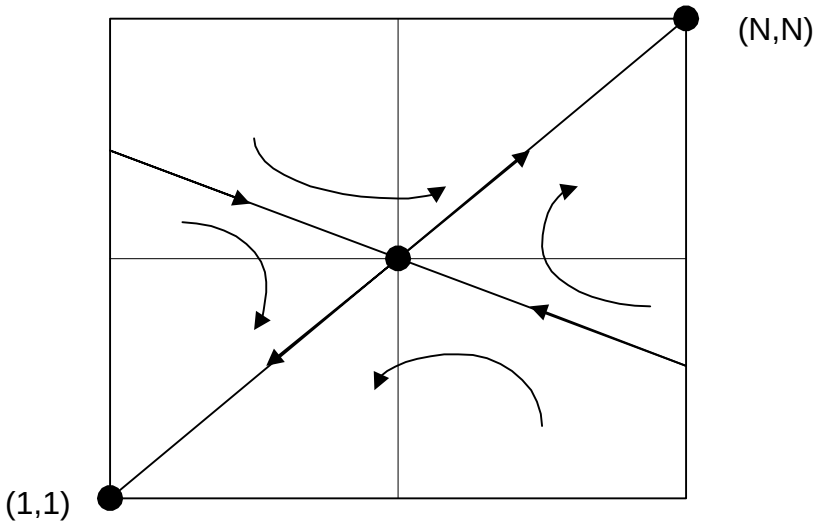
As throughout this chapter, the long run behavior of the perturbed system with very small mutation rates is determined by computing the number of mutations required to jump in and out of two basins. Moreover, for monotone systems, the lowest-cost transition is the immediate one:

Proposition 5.6: (Hahn [1995]) If the deterministic dynamics is monotone the lowest cost transition from one equilibrium to the basin of the other is to immediately jump to a state in the basin of the "target" equilibrium.

---

[113] Since there are only finitely many agents in each population, this assumption is satisfied for generic payoffs.

In other words, to determine the stochastically stable outcome we need only determine the basins of the two equilibria, and compute the minimum distance of each state corresponding to a pure-strategy equilibrium to the basin of the other. At this point the difference with the one-population model emerges: in the one-dimensional system d, the basins of the two equilibria under any Darwinian dynamics are the sets $\{\theta|\theta < \alpha^* N\}$ and $\{\theta|\theta > \alpha^* N\}$. In contrast, even strengthening the "Darwinian" assumption to the monotone assumption does not pin down the location of the basins in the two-dimensional case corresponding to two populations of players.[114] This can easily be seen by reference to the figure below



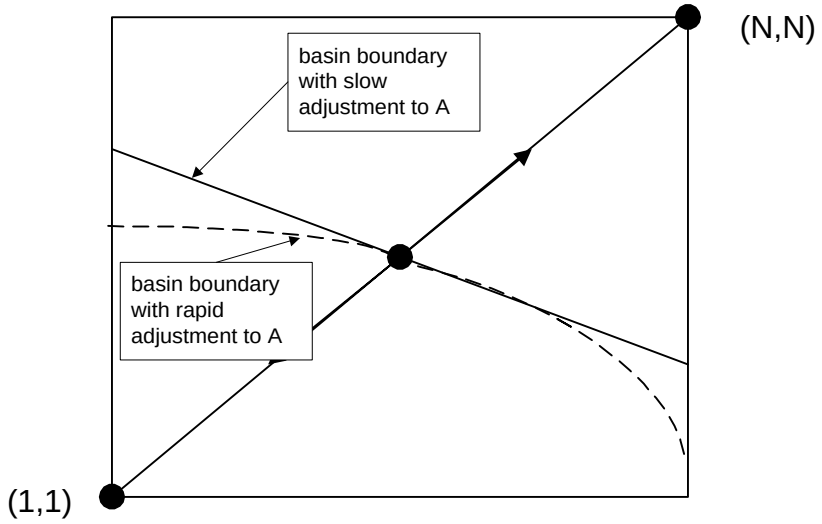
The Darwinian assumption implies that the lower left-hand region lies in the basin of (0,0), where all players play B and that the upper right-hand region lies in the basin of (N,N) where all players play A, but it does not pin down the eventual destinations of paths that start in the other two regions. In the upper left-hand box, for example, all

[114]In a similar paper, Romaldo [1995] notes that the Darwinian assumption does not determine the basins of attraction in one-population models with 3 or more actions per player.

trajectories have θ^1 increasing and θ^2 decreasing, until the first time that either $\theta_t^1 < \alpha^* N$ or $\theta_t^2 > \alpha * N$, but which if these occurs first depends on the relative speeds of adjustment of the two components of the state variable in this region. Moreover, the details of this specification can matter for the long-run outcome, as the shortest path from (0,0) to the basin of (N,N) need not be along the diagonal connecting the two equilibria As an example of the possibilities this generates, suppose that both populations adjust faster towards strategy A when A is optimal than they do towards B when B is optimal:

$$\begin{cases} \theta_{t+1}^i - \theta_t^i = \beta^A \text{ if } \theta_t^j > \alpha^{*j} N \\ \theta_{t+1}^i - \theta_t^i = \beta^B \text{ if } \theta_t^j < \alpha^{*j} N \end{cases},$$

with $\beta^A > \beta^B$. Then if the game is symmetric (or nearly so) the stochastically stable outcome is the state (N,N) where all agents play A. This change in the basin when the adjustment rates change is illustration below:



Hahn's focus is on asymmetric game like battle of the sexes, in which player 1 prefers the equilibrium (A,A) and player 2 prefers (B,B). He gives an upper bound on the relative

speeds of adjustment that is sufficient for the equilibrium selected to be invariant to other details of the dynamic.

5.9. Stochastic Replicator Dynamics

As we remarked in the introduction to this chapter, the evolutionary games literature traditionally considers only deterministic systems, with "mutations" being an unmodeled explanation for restricting attention to stable steady states. If one believes that "mutations" are real and recurring, phenomenon, it might seem more appropriate to include them explicitly in the model. From this viewpoint, an obvious way to study the effects of stochastic shocks is to introduce a stochastic term into the standard replicator equations, and indeed the first paper to study a stochastic adjustment model, Foster and Young [1990], did exactly that.

As noted by Fudenberg and Harris [1992], there is an important difference between stochastic replicator dynamics and the finite-population models discussed earlier in the chapter: If the noise has continuous sample paths, the evolutionary system will too. Consequently, the "cost" of transitions between the basins of various equilibria, and thus the nature of the long-run distribution, can depend on the "strength" or size of the "flows" (that is, the deterministic part of the dynamics) and not just on the directions of movement. This contrasts with the finite-population models discussed so far, where the transitions can occur by "jumps", and the costs of transitions depended on the shapes of the various basins but does not depend on the exact specification of the deterministic process within the basins. In the analogy used by Fudenberg and Harris, the various ω-limit sets can be viewed as "ponds," with the deterministic process corresponding to various streams of water. In the models driven by many simultaneous individual

mutations, the state moves from one pond to another by "jumping upstream" *over* the flow, so the strength or speed of the stream does not matter; in models with continuous sample paths, the state must "swim" against the flow, and the identity of the stochastically stable set is determined by comparing the relative size of expressions that involve integrals of both the noise and deterministic forces.

Foster and Young [1990] start with the single-population replicator model, and add a Wiener process with no cross-covariance and a general state-dependent variance to arrive at the system of stochastic differential equations of the general form

$$d\theta_t(s) = \theta_t(s)[u_t(s) - \bar{u}_t\ ] + \sigma(s|\theta)dW_t\ (s)$$

where each $W(k)$ is a standard Wiener process, and, as in chapter 3, the arguments of the variables denote the actions.

If the variance functions $\sigma(s|\theta)$ do not shrink quickly enough as the boundary is approached then solutions to this system have positive probability of hitting the boundary of the state space in finite time, so that the boundary behavior of the system matters. Foster and Young specify that the system has instantaneous reflection at the boundary.[115] They then give a general description of how to compute the limit of the long-run distributions as the variances of the Wiener processes shrink to 0, and use it to argue that in the case where the $\sigma(s|\theta)$ are constant, the distributions in 2x2 coordination games converge to a point mass on the Pareto-efficient and risk-dominant equilibrium.[116]

---

[115] Roughly speaking, instantaneous reflection in a continuous time stochastic system means that when the system hits the boundary, it continues to move with the same speed, but discontinuously reverses the direction with which it hit the boundary.

[116] In symmetric 2x2 coordination games, the risk dominant equilibrium must also be Pareto dominant. Unfortunately Foster and Young's proof of the general result cites a chapter of Freidlin and Wentzell which does not apply to their problem. In a private communication, Foster and Young have indicated that they are preparing a new proof that relies on another source.

Instead of adding stochastic terms directly to the replicator equations for population shares, Fudenberg and Harris [1992] take a different methodological approach, and add the stochastic terms to the equations governing the absolute population sizes, and then derive the corresponding equations for the evolution of population shares. That is, they start with the standard deterministic equations

$$\dot{\phi}_t(s) = \phi_t(s)u(s,\theta_t)$$

for the evolution of population size[117], and then suppose that the payoff to strategy s at date t is given by $u(s,\theta_t) + \sigma dW_t(s)$, where the W are independent standard Wiener processes; for notational simplicity we set the variance coefficient to be the same for each strategy.

The resulting stochastic differential equation then becomes

$$d\phi_t(s) = \phi_t(s)u_t(s,\theta)dt + \phi_t(s)\sigma dW_t(s).$$

The formulation using payoff shocks has the advantage of being consistent with a nonnegligible amount of noise in models with continuum of agents, while i.i.d shocks to individual agents might be expected to become almost deterministic in the aggregate as the population becomes large, just as the transition and convergence times in Kandori, Mailath and Rob [1993] grow exponentially with the population size.[118] As we see below,

[117] As we discussed in chapter 3, this formulation allows the possibility that payoffs and hence population growth rates may be negative, but the growth rate in the replicator dynamics can be thought of as the net difference between births and deaths.

[118] While we emphasize that some form of correlation seems necessary to explain nonnegligible noise in a model with a continuum of players, Binmore, Samuelson and Vaughn [1993] note that a stochastic differential equation can be used to approximate the limit of the long-run distribution of a discrete-time, finite-population model in which only one agent moves at a time, so that the limiting sample paths are continuous. The stochastic differential equation arises when the limit is taken in the following order: first time goes to infinity, then period length goes to 0, then population size grows to infinity, and finally a mutation rate goes to 0. The resulting stochastic differential equation is then not used to model a situation with nonnegligible noise, but only to compute the long-run limit of the system as the noise become negligible.

this formulation, in which variations in play are caused by variations in payoffs, can have very different implications than the "mutations" of KMR; it is not clear to us that either source of noise should be expected to always overwhelm the other one.

The stochastic system for the evolution of population shares can be derived by applying Ito's lemma applied to the function

$$\theta_i(s) = \frac{\phi(s)}{\sum_{s'} \phi(s')}.$$

This yields, in the 2x2 case, the equations

$$d\theta_t(s) = \theta_t(s)\theta_t(s')\left[(u_t(s) - u_t(s'))dt + (\sigma^2(\theta_t(s') - \theta_t(s))dt + \sqrt{2}\sigma d\tilde{W}_t\right],$$

where $\tilde{W} = (W(s) - W(s'))/\sqrt{2}\sigma$ is another standard Wiener process.

Observe that the deterministic part of the system (the coefficient of *dt*) is not the same as in the deterministic replicator dynamics, but includes an additional term corresponding to the weighted difference of the variances. In addition, when the shocks to the underlying payoff process have a constant variance, the shocks to the population shares have variance that shrinks as the boundary is approached, and it is easily seen that the boundary is never reached in finite time, so that the boundary behavior is irrelevant. This should be intuitive: regardless of the realization of the payoff shocks, and of the resulting absolute sizes of the population using each strategy, the share of each strategy is by definition nonnegative.

Fudenberg and Harris solve for the long-run behavior of this system in 2x2 games.[119] If the game has 2 strict equilibria, the system is not ergodic. Rather, the system converges with probability 1 to one of the two pure equilibria, but the relative

[119]In contrast to most of the papers discussed in this chapter, the results in Fudenberg and Harris are not based on the perturbation methods of Freidlin and Wentzell, but rather on an analysis of stochastic differential equations by Gihman and Skohorod [1972]. Unfortunately that analysis becomes very difficult in higher-dimensional systems, so it may not prove as useful in further work.

probabilities depend on the initial condition. Intuitively, because the replicator dynamics says that the absolute growth rate of a small population must be small, the assumed shocks to payoffs do not do very much to perturb the population shares in the neighborhood of a point where almost everyone is using the same action.

Fudenberg and Harris go on to consider a further modification of the replicator dynamics intended to capture the effects of a deterministic flow of mutations (or more generally an inflow of new players), as in Boylan [1994]. This flow serves to keep the system from approaching the boundaries, and thus makes the system ergodic. Moreover, the ergodic distribution can be found by calculating an integral which depends on the strength of the flow and the variance of the system (see for example Skohorod [1989]). In 2x2 games with 2 strict equilibria, the limit of the ergodic distribution as the variances of the payoffs and the flow of "mutations" both go to 0 is a point mass on the risk-dominant strategy. While this seems to demonstrate the robustness of the Kandori, Mailath, and Rob result, the degree of confirmation implied may be less than it appears since the equilibrium selected depends on the strength of the flows of the unperturbed adjustment process at each state and not only on the direction of adjustment. More precisely, there are many "Darwinian" processes with the same basin of attraction as replicator dynamic that select the risk-dominated equilibrium; an easy but artificial example is a process with very rapid adjustment in the basin of the dominated equilibrium and very slow adjustment in the basin of the dominant one.

Cabrales [1993] extends Fudenberg and Harris' analysis to general *n*-player games, and shows that even the stochastic replicator dynamics need not select the equilibria with the largest basin of attraction in symmetric one-population models of games with more than two players. The proof follows from computing the integral alluded to above. The technical reason that the answer here is different is that the payoff

to a given strategy is a linear function of the population fractions in a two-player game, but with more than two players its is a higher-order polynomial. To obtain a more satisfactory explanation, we must relate this observation to the foundations of the models. The driving force in the models we discussed in previous sections is the probability that enough players simultaneously "mutate" that the remaining players wish to switch as well. Since the fraction required depends only on the size of the basin, the size of the basin determines the stochastically stable outcome. In contrast, the driving force in models with shocks to payoffs is the probability of a sufficiently large change in payoffs that players choose to change their action. In a symmetric two-player game, these two criteria are identical because payoffs are linear in the population fraction playing each action but in *n*-player games payoffs are polynomial in the population fraction playing different actions.

To get an intuition about why the polynomial dependence of utility on population fractions makes a difference, consider the "Stag-Hunt" game, where the two strategies are "Hare," which pays 1 regardless of opponents' play, and "Stag," which pays $a > 1$ if *all* opponents play Stag, but pays 0 otherwise.[120] If there are only two players, the Pareto-dominant equilibrium "All Stag" is risk-dominant if and only if $a > 2$; but for any $a > 2$ "All Hare" is risk dominant if the number of players *n* is large enough that $a < 2^{n-1}$. Now consider the simple case where only the payoff to "all Stag" is stochastic, and where the fractions playing Stag and Hare are bounded below by $\varepsilon_s, \varepsilon_H > 0$ due to the inflow of new players. At the state "all Stag", the payoff to "Stag" is $a(1-\varepsilon_H)$ and the payoff to Hare is 1, so the payoffs would need to change by $a(1-\varepsilon_H)-1$ to make Hare the optimal choice. At the state "all Hare," the payoff to Hare is 1, and the payoff to Stag is $a\varepsilon_S$, so

[120] In Rousseau's story, all players must work together in order to catch the stag. This game is very similar to the example of a team problem that Cabrales used in his paper

payoffs would need to change by $1-a\varepsilon_S$ to make Stag optimal. Thus, regardless of the number of players, the change in payoffs required for a shift from Stag to Hare is larger than that for the reverse shift iff $a(1-\varepsilon_H)-1-(1-a\varepsilon_S)>0$, e.g. if $a-2>a(\varepsilon_H-\varepsilon_s)\approx 0$. This highlights the differing effects of the sources of noise in the two formulations.

APPENDIX 1: REVIEW OF FINITE MARKOV CHAINS

We consider discrete-time, finite-state Markov processes with Markov transition matrix *P*. Then if $\theta, \xi \in \Theta$, the element $P_{\theta\xi}$ of this matrix is the probability that the state is θ at date *t*+1 conditional on the state being ξ at date *t*. With this convention, probability distributions over states are represented by column vectors φ, and $\varphi_{t+1} = P\varphi_t$. Note that this system is "autonomous" or "stationary" meaning that *P* does not depend on the time *t*. This rules out processes such as fictitious play where the step size shrinks over time..

What can be said about the long-run behavior of the system? Under certain conditions developed below, this behavior is described by its "invariant distribution." We say that φ is an *invariant distribution* if $P\varphi = \varphi$.

Every finite Markov chain has at least one invariant distribution (P is a continuous operator on the compact convex set $\Delta(\Theta)$) but in general this distribution need not be unique. Consider, for example, the deterministic process corresponding to the Markov operator $P = I$ (the identity matrix). Here every probability distribution is an invariant distribution. Notice however, that only the point masses on a single state make sense as descriptions of the long-run behavior of this system; the other invariant distributions are rather descriptions of which beliefs would be "stable" (constant over time) for an outside observer whose initial beliefs are exogenous and who cannot observe the system itself.

This example shows that some conditions are needed in order to be able to interpret the invariant distributions as sensible descriptions of long-run behavior. A system is *ergodic* if it satisfies all of the following three conditions:

1) The invariant distribution $\hat{\varphi}$ is unique.

2) Convergence of time averages

$$\lim_{T\to\infty}(1/T)\sum_t 1(\theta_t = \theta) = \hat{\varphi}(\theta) \text{ almost surely}$$

where the indicator function $1(\cdot)$ is equal to one if the condition is true, zero otherwise.

3) Convergence of the date-*t* distributions:

$$\forall \varphi, \lim_{t\to\infty} P^t\varphi = \hat{\varphi}.$$

Some examples will help clarify the implications of these conditions. Looking first at the uniqueness of an invariant distribution, suppose that

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

This system cycles back and forth between the two states. It also has the unique invariant distribution $\hat{\varphi} = [1/2, 1/2]$. The sense in which this is a good description of long-run behavior is given by properties (2) and (3). Property (2), the convergence of the long-run average, means that the system spends half of its time in each state. This property is satisfied by the example. Property (3) would mean that even if the initial state is known perfectly, beliefs about the state at a sufficiently far distant time are 50-50. This property is not satisfied by the example: if the initial condition is known, then it is possible to predict exactly where the system will be at each future time, since it is deterministic.

Two examples show how a system that is not ergodic may be perturbed slightly so that it is. The first is a perturbation of the identity map,

$$\begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix}$$

which can be described as a persistent state: the system tends to remain in the current state, but has a slight chance of moving to the other state. The second is a perturbation of a deterministic two-cycle

$$\begin{bmatrix} \varepsilon & 1-\varepsilon \\ 1-\varepsilon & \varepsilon \end{bmatrix}$$

which can be described as a near cycle. This system cycles (switches to the other state) with high probability, but with a small probability it instead remains in the current state. Although it may not be immediately obvious, both of these systems are ergodic.

Taking their ergodicity as given for the moment, notice that these examples show that even if the system is ergodic, the beliefs of a player who observes the system as it evolves need not correspond to the ergodic distribution: In particular, while condition (3) says that knowing yesterday's state does not help with forecasting the system's long-run behavior, knowing yesterday's state can help forecast today's, as it does in the two examples here. Put differently, convergence to the invariant distribution does not imply the convergence of the time-*t* distributions over outcomes conditional on history until that point, but only the convergence of the unconditional distributions.

The easiest sufficient condition for ergodicity is the "strict positivity" condition that $P > 0$. The two examples above satisfy this condition, as do most of the models in this chapter. A weaker sufficient condition is $P^n > 0$ for some *n*. Another condition weaker than strict positivity is that there is a state that can be reached from any other state, and such that when this state is reached, there is a positive probability of remaining there the next period. That is,

(1) there exists a state θ such that for any θ' there exists a time *n* such that $\left(P^n\right)_{\theta\theta'} > 0$.
(2) $P_{\theta\theta} > 0$.

This condition may be understood in terms of the notion of a *recurrent class*, which is a stochastic analog of the invariant sets of the deterministic theory: A subset of states is recurrent if it has the property that once it is reached, the state must remain in the set with probability one. A recurrent class has the stronger property that it is a minimal recurrent set. Property (1) above implies that there is only one recurrent class, since

recurrent classes must be disjoint, and all contain the special state θ. Property (2) says that there is a positive probability of remaining in this state from one period to the next. This rules out deterministic cycles, and leads to the conclusion that the recurrent class is "aperiodic."

More generally, if there is a unique recurrent class, then there is a unique invariant distribution and this is sufficient for the time averages to converge. Consequently the first two conditions for ergodicity are satisfied. (In some treatments, ergodicity is defined by the first two conditions alone.) The addition of the condition that the recurrent class is aperiodic assures that the distant future is not terribly sensitive to current conditions, a condition known as mixing, and this leads to the final condition for ergodicity, the convergence of the long run distribution.

An example shows how the weak condition may be satisfied even when the transition matrix is not strictly positive. Consider two sequences of short-run players playing a 2-player game; each individual plays only once, but knows play in past K periods. Each player *i* "intends" to choose the strategy that maximizes his expected payoff against the distribution corresponding to the last K periods of play. The realized strategy is the intended one with probability 1-ε, and some fixed strategy $\hat{s}^i$ with probability ε. Here the state is the realized strategy profiles in the last K periods, so that not all transitions are possible in a single period: for example, the transition from a history of both players always playing "1" to a history of both players always playing "2" is impossible in a single period. Moreover, if (s^{1*}, s^{2*}) is a profile of strictly dominated strategies, the state has probability 0 of being to "K observations of (s^{1*}, s^{2*})" in any period after the first. However the state "*K* observations of $(\hat{s}^1, \hat{s}^2)$" satisfies the 2-part condition for ergodicity given above.

Appendix 2: Stochastic Stability Analysis

The basis of proposition 5.2 is the following result of Freidlin and Wentzell, which characterizes the invariant distribution of any finite-state irreducible Markov chain in terms of the trees defined in the text, where here the trees include all states θ and not just that are the ω-limit sets of some (as yet unspecified) deterministic system. Let $(\theta', \theta'') \in h_\theta$ mean that the tree h_θ has a transition from θ' to θ''.

Lemma 3.1: (Freidlin and Wentzell [1982]) If Q is an irreducible finite-dimensional matrix, the unique invariant distribution μ of Q is given by

$$\mu_\theta = \frac{Z_\theta}{\sum_{\theta'} Z_{\theta'}}, \text{ where}$$

$Z_\theta = \sum_{h_\theta \in H_\theta} \prod_{(\theta', \theta'') \in h_\theta} Q_{\theta'' \theta'}$, and H_θ is the set of all tress with root θ.

The proof of this lemma is not very revealing, as it consists of simply verifying that the distribution constructed is indeed invariant. Intuitively, the reason that the formula involves a sum over all the h_θ-trees is that each $\omega-$tree represents one way that the state might arrive at θ; each path is then weighted by its probability. Of course, the weight attached to a transition from θ' to θ depends on the probability of θ' as well as on the conditional probability of the transition, which is just another way of saying that invariant distribution is a fixed point. A brute-force computation of the distribution would involve inverting the matrix Q, and thus introducing a term corresponding to $\frac{1}{\det(Q)}$; this corresponds to the summation in the denominator of the expression for μ_θ.

With lemma 3.1 in hand, we now turn to our case of interest, where the perturbed matrices P^ε play the role of the irreducible matrix Q, so that the transition probabilities $Q_{\theta' \theta}$ are approximately $k_{\theta'' \theta'} \varepsilon^{c(\theta'' | \theta')}$, where the $k_{\theta'' \theta'}$ are independent of ε. Inspecting the

formula $\mu_\theta = \frac{Z_\theta}{\sum_{\theta'} Z_{\theta'}}$ shows that the support of the limit distribution will be concentrated on the states θ for which Z_θ is the lowest order in ε. Furthermore, the order of

$$Z_\theta = \sum_{h_\theta \in H_\theta} \prod_{(\theta',\theta'') \in h_\theta} Q_{\theta''\theta'} = \sum_{h_\theta \in H_\theta} \left[\prod_{(\theta',\theta'') \in h_\theta} k_{\theta''\theta'} \right] \varepsilon^{\sum_{(\theta',\theta'') \in h_\theta} c(\theta''|\theta')}$$

will be determined by the lowest-order elements in the summation, so

$$o(Z_\theta) = \arg\min_{h_\theta \in H_\theta} \varepsilon^{\sum_{(\theta',\theta'') \in h_\theta} c(\theta''|\theta')}$$

Thus we conclude that the limit distribution is concentrated on the states whose trees have the lowest cost.

This formula, while correct, requires one to consider all of the states of the process; part of the appeal of Proposition 5.2 is that it shows that it is sufficient to build trees whose elements are the ω-limit sets of the unperturbed process *P*. It is easy to see that if state θ' is in the basin $D(\theta)$ of state θ under *P*, then transitions from θ' to θ have cost 0, and can be ignored in computing the minimum. Therefore, if we construct a tree on the ω-limit sets whose cost is $\min_{\omega \in \Omega} \min_{h \in H_\omega} \Sigma_{\omega' \in \Omega/\omega} \vec{c}(D(h(\omega'))|\omega')$, we can construct a tree of the same cost over all the states adding an initial step in which every state θ is mapped to its ω-limit. The final step is to verify that no other tree on the whole state space can have lower cost than $\min_{\omega \in \Omega} \min_{h \in H_\omega} \Sigma_{\omega' \in \Omega/\omega} \vec{c}(D(h(\omega'))|\omega')$; this done by a straightforward but tedious "tree surgery" argument that we will omit.

References

Bergin, J. and B. Lippman [1995]: "Evolution with State Dependent Mutations," Queens University.

Binmore, K., L. Samuelson and K. Vaughn [1995]: "Musical Chairs: Modelling Noisy Evolution," *Games and Economic Behavior*, 11: 1-35.

Blume, L. [1993]: "The Statistical Mechanics of Strategic Interaction," *Games and Economic Behavior*, 5: 387-424.

Boylan, R. [1994]: "Evolutionqry Equilibria Resistant to Mutations," *Games and Economic Behavior*, 7: 10-34.

Cabrales, A. [1993]: "Stochastic Replicator Dynamics," UCSD.

Canning, D. [1992]: "Average Behavior in Learning Models," *Journal of Economic Theory*, 57: 442-472.

Ellison, G. [1993]: "Learning, Local Interaction, and Coordination," *Econometrica*, 61: 1047-1071.

Ellison, G. [1995]: "Basins of Attraction and Long-Run Equilibria," MIT.

Ely, J. [1995]: "Local Conventions," Northwestern University.

Foster, D. and P. Young [1990]: "Stochastic Evolutionary Game Dynamics," *Theoretical Population Biology*, 38: 219-232.

Friedlin, M. and A. Wentzell [1982]: *Random Perturbations of Dynamical Systems*, (New York: Springer Verlag).

Fudenberg, D. and C. Harris [1992]: "Evolutionary Dynamics with Aggregate Shocks," *Journal of Economic Theory*, 57: 420-441.

Fudenberg, D. and J. Tirole [1991]: *Game Theory*, (Cambridge: MIT Press).

Futia, C. [1982]: "Invariant Distributions and the Limiting Behavior of Markovian Economic Models," *Econometrica*, 50: 377-408.

Hahn, S. [1995]: "The Long Run Equilibrium in an asymmetric Coordination Game," Harvard.

Kandori, M. and R. Rob [1993]: "Bandwagon Effects and Long Run Technology Choice," U. Tokyo DP 93-F-2.

Kandori, M. and R. Rob [1995]: "Evolution of Equilibria in the Long Run: A General Theory and Applications," *Journal of Economic Theory*, 65: 383-414.

Kandori, M., G. Mailath and R. Rob [1993]: "Learning, Mutation and Long Run Equilibria in Games," *Econometrica*, 61: 27-56.

Kim, Y. [1993]: "Equilibrium Selection in n-person Coordination Games," *Games and Economic Behavior*.

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

Morris, S., R. Rob and H. Shin [1993]: "p-dominance and Belief Potential," *Econometrica*, 63: 145-158.

Romaldo, D. [1995]: "Simularities and Evolution," mimeo.

Skohorod, A. V. [1989]: *Asymptotic Methods in the Theory of Stochastic Differential Equations* (translated by H.H. McFadden), (: American Mathematical Society).

Young, P. [1993]: "The Evolution of Conventions," *Econometrica*, 61: 57-83.

6. Extensive form games and self-confirming equilibrium

6.1. Introduction

So far, we have limited attention to simultaneous-move games, where a player's strategy is simply a choice of a single uncontingent action. In such games, it is natural to assume, as we have done, that at the end of each play of the game each player observes the strategies used by each of his opponents. We now wish to consider learning in non-trivial extensive form games. The most natural assumption in many such contexts is that agents observe the terminal nodes (that is, outcomes) that are reached in their own plays of the game, but that agents do not observe the parts of their opponents' strategies that specify how the opponents would have played at information sets that were not reached in that play of the game.[121] The only setting we can imagine in which players observe more information than the realized terminal node is if the players are forced to write down and commit themselves to contingent plans, and even in that case the most natural interpretation is that the game has been changed to one in which the "actions" are commitments to strategies of the original game. On the other hand, in many settings, players will not even observe the realized terminal node, as several different terminal nodes may be consistent with their observation. For example in a first-price sealed-bid auction, players might observe the winning bid but not the losing ones. We say more about this possibility below. In large population settings, we will also assume that agents observe no signals at all about the outcomes of matches they do not participate in. This is the case in most game theory experiments, but it is less compelling as a description of

[121] Recall that each terminal node is associated with a unique path through the tree, and so with a unique sequence of actions.

real-world games, as in many cases agents may receive information about the outcomes in other matches.

Given that players do not observe play at unreached information sets, it is possible for the observed outcome to converge while the players maintain incorrect beliefs about off-path play. As a result, the learning process can converge to outcomes that cannot be generated by any Nash equilibrium of the game. We first illustrate this in section 6.2 with an example. After setting up the basic notation of an extensive form game in section 6.3, we recapitulate the simple learning model of chapter 2 in the extensive form setting in section 6.4. Section 6.5 introduces a weakening of Nash equilibrium, *self-confirming equilibrium*, that allows differences in beliefs off the equilibrium path. The stability of this concept of equilibrium in the basic learning model is explored in section 6.6. In section 6.7 we further weaken the notion of self-confirming equilibrium to allow the possibility that when there is a large population of players who share a single role different players who play the same role may have different beliefs off the equilibrium path.

We also consider several possible ways of strengthening self-confirming equilibrium. In section 6.8 we consider the possibility that opposing players randomize (or "tremble). The resulting notion of consistent self-confirming equilibrium is relatively close to Nash equilibrium; we consider the exact connection in section 6.9. Finally, players may know (or be fairly confident of) one another's payoffs. and use this knowledge to deduce restrictions on the likely play of their opponents; for example, that their opponents will not play dominated strategies. This leads to the notion of rationalizable self-confirming equilibrium, a notion we explore in section 6.10.

6.2. An Example

The possibility of non-Nash outcomes persisting in the long run is illustrated in the following example from Fudenberg and Kreps [1988].

Example 6.1 [Fudenberg and Kreps]: In the three player game illustrated in Figure 6.1, player 3 moves last, and cannot tell whether he has the move because player 1 played D_1, or because player 1 played A_1 and player 2 played D_2.

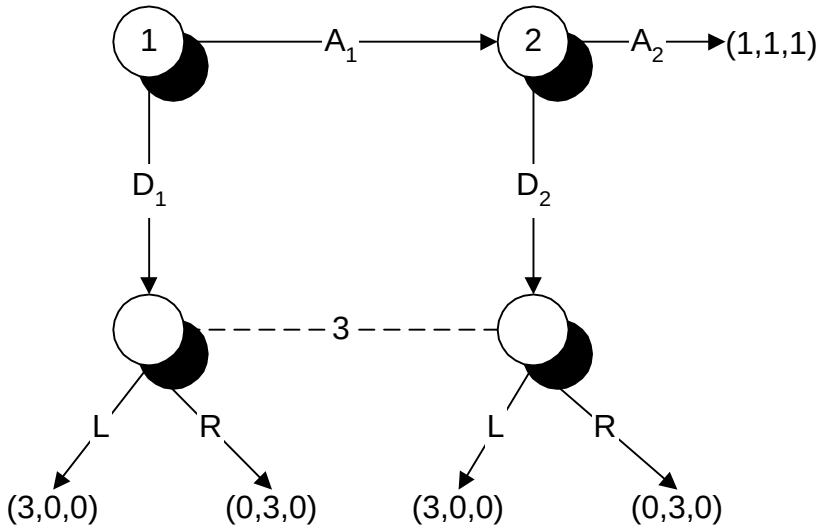


Figure 6.1

Suppose that player 1 expects player 3 to play R with probability exceeding 2/3, and player 2 to play A_2 with high probability, while player 2 expects player 3 to play L with at least this same 2/3 probability. and expects player 1 to play A_1.[122] Given these beliefs,

[122] For a more precise specification, suppose that players 1 and 2 form and update their beliefs as follows. Player 1's prior beliefs over the mixed strategies of players 2 and 3 are given by $\text{Prob}[\pi_2(A_2) \le p, \pi_3(R) \le q] = p^{100}q^{100}$; player 2's beliefs are given by $\text{Prob}[\pi_1(A_1) \le p, \pi_3(L) \le q] = p^{100}q^{100}$. Given these beliefs, player 1 assigns marginal probability $\int_0^1 100q^{100}dq = \frac{100}{101}$ that 3 chooses R, and assigns this same probability to player 2 choosing A_2; player 2

it is myopically optimal for players 1 and 2 to play A_1 and A_2, and the first-period outcome will be (A_1,A_2). Moreover, provided that player 1's initial beliefs about the play of players 2 and 3 are independent (that is, are a product distribution), the observed outcome gives player 1 no reason to change his beliefs about the play of player 3–player 1 becomes all the more convinced that 2 plays A_2. Likewise, if player 2's initial beliefs are a product distribution, player 2's beliefs about player 3 remain unchanged as well. Consequently, the outcome in the second and subsequent periods will also be (A_1,A_2), so that this outcome is a steady state. However, it is not a Nash equilibrium outcome: Nash equilibrium requires players 1 and 2 to make the same (correct) forecast of player 3's play, and if both make the same forecast, at least one of the players must choose D.

This example shows that learning can lead to non-Nash steady states unless there is some mechanism that leads players to have correct beliefs about off-path play. In this chapter we will focus on settings without such mechanisms, so that Nash outcomes cannot be expected. In the next chapter we examine one important reason why such off-path learning might occur: players might not be myopic, and consequently might choose to "experiment" with off-path actions that have a lower current period payoff in order to gain information that can be used in future plays. It is also possible that, players learn about off-path play from on-path observations if they believe on-path and off-path play are sufficiently correlated. While we do not wish to rule out such correlation, we do not wish to assume it, as we are not convinced that it is more reasonable than the opposite

---

assigns probability 100/101 to the event that 1 plays A_1 and assigns this same probability to 3 playing L. Consequently the myopic best responses for players 1 and 2 are A_1 and A_2, and this is the first-period outcome. Moreover, given the product structure of the beliefs, neither player 1 nor player 2 is led to change their beliefs about player 3, so the outcome is a steady state. Section 6.3 gives a more general discussion of the extension of fictitious play to extensive form games.

polar case of independent beliefs. We will discuss this point further when we revisit example 6.1

6.3. Extensive Form Games

We will examine extensive form games with *I* players; the game tree *X*, with nodes $x \in X$, is finite. Terminal nodes are $z \in Z$. For notational convenience, we represent nature by player 0, and suppose that Nature moves only at the initial node of the tree. Information sets, denoted by $h^i \in H$ are a partition of $X \setminus Z$. The information sets where player *i* has the move are denoted by $H^i \subset H$. The feasible actions at information set $h^i \in H$ are denoted $A(h^i)$. We continue to use $-i$ for all players except player i, so that for example H^{-i} are information sets for all players other than *i*. A pure strategy for player *i*, s^i, is a map from information sets in H^i to actions satisfying $s^i(h^i) \in A(h^i)$; S^i is the set of all such strategies.; mixed strategies are $\sigma^i \in \Sigma^i$. Each player except nature receives a payoff that depends on the terminal node, denoted $r_i(z)$.

In addition to mixed strategies, we define behavior strategies $\pi^i \in \Pi^i$. These are probability distribution over actions at each information set for player *i*. For any given mixed strategy σ^i for player *i*, and any information set for that player, we can define a behavior strategy by Kuhn's theorem–we denote this as $\hat{\pi}(h^i \mid \sigma^i)$. For any given behavior strategy π it is also useful to define the induced distribution over terminal nodes $\hat{\rho}(\pi)$. We will also use the shorthand notation $\hat{\rho}(\sigma) \equiv \hat{\rho}(\hat{\pi}(\sigma))$.

We assume that all players know the structure of the extensive form and their own payoff function, so that the only uncertainty each player faces concerns the strategies opponents will use and that of nature.[123] To avoid complications, we suppose

[123] As usual, one way to model cases where players are uncertain of the structure of the extensive form is to include a move by Nature in which the extensive form is chosen. This permits a player who is consistently

that the distribution of Nature's moves is known; any unknown but exogenous distributions can be represented as arising from the choice of a "dummy" player. To model the "strategic uncertainty" about players' strategies, we let μ^i be a probability measure over Π^{-i}, the set of other players' and Nature's behavior strategies. As discussed in chapter 2, assuming that the support is Π^{-i} and not $\Delta(\Pi^{-i})$ implies that players are certain that opponents do not correlate their play and will maintain that belief regardless of any evidence to the contrary. That is, any correlating devices that may be available to any subset of the players are explicitly included in the description of the extensive form. We are somewhat concerned by this restriction, but we impose it anyway to limit the number of complications that need to be addressed.

Again following chapter 2, the beliefs, which are distributions over strategies, must be integrated to obtain the player's predictions about expected play. For example, the probability that *i* assigns to terminal node *z* being reached when he plays π^i is $\gamma^i(z|\pi^i,\mu^i) = \int_{\Pi^{-i}} \hat{\rho}(z|\pi^i,\pi^{-i})\mu^i[d\pi^{-i}]$. This allows us to compute the expected utility $u^i(\pi^i,\mu^i) = \sum_Z r^i(z)\mu^i(z|\pi^i,\mu^i)$.

For any mixed profile σ, we let $\overline{H}(\sigma) \subset H$ be the information sets that are reached with positive probability when σ is played. Note that this set is entirely determined by the distribution over terminal nodes ρ, so we may equally well write $\overline{H}(\rho) = \overline{H}(\hat{\rho}(\sigma))$. We also denote by $H(s_i)$ the set of information sets that can be reached when player *i* plays s_i, that is the set $\{h^i | \exists s^{-i} s.t. h^i \in H(s^i, s^{-i})\}$. For any subset $J \subset H$ and any profile σ we may define the subset of behavior strategies consistent with players other than *i* playing σ_{-i} at the information sets in J by $\Pi^{-i}(\sigma^{-i} \Big| J) \equiv \{\pi^{-i} \Big| \pi^j(h^j) = \hat{\pi}(h^j \Big| \sigma^j), \forall h^j \in H^{-i} \cap J\}$.

---

outguessed when he *thinks* he is playing the simultaneous-move game "matching pennies" to eventually infer that his opponent is somehow observing and responding to the player's choice.

6.4. A Simple Learning Model

We now consider an extensive-form analog of the generalized version of fictitious play discussed in Chapter 2. To keep things simple, at this point we will suppose that there is only 1 agent in each player role, and that all agents are completely myopic; both of these restrictions will be relaxed in Chapter 7.

Each play of the game results in a particular terminal node *z* being reached. We assume that all players observe this terminal node, so at the start of round *t* all players know the sequence $(z_1, z_2, \ldots z_{t-1})$; this is called the *history at t* and is denoted h_t; h_∞ denotes an infinite history of play, and when a particular infinite sequence h_∞ has been fixed, h_t will mean the first *t* observations in that sequence. Note that h^i are information sets, while h_t are histories. A *belief rule* for player i is a function from histories to beliefs μ^i; in a slight abuse of notation we will denote this function by μ^i as well, so that $\mu_t^i(h_t)$ denotes player i's beliefs at date *t* given history h_t.[124]

Our next step is to specify how players update their beliefs and choose their actions in the course of the dynamic learning process.

6.4.1. Beliefs

To model beliefs, we will extend the strategic-form definition of asymptotic empiricism (given in Chapter 2) to the current setting of extensive-form games. (Recall that that definition said that players' beliefs corresponded to the empirical distribution.) Following Fudenberg and Kreps [1995a], we suppose that player i's estimates of play at a given information set for player *j* converges to the empirical distribution of play at that

[124] To model situation where players need not observe the terminal node at the end of each round, we could suppose that each player i observed some element $\lambda^i(z)$ of a partition of the *z's*, where each player's own payoff function is measurable with respect to his partition. This sort of more general learning model is implicit in the equilibrium concept proposed by Battigalli [1987].

information set as the number of observations of play at that information set converges to infinity.

To make this more precise, let $\widehat{H}(h)$ denote the information sets that are reached a positive fraction of the time along history h_∞, and let $d(h^j|h_t)$ be the empirical distribution of play at information set h^j.

Definition 6.1: Player *i's* belief rule is *asymptotically empirical in the extensive form* if for every $\varepsilon > 0$, infinite history h_∞, $j \neq i$, and information set $h^j \in \widehat{H}(h_\infty) \bigcap H^j$,

$$\lim_{t\to\infty} \mu_t^i(h_t)\left(\left\{\pi^{-i} \,\middle|\, \left\|\pi^j(h^j) - d(h^j|h_t)\right\| < \varepsilon\right\}\right) = 1.$$

In one-shot simultaneous move games, this definition reduces to that of asymptotic empiricism in the strategic form *provided* that the assessments are assumed to be the product of independent marginals. To see this, suppose that player 1 and player 2 each have a single information set, and those information sets are reached a positive fraction of the time. Then the probability that any third player assigns to the event (1 and 2 both play *L)* must converge to the product of the corresponding empirical marginal distributions, even if the empirical *joint* distribution is correlated.

6.4.2. Behavior Given Beliefs

For simplicity, we will assume here that players are completely myopic, and in each period choose a strategy that is a best response to their current beliefs. More precisely, we suppose that the strategy chosen by player *i* at date *t* is a maximizer of $u^i(\pi^i, \mu_t^i) = \sum_Z r^i(z)\mu_t^i(z|\pi^i, \mu^i)$. We should emphasize that this is an *ex-ante* notion of maximization, as is the definition of asymptotic myopia in strategic-form games given in chapter 4: with this notion of maximization, a maximizing strategy may prescribe an action that is suboptimal at an information set that has probability 0 given π^i, μ_t^i.

Note also that this assumption here is more restrictive than it was in the case of strategic-form games, for here myopia is *not* an implication of large population models with random matching. Such models do imply that players should not sacrifice utility in the current match to influence play in future matches, but in the present setting there is an additional reason that players might choose to sacrifice current utility, namely to gain information that may be useful in future play. That is, players might choose to "experiment" to learn more about their opponents' strategies.

To see this, consider the game in figure 6.2

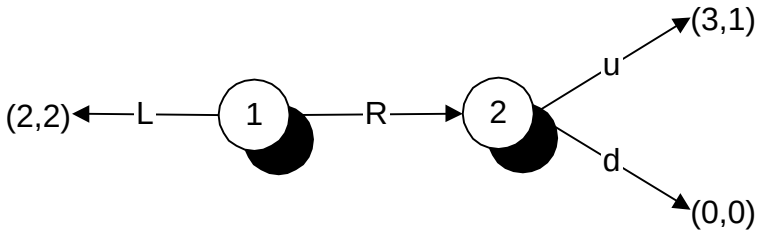


Figure 6.2

Suppose that player 1's beliefs μ^1 are that with probability ½ player 2 plays u in every period, and with probability ½ player 2 plays d in every period. Then player 1's current assessment of 2's play corresponds to the mixed strategy (½ u, ½ d), and the expected payoff from R is 1.5, which is less than the payoff to L. Hence a myopic player 1 would play L in the first period; since this results in no new information about player 2' s play, player 1 would then play L in all subsequent periods as well. However, if player 1 plays R a single time, he will learn exactly how player 2 is playing, so that the decision rule "play R in the first period, and play R thereafter if and only if 2 plays u" has probability ½ of yielding 3 in every period, and probability ½ of yielding 0 in period 1, followed by 2 at all future dates, which will have a higher present value provided that player 1's discount factor is not too large.

This shows that we should not expect players to behave myopically in the initial periods of the game, as they may choose to "experiment" with strategies that do not maximize their short-run expected payoff. However, the results we will present extend to situations where players satisfy the much weaker condition of "asymptotic myopia," meaning that they *eventually* stop experimenting and play to maximize their current period's payoff. Fudenberg and Kreps [1995a,b] formulate several variants of this condition, and use it in place of the exact myopia we assume here.[125] From the literature on the "bandit problem," we would expect that players with any discount factor less than 1 would eventually stop experimenting, and be "asymptotically myopic" in this sense. Chapter 7 discusses the results of Fudenberg and Levine [1993b] who show that this is true for Bayesian present-value maximizers in a closely related model of learning in extensive form games

To summarize this section, then, we will be interested in fictitious-play-like processes

in which the strategies chosen in each period are a best response to expected play in that period, and where beliefs about opponents' actions are asymptotically empirical in the extensive form.

6.4.3. Equilibrium Notions

A number of notions of "conjectural," "subjective" or "self-confirming" equilibrium have been introduced to capture the relationship between steady states of a learning process in an extensive form game and equilibria of the static game. The notion

[125] In fact, a major concern of that paper is formulating and investigating definitions of "asymptotic myopia" that seem general enough to be plausible while still strong enough that the experimentation does not "show up" in the long run outcome. Another issue that that paper addresses, and we will skip over, is the extent to which very weak forms of asymptotic myopia may conflict in spirit with some strengthenings of asymptotic empiricism that might otherwise seem natural, for example requiring beliefs to correspond to empirical distribution at all information sets that are reached infinitely often but with vanishing frequency.

originates in Hahn [1977] in the context of production economies, and is discussed in Battigalli [1987] in the context of a game.[126] Self-confirming equilibrium in which only the terminal nodes of the game are observed is developed in Fudenberg and Levine [1993a] and in Fudenberg and Kreps [1995a]. The notion of subjective equilibrium is developed in the context of repeated games in Kalai and Lehrer [1993]. There are also some related concepts that include elements of rationalizability; we discuss these later in this chapter. All of these definitions are intended to capture and generalize some version of the point raised by example 6.1, namely that each time the game is played players get only incomplete information about the strategies used by their opponents, so that incorrect beliefs about off-path play can persist.Our presentation follows Fudenberg and Kreps [1995a] with the goal of providing analogs of the results in chapter 2 on the relationship between steady states of this learning process and equilibria of the underlying game.

Nash equilibrium is usually defined as a strategy profile such that each player's strategy is a best response to his or her opponents. For our purposes, though, it is instructive to give an equivalent definition that parallels the way in which we will define self-confirming equilibrium.

Definition 6.1: A *Nash equilibrium* is a mixed profile σ such that there exist beliefs μ^i and for each $s^i \in \text{supp}(\sigma^i)$

- $u^i(s^i|\mu^i) \geq u^i(s'^i|\mu^i)$ for all $s'^i \in S^i$, and
- $\mu^i(\Pi^{-i}(\sigma^{-i})|H)) = 1$.

[126] Battigalli's concept allows more general "signals" that correspond to the partitions discussed in footnote 5 above.

In this definition, the first condition requires that each player's strategy be optimal given his beliefs about the opponents' strategies. The second requires that each player's beliefs are correct at every information set.

If, as we suppose, players observe only the terminal nodes that are reached, and not how opponents would have played at unreached information sets, then even if player *i* continually plays σ^i, he will only observe opponents play at information sets in $\overline{H}(\sigma)$, and will not learn about his opponents' play at other information sets. This leads us to the following equilibrium concept:

Definition 6.2: [Fudenberg and Levine 1993a] A unitary self-confirming equilibrium is a mixed profile σ such that there exist beliefs μ^i and for each $s^i \in \text{supp}(\sigma^i)$

- $u^i(s^i \mid \mu^i) \geq u^i(s'^i \mid \mu^i)$ for all $s'^i \in S^i$, and
- $\mu^i(\Pi^{-i}(\sigma^{-i} \mid \overline{H}(\sigma))) = 1$.

6.5. Stability of Self-Confirming Equilibrium

We turn now to stability analysis in the simple learning model introduced above. As in Chapter 4, say that a profile is *unstable* if for every positive ε, players' behavior is almost surely more than ε away from the profile infinitely often.

Proposition 6.1: [Fudenberg and Kreps 1995a]: If σ is not a self-confirming equilibrium, then it is unstable with respect to any behavior rules that are myopic with respect to asymptotically empirical assessments.

The intuition for this is simple: If play converges to σ, then from the strong law of large numbers we expect that every information set in the support of σ should be reached a nonvanishing fraction of the time, and that the distribution of actions at such information sets should converge to that generated by σ. Asymptotic empiricism then

implies that players' assessments converge to σ along the path of play, and a standard continuity argument shows that some player eventually perceives a benefit to deviating from σ.

This result only shows that *strategy profiles* cannot converge to a limit that is not a self-confirming equilibrium; it does not preclude a situation in which the strategy profiles fail to converge while the *outcome* converges to a limit that cannot be generated by any self-confirming equilibrium. Since only outcomes are observed, it is of some interest to know that the proposition can be extended: Say that an outcome ρ is unstable if there is an $\varepsilon > 0$ such that there is probability 0 that the distribution of outcomes generated by the players' strategies is always within ε of ρ

Proposition 6.2: (Fudenberg and Kreps 1995a): If outcome ρ is not generated by any self-confirming equilibrium, then it is unstable with respect to any behavior rules that are myopic with respect to asymptotically empirical assessments.

Note that this result compares the probability law generating outcomes to the specified outcome distribution ρ, as opposed to comparing the observed empirical distribution to ρ; but arguments in the spirit of the strong law of large numbers can be combined with proposition 6.2 to show that there is probability 0 that the empirical distribution of outcomes remains within ε of ρ.

The discussion of example 6.1 already gives an example of a stable profile that is not a Nash equilibrium. A more formal statement of this requires a definition of local stability that allows for randomness:

Definition 6.3: A strategy profile π is *locally stochastically stable* under a given behavior rule (and initial condition) if there is positive probability that the strategy profile chosen by the players converges to π.

Proposition 6.3: Every self-confirming strategy profile π is locally stable for some behavior rules that are myopic with respect to asymptotically empirical assessments.

The proof of this parallels the construction in chapter 4 in which players start out with a strong prior belief in the particular equilibrium, and maintain that belief unless they receive overwhelming evidence to the contrary.

6.6. Heterogeneous Self-Confirming Equilibrium

In Chapter 7 we will discuss learning in the extensive form in a model where players are randomly matched with one another and observe only the results of their own match, as in most game theory experiments. In this case, there is no reason that two subjects assigned the same player role should have the same prior beliefs.. Moreover, given that players only observe the outcomes in their own matches, if two players have always played different pure strategies, their beliefs may remain different.[127] Fudenberg and Levine [1993a] introduce the following weaker notion of self-confirming equilibrium to capture this notion.

Definition 6.4: A *heterogeneous self-confirming equilibrium* is a mixed profile σ such that for $s^i \in \text{supp}(\sigma^i)$ there exist beliefs μ^i such that

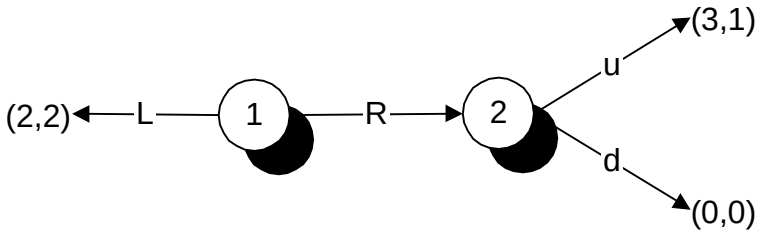
- $u^i(s^i \,|\, \mu^i) \geq u^i(s'^i \,|\, \mu^i)$ for all $s'^i \in S^i$, and
- $\mu^i(\Pi^{-i}(\sigma^{-i} \,|\, \overline{H}(s^i, \sigma^{-i}))) = 1$.

This definition allows different beliefs be used to rationalize each pure strategy in the support of σ^i, and allows the beliefs that rationalize a given strategy to be mistaken at

[127] On the other hand, we would expect all players to eventually have the same beliefs if they observe the aggregate distribution of outcomes in the whole population.

information sets that are not reached when the strategy is played, but are reached under reached under a different strategy that is also in the support of σ^i. A simple example from Fudenberg and Levine [1993a] shows how this allows outcomes that cannot arise with unitary beliefs:

Consider again the game in Figure 6.2.



The game has two types of Nash equilibria: the subgame perfect *Ru* and the equilibria in which player 1 plays *L* and player 2 plays *d* at least 50% of the time. However, there is no Nash equilibrium in which player 1 randomizes between *L* and *R*, nor is there a unitary self-confirming equilibrium of that form. This is a consequence of a more general theorem that we present in section 6.z, which gives conditions for the outcomes of unitary SCE to coincide with the set of Nash outcomes, but the argument can be made directly in this example: If a single player 1 randomizes between *L* and *R*, unitary SCE requires that he know how player 2 responds to *R*, and since player 2 is reached a positive fraction of the time, player 2's will always play *u*.

There is however a heterogeneous self confirming equilibrium in which player 1 does randomize: player 2 plays *U*, and while those player 1's that play *R* know this, those that play *L* incorrectly believe that player 2 would play *D*.

Note that in a one-shot simultaneous-move game, all information sets are on the path of every profile, so the sets $\overline{H}(s^i,\sigma^i)$ are all of *H*, and so even heterogeneous self-

confirming equilibrium requires that beliefs be exactly correct. Hence, in these games, all self-confirming equilibria are Nash.

6.7. Consistent Self-Confirming Equilibrium

So far we have considered various ways of weakening the notion of equilibrium to capture steady states of a learning process where the entire opponents strategy is not observed. We now wish to consider ways in which we can strengthen the notion of self-confirming equilibrium to reflect additional information that may be available to players. Our first considerations is what happens when a player faces an opponent whose hand trembles, or a sequence of different opponents a small fraction of which have different preferences. In this case the player in question will learn not only what will happens on the equilibrium path, but also what will happen at all information sets that are actually reachable given his own strategy.

Definition 6.5: A *consistent unitary self-confirming equilibrium* is a mixed profile σ such that for each $s^i \in \text{supp}(\sigma^i)$ there exist beliefs μ^i such that

- $u^i(s^i \mid \mu^i) \geq u^i(s'^i \mid \mu^i)$ for all $s'^i \in S^i$, and
- $\mu^i(\Pi_{-i}(\sigma^{-i} \mid H(\sigma^i)) = 1$

Consistent self-confirming equilibria have stronger properties and are more "Nash-like" than inconsistent self-confirming equilibrium. Although consistency may be a reasonable condition to impose in some circumstances, such as those mentioned above, the main uses of the condition so far have been consequences of the fact that in some classes of games, all self-confirming equilibria are necessarily consistent.

Obviously Nash equilibrium requires consistency, that is, that two players agree about the play of a third player. This is also required by any sort of correlated equilibrium. However, not all inconsistent beliefs lead to departures from Nash equilibrium. In particular, in order for inconsistent beliefs of players 1 and 2 about the play of player 3 to support an outcome that cannot occur in a Nash equilibrium, both player 1 and player 2 need to be able to *unilaterally* deviate from the path of play and cause the information set in question to be reached, which is only possible if player 3 is unable to distinguish between deviations by the two players. Fudenberg and Levine [1993a] define a class of games in which this cannot happen.

Definition 6.6: A game has *observed deviators* if for all players *i*, all strategy profiles *s* and all deviations $s'^i \neq s^i$, $h^i \in \overline{H}(s'^i, s^{-i}) \setminus \overline{H}(s)$ implies that there is no s''^i with $h^i \in \overline{H}(s^i, s'^{-i})$.

What this definition requires is that if a deviation by player *i* leads to a new information set off the equilibrium path there is no deviation by *i*'s opponents that leads to the same information set. Games of perfect information satisfy this condition, as do repeated games with observed actions. More generally, the conditions is satisfied by all multi-stage games with observed actions, as defined by Fudenberg and Tirole [1991]. Moreover, Fudenberg and Levine establish that two person games of perfect recall satisfy this condition: with two players, both players must know whether it was their deviation other opponents that led them to a particular information set.

Proposition 6.4 **(**Fudenberg and Levine [1993a]): In games with observed deviators, self-confirming equilibria are consistent self-confirming.

The idea is that with observed deviators, the information sets off the equilibrium path that are reachable when a player's opponents deviates (as described in the definition of

consistency) cannot be reached when the player himself deviates, so that beliefs about play at such information sets are irrelevant.

6.8. Consistent Self Confirming Equilibria and Nash Equilibria

Even consistent self-confirming equilibria, however, need not be Nash. There are two reasons for this difference. First, consistent self-confirming equilibrium allows a player's uncertainty about his opponents' strategies to be correlated, while Nash equilibrium requires that the beliefs be a point mass on a behavior strategy profile.

Example 6.2 [Untested Correlation]: In the game in Figure 6.3 player 1 can play *A* which ends the game, or make any other of three moves, all leading to a simultaneous move game by player 1's opponents, player 2 and player 3, neither of whom observes player 1's move. In this game it is verified in Fudenberg and Levine [1993a] that A is not a best response for player 1 to any behavior strategy of his opponents, but it is a best response to the correlated distribution with puts equal weight on $(L_2, L_3), (R_2, R_3)$. Making use of this observation, we see that in fact the only Nash equilibrium of this game has player 1 playing R_1 and players 2 and 3 giving both actions equal probability. However, player 1 can play A in a consistent self-confirming equilibrium provided his beliefs are the correlated distribution given above.

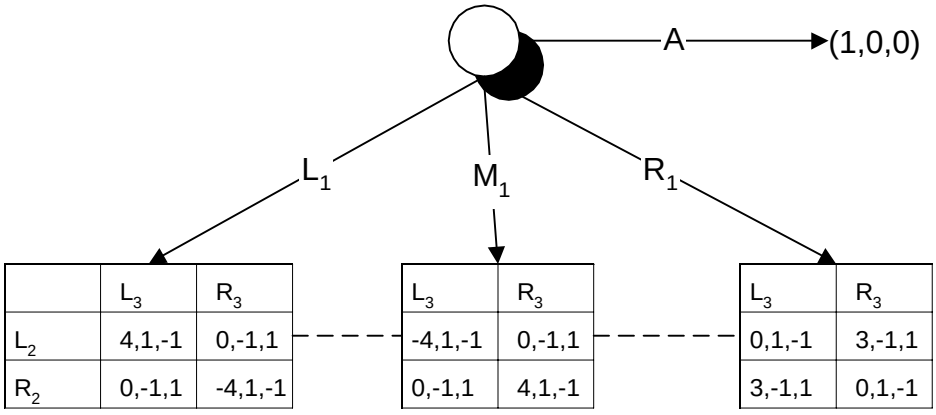


Figure 6.3

The non-Nash outcome in this example arises because of player 1's correlated uncertainty about the play of players 2 and 3. Note well that the support of player 1's beliefs is the (uncorrelated) mixed strategies of players 2 and 3, so that player 1 does not believe that the actual play of his opponents is correlated. Rather, the correlation lies in player 1's subjective uncertainty about his opponents' play.

Of course, this subjective correlation can only arise in games with three or more players. There is a second way that consistent self-confirming equilibria can fail to be Nash that arises even in two-player games. This is because the heterogeneous self-confirming concept allows each s_i that player i assigns positive probability to be a best response to different beliefs. The most immediate consequence of these differing beliefs is a form of convexification, as in the following example.

Example 6.3 [Public Randomization]: In the game in Figure 6.4, player 1 can end the game by moving L or he can give player 2 the move by choosing R.

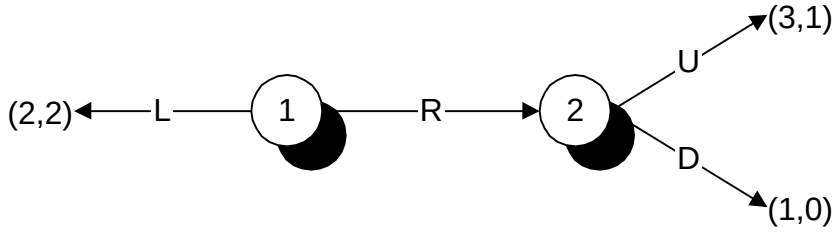


Figure 6.4

Player 1 should play L if he believes 2 will play D, and should play R if he believes 2 will play U. If player 1 plays R with positive probability, player 2's unique best response is to play U, so there are two Nash equilibrium outcomes, (L) and (R,U). The mixed profile ((1/2 L, 1/2 R), U) is a self-confirming equilibrium whose outcome is a convex combination of the Nash outcomes: Player 1 plays L when he expects player 2 to play D, and R when he expects 2 to play U, and when he plays L his forecast of D is not disconfirmed. (Moreover, this equilibrium is clearly independent.)

Although we are unaware of a formal proof, we believe that in all two person games of perfect information the only possible heterogeneous self-confirming equilibria are public randomizations over Nash equilibrium. Absent the restriction to perfect information, self-confirming equilibria in two player games can involve more than convexification over Nash equilibria. The idea is that by embedding a randomization over equilibria as in Example 6.3 in the second stage of a two-stage game, we can induce one player to randomize in the first stage even though such randomization cannot arise in Nash equilibrium. Moreover, this randomization may in turn cause the player's opponent to take an action that would not be a best response without it.

Both the off-path correlation of the first example and the "extra randomization" of the second one can occur in the *extensive-form correlated equilibria* defined by Forges [1986]. These equilibria, which are defined only for games whose information sets are ordered by precedence, are the Nash equilibria of an expanded game where an "autonomous signaling device" is added at every information set. The joint distribution over these signals is assumed to be independent of the actual play of the game, and common knowledge to the players, and the player on move at each information set h is

told the outcome of the corresponding device before he chooses his move. Extensive-form correlated equilibrium includes Aumann's [1974] correlated equilibrium as the special case where the signals at information sets after stage 1 have one-point distributions and so contain no new information. The possibility of signals at later dates allows the construction of extensive-form correlated equilibria that are not correlated equilibria, as in Myerson [1986].

Proposition 6.5: [Fudenberg and Levine, 1993a] For each consistent self-confirming equilibrium of a game whose information sets are ordered by precedence, there is an equivalent extensive-form correlated equilibrium.

Here equivalent means they have the same distribution over terminal nodes, that is, the same outcome. Note that the converse is false in general: even "ordinary" correlated equilibria need not be self-confirming, as is easily seen by considering one-shot simultaneous-move games, where self-confirming equilibrium reduces to Nash.

Corollary 6.1: In two player games every self-confirming equilibrium outcome is the outcome of an extensive form correlated equilibrium.

The discussion and examples above show that there are at least three possibilities that allow non-Nash outcomes to occur in a self-confirming equilibrium: two players may have different (i.e. inconsistent) beliefs about the play of a third one; subjective correlation in a players' beliefs about the play of two or more opponents; and multiple (heterogeneous) beliefs for a single player role. The following result shows that these are the only reasons that a non-Nash outcome can be self-confirming.

Proposition 6.6: [Fudenberg and Levine, 1993a] Every consistent self-confirming equilibrium with independent, unitary beliefs is equivalent to a Nash equilibrium.

The idea, as in the proof of Proposition 6.3, is simply to specify that each player's off-path actions are exactly those that the player's opponents believe would be played. The consistency condition implies that all of player i's opponents expect him to play in the same way, and the independence condition enables us to conclude that we are looking not merely at a correlated equilibrium, but actually a Nash equilibrium.

6.9. Rationalizable SCE and Prior Information on Opponents' Payoffs

Because self-confirming equilibrium allows beliefs about off-path play to be completely arbitrary, it (like Nash equilibrium) corresponds to a situation in which players have no prior information about the payoff functions of their opponents. This may be a good approximation of some real-world situations; it is also the obvious assumption for analyzing game theory experiments in which subjects are given no information about opponents' payoffs. In other cases, both in the real world and in the laboratory, it seems plausible that players do have some prior information about their opponents' payoffs. In an effort to capture this idea, Dekel, Fudenberg and Levine [1996] introduce the notion of "rationalizable self confirming equilibrium".

Consider in particular the game in Example 6.3.

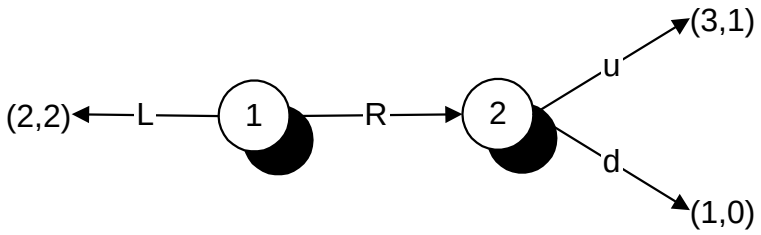


Figure 6.5

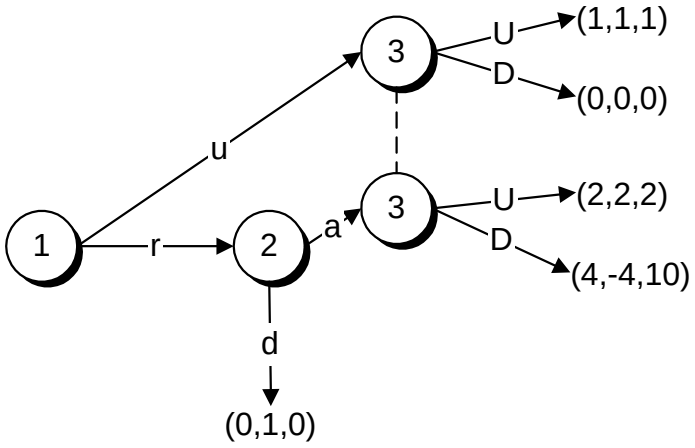
Self-confirming equilibrium allows 2 to play d so long as 2's information set is not reached in the course of play. As noted by Selten [1965], 2 can thus "threaten" to play d, and thus induce 1 to play L. However, this threat is not "credible" if 1 knows 2's payoff function, for then player 1 should realize that player 2 would play u if ever her information set is reached. For this reason, in many settings the weak rationality condition used by Nash and self-confirming equilibrium incorporates too little information about opponents' payoffs.

Although Selten used this example to motivate subgame perfection, it is important to note that the argument given in the last paragraph, taken on its own, only justifies the much weaker conclusion that a player should not use a strategy that is not a best response at the information set in question. In particular, this argument does not provide a rationale for subgame perfection's requirement that expectations about play in an off-path proper subgame should be a Nash equilibrium of that subgame. Dekel, Fudenberg and Levine [1996] propose that the appropriate use of information about opponents' payoffs is through a version of extensive-form rationalizability. The idea is that players should exclude certain strategy profiles from consideration before they observe any information about how the game is actually being played.

The key issue involved in modeling this idea is determining what sort of prior information about payoffs should be considered, as this will determine which strategy profiles are ruled out by the players. One possibility would be to consider predictions consistent with common certainty about payoffs. However, it is well known that predictions of this type are not robust to even a small amount of uncertainty. Since we believe that exact common certainty is more prior information than is reasonable, we focus instead on the strongest restrictions on players' beliefs that are robust to small

amounts of payoff uncertainty. Past work suggests that this assumption should be that payoffs are almost common certainty in the sense of Monderer and Samet (1989).[128] This is captured by a preliminary concept–rationalizability at reachable nodes–that incorporates almost common certainty of the payoffs, and as a result is robust to the introduction of a small amount of uncertainty. In particular, players believe that their opponents' actions will maximize their presumed payoff functions so long as the opponents have not been observed to deviate from anticipated play, but once an opponent has deviated this restriction is no longer imposed. [129]

Before getting into the details, consider the following example, which illustrates some of the possibilities that occur when rationalizability is combined with self-confirming equilibrium. An example from Dekel., Fudenberg and Levine [1996] illustrates the issues involved.



[128] This can be seen, for example, by relating the results of Dekel and Fudenberg [1990] and Börgers [1994]. More specifically, Dekel and Fudenberg applied the FKL notion of robustness to show, roughly speaking, that the tightest robust solution concept that does not impose the common prior assumption is given by deleting one round of weakly dominated strategies and then iteratively deleting strongly dominated strategies. Subsequently, Börgers [1994] showed that this solution concept is characterized by almost common certainty of caution and of payoffs / rationality in the strategic form.

[129] This assumes, in addition to almost common certainty of the payoffs, that the payoffs are determined independently, so that the signal refers only to the deviator's payoffs

Figure 6.6

In this example (u,U) is a Nash outcome (and so certainly self-confirming) since 2's information set is off the equilibrium path, and so he may play d. Intuitively, however, if this were the long-run outcome of as learning process, player 1 should realize that 2 knows that 3 is playing up, and player 1 can use this knowledge and his knowledge of player 2's payoffs to deduce that 2 will play a.

6.9.1. Notation

In order to deal formally with rationalizability and self-confirming equilibrium, it is necessary to introduce additional notation concerning beliefs in extensive form games. An *assessment* a^i for player i is a probability distribution over nodes at each of his information sets. A *belief pair* for player i is a pair $b^i = (a^i, \pi^{i:-i})$ consisting of i's assessment over nodes a^i and i's expectations of opponents' play $\pi^{i:-i} = (\pi^{i:j})_{j \neq i}$. Notice that we are now imposing the independence restriction that a player must belief that his opponents play independently of one another. The belief $b^i = (a^i, \pi^{i:-i})$ is *consistent* (Kreps and Wilson [1982]) if the assessment a^i can be derived from full support approximations to $\pi^{i:-i}$; .

Given a consistent belief by player i, player i's information sets give rise to a decision tree in a perfectly natural way. Moreover, each information set has associated with it a well defined sub-tree that follows after that information set. Each behavior strategy induces a strategy in that sub-tree in a natural way. A behavior strategy is a *conditional best response at* h^i by a player i to consistent beliefs $b^i{}_i$ if the restricted strategy is optimal in sub-tree that follows h^i. (This implicitly supposes that the player will play optimally at subsequent nodes, so a choice that will yield 1 given optimal future play, and 0 otherwise, is just as good as a choice that guarantees a payoff of 1.)

6.9.2. Belief-Closed Sets and Extensive-Form Rationalizability

The basic idea of rationalizability, due to Bernheim [1984] and Pearce [1984], is that, based on his knowledge of the payoffs, each player should have a consistent sequence of conjectures about how his opponent thinks he thinks he should play, and so forth. One method of formalizing this idea is to assign to each player a set of strategy-belief pairs. Each strategy should be a best response to the corresponding beliefs, and each belief in this set should be "rationalized" by the existence of strategies that are in the set of consistent strategy-belief pairs for other players. It is convenient to separate out this latter idea of "belief-closedness" in a separate definition. When combined below with the requirement that the strategies be best-responses to beliefs, we get a definition of rationalizability.

Definition 6.7: The sets of strategy-belief pairs $SB^1,\ldots,SB^n$ are *belief-closed* if $(\pi^i,(a^i,\pi^{i:-i}))\in SB^i$ implies that $\pi^{i:j}$ is in the convex hull of $\left\{\tilde{\pi}^j|(\tilde{\pi}^j,b^j)\in SB^j \text{ for some } b^j\right\}$.

In words, if *i* believes *j* can choose some behavior strategy then that strategy must be in *j*'s set of possible choices. As we indicated above, the elements of the sets SB^j are better viewed as "things that player *i* might think player *j* will do" than as "things *j* is likely to do ex-ante." For example, if *j*'s strategy specifies an action at some off-path information set that is not optimal given *j*'s specified payoffs, the interpretation is that this is something *i* plausibly thinks that *j* would do if that information set is reached.

Using the idea of belief closed sets, Dekel., Fudenberg and Levine [1996] define notion of rationalizability for extensive form games.[130]

[130] Related notions can be found in Basu [1985], Reny [1992], Rubinstein and Wolinsky [1994] or Greenberg [1994].

Definition 6.8: An *n*-tuple of strategy-belief pair sets $SB^1,\ldots,SB^n$ is *rationalizable at reachable nodes* if for all *i*:

1'. If $(\pi^i,b^i)\in SB^i$ then π^i is a best response to b^i at information sets reachable under π^i.

3. $SB^1,\ldots,SB^n$ are belief closed.

For reasons of robustness that we discuss further below, this notion does not require rationalizability at all nodes.

It is useful to provide an equivalent definition of self-confirming equilibrium that incorporates the notion of belief closedness; the point is that without the additional requirement of rationalizability, the belief closedness itself has no force.

Proposition 6.7 Profile $\hat{\pi}$ is a *self-confirming equilibrium* iff and only if there are sets of strategy-belief pairs, $SB^1,\ldots,SB^n$ such that, for all players *i,*

1. If $(\pi^i,b^i)\in SB^i$ then π^i is a best response to b^i at information sets that are reached with positive probability under $(\pi^i,\pi^{i:-i})$.

2. Every $(\pi^i,b^i)\in SB^i$ has the distribution over outcomes induced by $\hat{\pi}$.

3. $SB^1,\ldots,SB^n$ are belief closed.

If we now add the requirement that π^i is a best response at all reachable information sets, rather than merely all information sets that are reached with positive probability, we then get

Definition 6.9: Profile $\hat{\pi}$ is a *rationalizable self-confirming equilibrium* iff and only if there are sets of strategy-belief pairs, $SB^1,\ldots,SB^n$ such that, for all players *i,*

1'. If $(\pi^i, b^i) \in SB^i$ then π^i is a best response to b^i at information sets reachable under $(\pi^i, \pi^{i:-i})$.

2. Every $(\pi^i, b^i) \in SB^i$ has the distribution over outcomes induced by $\hat{\pi}$.

3. $SB^1, \ldots, SB^n$ are belief closed.

Turning back to the game of example 6.3

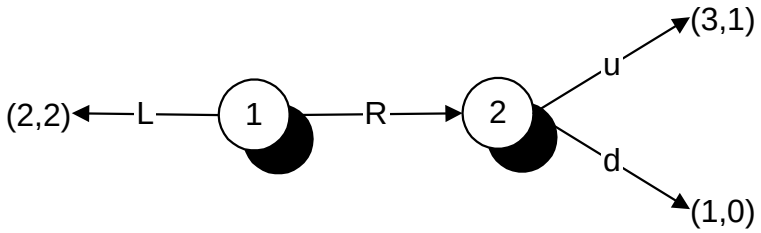


Figure 6.7

we see that the rationalizable self-confirming equilibrium notion captures what we wanted: L is not part of any beliefs that are rationalizable at reachable nodes. To see this, observe that 2's information set is always reachable, so condition 1' implies that the only strategy in SB^2 is u. From condition 3, player 1 must believe this, and so he plays R.

6.9.3. Robustness

An important feature of rationalizable self-confirming equilibrium is that a strategy need not be optimal at information sets that the strategy itself precludes. The reason that we do not wish to impose optimality at such information sets is that this stronger requirement is not robust to the presence of a small amount of payoff uncertainty. To see this, consider the game in Figure 6.8.

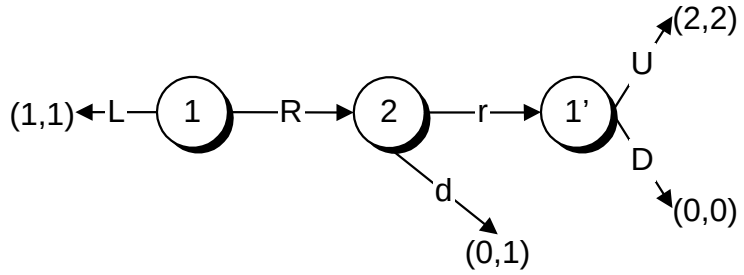


Figure 6.8

In this game the outcome L occurs in the Nash equilibrium (LD,d), but not in any subgame-perfect equilibrium. However in the game of incomplete information in Figure 6.9,

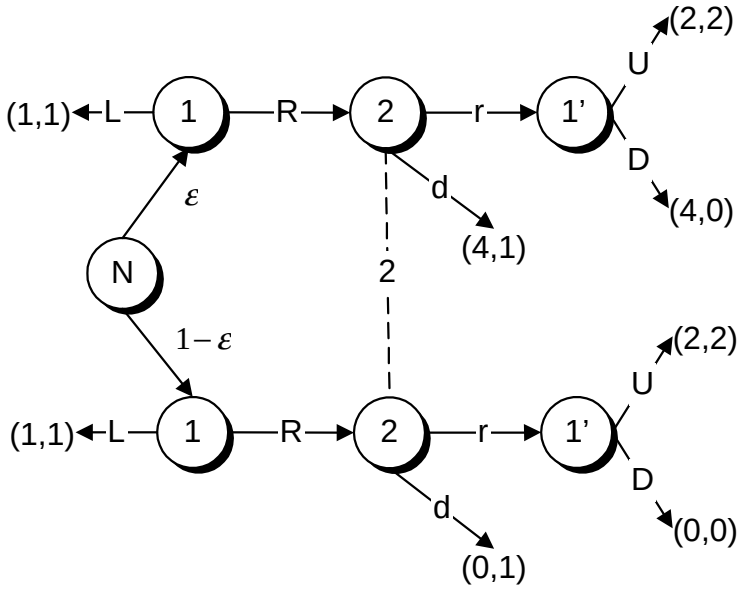


Figure 6.9

where payoffs are very likely to be as in figure 6.8, the outcome L occurs in a sequential equilibrium. So requiring optimality at all information sets rules out the outcome L in Figure 6.8 but not in 6.9; hence this requirement is not robust to small payoff

uncertainties.[131] It is easy to see that by construction rationalizable self confirming equilibrium achieves our objectives in Figure6.8: since player 1's second information set is not reachable when 1 plays L, the outcome L can occur in a rationalizable self confirming equilibrium

6.9.4. Example 6.1 revisited

Ordinary self-confirming equilibrium allows two players to disagree about the play of the third. This example demonstrates the intuitive idea that the possibilities for such disagreements are reduced when players must believe that opponent's play is a best-response at reachable nodes. Consider the following version of the extensive-form game Fudenberg and Kreps [1988] used to show that mistakes about off-path play can lead to non-Nash outcomes:

[131] Just as in previous work related to this notion of robustness, one may be able to identify a smaller set of robust predictions if one feels confident that certain forms of payoff uncertainty are much less likely than others. We say more about this in the next section. For more about the idea of robustness, see Fudenberg, Kreps and Levine [1988] and Dekel and Fudenberg [1990].

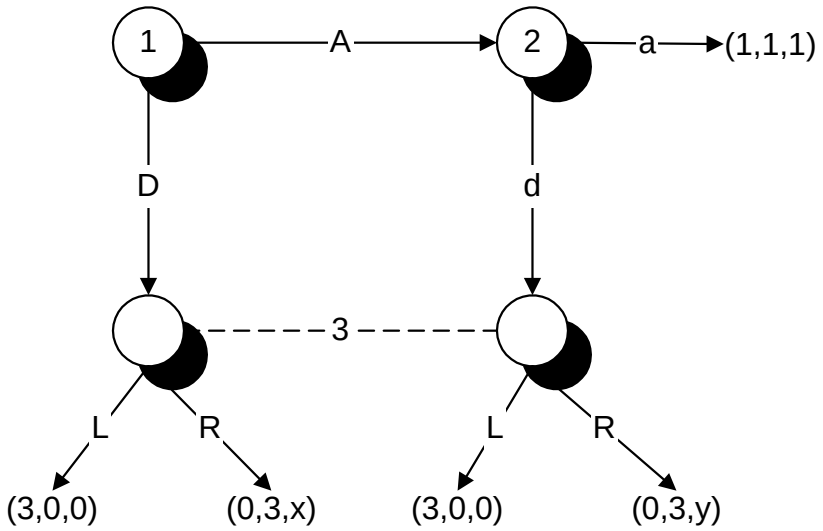


Figure 6.10

Here the outcome (A,a) is self-confirming for any values of x and y. It is supported by player 1 believing that player 3 will play R and player 2 believing that player 3 will play L. However, because 3's information set is reachable, this outcome is not RSCE if both x and y have the same sign: If $x,y>0$ then players 1 and 2 forecast that 3 will play R, and so 2 plays d; if $x,y<0$ then 3 plays L so 1 plays D. However, if x and y have opposite signs, then (A,a) is a RSCE outcome, since 1 and 2 are not required to have the same beliefs about player 3's off-path assessment of the relative probability of the nodes w and w', and player 1 can think that 3's assessment makes R optimal, while player 2 can think that 3's assessment induces her to play R. This example shows that even a sequentially RSCE need not be Nash.

6.9.5. Experimental Evidence

Perhaps the best motivation for rationalizable self-confirming equilibrium is a pair of experiments by Prasnikar and Roth [1992] on the "best-shot" game, in which two players sequentially decide how much to contribute to a public good. The only

rationalizable self-confirming equilibrium of this game is its backwards-induction solution, in which the first mover contribute nothing; there is also an imperfect Nash equilibrium in which the first mover contributes and the second does not. Prasnikar and Roth ran two treatments of this game. In the first one, players were informed of the function determining opponents' monetary payoffs. Here, by the last few rounds of the experiment the first movers had stopped contributing, which is the prediction made by rationalizable self-confirming equilibrium. In the second treatment, subjects were not given any information about the payoffs of their opponents. In this treatment even in the later rounds of the experiment many first movers contributed to the public good. This is not consistent with rationalizable self-confirming equilibrium, but it is consistent with an (approximate, heterogeneous) self confirming equilibrium (Fudenberg and Levine [1996]). Thus these experiments provide evidence that information about other players' payoffs makes a difference, and that this difference corresponds to the distinction between self-confirming equilibrium and rationalizable self-confirming equilibrium.

References

Aumann, R. [1974]: “Subjectivity and Correlation in Randomized Strategies,” *Journal of Mathematical Economics*, 1: 67-96.

Basu, K. [1985]: “Strategic Irrationality in Extensive Games,” Princeton University.

Battigalli, P. [1987]: “Comportamento Razionale ed Eqiulibrio nei Giochi e nelli Situatzioni Sociali,” Bocconi University Ph.D. Dissertation.

Bernheim, D. B. [1984]: “Rationalizable Strategic Behavior,” *Econometrica*, 52: 1007-1028.

Dekel, E. and D. Fudenberg [1990]: “Rational Behavior with Payoff Uncertainty,” *Journal of Economic Theory*, 52: 243--267.

Dekel, E., D. Fudenberg and D. K. Levine [1996]: “Payoff Information and Self-Confirming Equilibrium,” HIER D.P. 1774 Harvard University.

Forges, F. [1986]: “An Approach to Communication Equilibrium,” *Econometrica*, 54: 1375-1386.

Fudenberg, D. and D. K. Levine [1993a]: “Self-Confirming Equilibrium,” *Econometrica*, 61: 523-546.

Fudenberg, D. and D. K. Levine [1993b]: “Steady State Learning and Nash Equilibrium,,” *Econometrica*, 61 (May): 547-573.

Fudenberg, D. and D. K. Levine [1996]: “Measuring Subject’s Losses in Experimental Games,” *Quarterly Journal of Economics*, forthcoming.

Fudenberg, D. and D. M. Kreps [1988]: “Learning, Experimentation and Equilibrium in Games,” Stanford.

Fudenberg, D. and D. M. Kreps [1995b]: “Learning in Extensive Games, I: Self-Confirming Equilibrium,” *Games and Economic Behavior*.

Fudenberg, D. and D. M. Kreps [1995a]: "Learning in Extensive Games, II: Experimentation and Nash Equilibrium," Harvard.

Fudenberg, D. and J. Tirole [1991]: *Game Theory*, (Cambridge: MIT Press).

Fudenberg, D., D. M. Kreps and D. K. Levine [1988]: "On The Robustness of Equilibrium Refinements," *Journal of Economic Theory*, 44: 354--380.

Greenberg, J. [1994]: "Social Situations Without Commonality of Beliefs: Worlds Apart-but Acting Together," McGill Univ. W.P. 8/94.

Hahn, F. [1977]: "Exercises in Conjectural Equilibrium," *Scandinavian Journal of Economics*, 79: 210-226.

Kalai, E. and E. Lehrer [1993]: "Rational Learning Leads to Nash Equilibrium," *Econometrica*, 61: 1019-1046.

Kreps, D. and R. Wilson [1982]: "Reputation and Imperfect Information," *Journal of Economic Theory*, 50: 253-79.

Myerson, R. [1986]: "Bayesian Equilibrium and Incentive Compatibility: An Introduction," In *Social Goods and Social Organization: Essays in Honor of Elizha Pazner*, Ed. L. Hurwicz, D. Schmeidler, and H. Sonnenschein, (Cambridge: Cambridge University Press).

Pearce, D. [1984]: "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52:1029-1050.

Prasnikar, V. and A. E. Roth [1992]: "Considerations of fairness and strategy: experimental data from sequential games," *Quarterly Journal of Economics*, 865-888.

Reny, P. [1992]: "Rationality in Extensive-Form Games," *Journal of Economic Perspectives*, 6: 103-118.

Rubinstein, A. and A. Wolinsky [1994]: "Rationalizable Conjectural Equilibrium: Between Nash and Rationalizability," *Games and Economic Behavior*, 6: 299-311.

Selten, R. [1965]: "Spieltheoretische Behandlung eines Oligopmodells mitnachfragetragheit," *Z. Ges. Staatswiss.*, 121: 301-324.

7. Nash Equilibrium, Large Population Models, and Mutations in Extensive Form Games

7.1. Introduction

As we have seen, there is no presumption that simple learning in extensive-form games

leads to Nash equilibrium outcomes, even when the learning process converges. However, a convergent learning process *will* converge to a Nash equilibrium outcome if it generates "enough" learning about off-path play. This chapter explores the related issues of just how much information is "enough," and what sorts of forces might lead to "enough" information being available. While we discuss several explanations, our focus is on the idea that players sometimes deliberately "experiment" with actions that do not maximize the current period's payoff expected payoff in order to gain information about how their opponents react to these little-played actions.

As the first step in this chapter, we address the question of how much information about opponents' play is "enough" to rule out all but Nash equilibrium outcomes. The usual definition of Nash equilibrium implies that players know the entire strategy profile used by their opponents, or equivalently the distribution of actions that would occur at any information set. However, this is more knowledge than is necessary, as a given player's beliefs about play at some information sets may have no impact at all on how he chooses to play. Instead, it suffices that players have correct beliefs at those information sets that are "relevant" to them. We formalize this idea in Section 7.2.

Section 7.3 develops sufficient conditions on exogenously specified behavior (in the spirit of fictitious play) that lead to Nash equilibrium. Section 7.4 then examines

these assumption, how they might be relaxed, and the connection between learning in games and in multi-armed bandit problems. Section 7.5 considers a model of fully rational, Bayesian learning in which experimentation rates are endogenous. In order to avoid some of the problems discussed in section 7.4, this is done in the context of a model of steady-state learning in large, heterogeneous populations. This model also provides a foundation for heterogeneous self-confirming equilibrium.

One obvious question in this area that has so far been little explored is the extent to which it is possible to establish convergence to a refinement of Nash equilibrium. Section 7.6 discusses the work of Noldeke and Samuelson [1993] that relates the stochastically stable outcomes of a particular learning process to the subgame-perfect equilibria in a special class of games.

We conclude with a discussion of cheap talk games and return to the idea (discussed in chapter 3) that players can give a "secret hand-shake", a signal that they intend to carry out a particular action. We give a critical overview of the literature on evolutionary dynamics in this game, and suggest that future work on this topic should take account of the extensive-form nature of the game.

7.2. Relevant Information Sets and Nash equilibrium

Self-confirming equilibrium need not have be Nash because some players may have incorrect beliefs about off-path play. However, to conclude that a particular self-confirming equilibrium is Nash, it is not necessary to assume that every players' beliefs are correct at every information set. In particular, since Nash equilibrium tests only for unilateral deviations, a player's beliefs about what would happen if some other player deviated are irrelevant. To capture this, Fudenberg and Kreps [1995b] introduce the following definition:

Definition 7.1: An information set h is *relevant to player i at profile π_** if there is some π^i such that (π^i, π_*^{-i}) assigns positive probability to *h*; the set of all such information sets is denoted $\hat{H}^i(\pi_*)$.

As in chapter 6, let $\Pi^{-i}(\pi_* \mid J) \equiv \{\pi^{-i} \mid \pi^j(h^j) = \pi_*^j(h^j \mid \sigma^j), \forall h^j \in H^{-i} \cap J\}$ be the subset of behavior strategies consistent with players other than *i* playing according to π_* at the information sets in *J*.

Proposition 7.1 (Fudenberg and Kreps (1995b)): A strategy profile π_* is a Nash equilibrium if there exist beliefs μ^i such that

- $u^i(\pi_*{}^i \mid \mu^i) \geq u^i(\pi^i \mid \mu^i)$ for all π^i, and
- $\mu^i(\Pi^{-i}(\pi_* \mid \hat{H}^i(\pi_*)) = 1$.

This shows that it's sufficient for Nash equilibrium that beliefs be correct at relevant information sets. It is obvious that even this condition is not necessary. For example, if player *i* doesn't get to move along the path of π_* his beliefs are irrelevant, and if the payoff inequalities in the definition hold strictly they will continue to hold if the beliefs are slightly incorrect at every information set.

However, the result does show that it in order for a non-Nash profile to be unstable in a learning model, it is sufficient that beliefs become approximately correct at those information sets which are relevant given the profile.

This in turn raises the question of when that will be the case. Intuitively, beliefs about play at an information set will be correct if that information is reached sufficiently often, so that players have "enough" observations about play at the information set to outweigh their possibly incorrect priors. Moreover, unless we are prepared to make assumptions about the *strength* of the players' prior convictions (that is, the size of the

fictitious initial sample in fictitious play) "enough" observations means infinitely many of them. Of course, any assumption that implies a positive probability that a (subjectively) suboptimal action will be played infinitely often is inconsistent with optimal behavior in the discounted multi-armed bandit problem, for which the optimal solution (with a full-support prior) has probability 1 that experimentation ceases in finite time, with a positive probability of "locking on" to the objectively "wrong" arm. (The appendix reviews the classic multi-armed bandit problem.)

Consequently any assumptions that imply probability 1 of all relevant information sets being reached infinitely often, regardless of the priors (and consequently probability 0 of convergence to a non-Nash self-confirming outcome), are not consistent with optimal behavior in the discounted bandit problem. The reason for interest in such assumptions is that they do correspond to the limit of behavior in the bandit problem as the discount factor goes to 1. Intuitively, as players become more patient, the value of information increases, so they do more and more experiments, and the probability of locking on to the wrong arm converges to 0. As a result, the "sufficient experimentation" conditions in the following section should be interpreted as an idealization of the limit behavior as the discount factor tends to 1.

7.3. Exogenous Experimentation

Fudenberg and Kreps [1995b] develop sufficient conditions for instability of non-Nash equilibrium and local stability in Nash equilibrium in a model of boundedly rational behavior that is in the spirit of fictitious play. Their assumptions imply that if the strategies played converge, then all relevant information sets (given the limit profile) are reached infinitely often, that beliefs at these information sets converge to the empirical

distribution of play there, and that the empirical distribution resembles the limit profile to which play is converging. The latter two conditions are imposed by strengthening the asymptotic myopia and asymptotic empiricism conditions developed in Chapter 6; the first condition, that all relevant information sets are reached infinitely often, is obtained by imposing lower bounds on the probabilities that players "experiment" in various ways.

The reason that the experimentation condition on its own is not sufficient is that the definitions of asymptotic empiricism and myopia given in chapter 6 impose no restrictions at all on beliefs or behavior at information sets that are reached infinitely often but a vanishing fraction of the time. It is easy to strengthen the empiricism condition. Fix an infinite history h_∞, and let $H^i_\infty(h_\infty)$ be the collection of player *i* information sets that are reached infinitely often along h_∞:

Definition 7.2: Player *i's* belief rule is *strongly asymptotically empirical in the extensive form* if for every $\varepsilon > 0$, infinite history h_∞, $j \neq i$, and information set $h^j \in H^j_\infty(h_\infty)$,

$$\lim_{t\to\infty} \mu^i_t(h_t)\left(\left\{\pi^{-i} \,\middle|\, \left\|\pi^j(h^j) - d(h^j|h_t)\right\| < \varepsilon\right\}\right) = 1.$$

This condition is satisfied by Bayesian learners who believe that opponents' play corresponds to a fixed but unknown distribution (that is, exchangeable draws), and have a non-doctrinaire prior over the set of all strategy profiles for the opponents.

In order for Nash equilibrium to be reached in the limit, players must engage in "enough" experimentation to learn about off-path play; in particular the "rate" of experimentation cannot decrease too quickly. At the same time, however, these experiments must vanish quickly enough that they are a nonnegligible component of asymptotic play. We first modify the definition of asymptotic myopia to include a limited and asymptotically negligible amount of experimentation. Let $\kappa(a, h_t)$ denote the

number of time the action *a* has been played in the history h_t, and $\kappa(h^i, h_t)$ the number of time the information set h^i has occurred.

Definition 7.3: For a particular forecast rule γ^i *a* behavior rule ρ^i is *strongly asymptotically myopic with experience-time limitations on experimentation* if it can be decomposed into two rules, a "myopic" rule $\hat{\rho}^i$ and an "experimentation" rule $\tilde{\rho}^i$ such that

(a) $\rho^i(h_t)(h^i) = a^i(h_t)(h^i)\hat{\rho}^i(h_t)(h^i) + (1 - a^i(h_t)(h^i))\tilde{\rho}^i(h_t)(h^i)$ for some $a^i(h_t)(h^i) \in [0,1]$

(b) $\hat{\rho}^i$ is asymptotically myopic

(c) there is a non-negative sequence $\eta_t \to 0$ such that $(1 - a^i(h_t)(h^i))\tilde{\rho}^i(h_t)(h^i)(a) > 0$ only if $\kappa(a, h_t) / \kappa(h^i, h_t) \leq \eta_{\kappa(h^i, h_t)}$.

In other words, the probability assigned to an "experimental" action must be zero unless the action has been tried infrequently.

The experience-time limitations on experimentation imply that, asymptotically, play is with high probability asymptotically myopic. In particular, the proportion of the time that *i* experiments at an information set must go to 0 as the number of times that the information set is reached becomes large. On the other hand, to attain Nash equilibrium, it is necessary also that there be "enough" experimentation.

Definition 7.4: For a given player *i* and information set h^i the behavior rule ρ^i satisfies the *minimal experience-time experimentation condition* at h^i if there is a constant $\beta > 0$ and a non-negative sequence $\nu_t \to 0$ with $t\nu_t$ non-decreasing such that

$$\rho^i(h_t)\left(a \in A(h^i) \middle| \kappa(a, h_t) / \kappa(h^i(a), h_t) \leq \nu_{\kappa(h^i(a), h_t)}\right) \geq \beta.$$

In other words, actions that have been played infrequently, should be tried with at least probability β.

The force of this condition can be seen from the following result:

Proposition 7.2: If player i's behavior satisfies the minimal experience-time experimentation condition at information set h^i, then for every $a \in A(h^i)$,

$$P\left(\{h_\infty | \lim_{t\to\infty} \kappa(h^i, h_t) = \infty \text{ and } \lim_{t\to\infty} \kappa(a, h_t) < \infty\}\right) = 0.$$

Roughly speaking, this says that if h^i is reached infinitely often, then every action that is feasible there must be taken infinitely often. This is quite a strong conclusion, and indeed it suggests that the so-called "minimal experience-time experimentation condition" may require more experimentation than is plausible. We discuss this issues in the next section. For now we note the following corollary: In a game of perfect information, if minimal experience-time experimentation is satisfied at *every* information set, then with probability 1 every information set is reached infinitely often.

Surprisingly, these assumptions are not enough to preclude convergence to non-Nash equilibrium in games of imperfect information. The following example, from Fudenberg and Kreps [1995b], illustrates the potential problem.

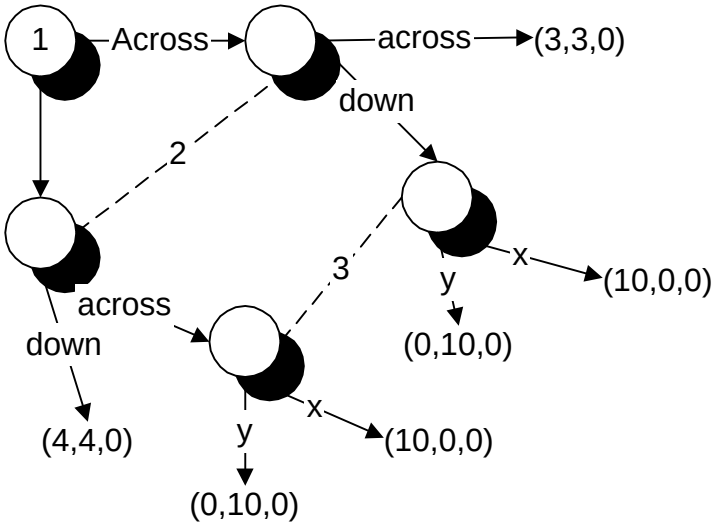


Figure 7.1

The outcome (Across, across) is not a Nash equilibrium, as for any strategy of player 3, at least one of them would prefer to deviate. Suppose that player 1 initially believes 3 will play y, so that Across is player 1's short-run optimum, while 2 believes that 3 will play x so that 2's myopic best response is to play across. Suppose moreover that both players choose to "experiment" with their other, apparently sub-optimal, action at dates 1, 10, 100, 1000, and so forth. The behavior rules satisfy the minimal experience-time experimentation condition, yet player 3's information set is never reached. Fudenberg and Kreps show that this problem can be avoided with either of two additional assumptions. The following is the simpler but less palatable of the two:

Definition 7.5: The behavior rule ρ^i is *uniformly non-experimental* iff the probability $a^i(h_t)(h^i)$ of following the nonexperimental rule $\hat{\rho}^i$ at information set h^i given history h_t is uniformly bounded below by some $\alpha > 0$.

This requires that there always be at least an α chance a player does not experiment, so that experimentation by the opposing player has a chance of revealing information about

his non-experimental play. However, this assumption is inconsistent with optimal play in a bandit problem. We discuss this in more detail below. As an alternative, Fudenberg and Kreps suggest that stability to be redefined so to exclude histories in which the players somehow perfectly coordinate their experiments; we say more about this in the next section. The next section also explains why an analog of uniform non-experimentation hold in models with anonymous random matching in a large population.

Proposition 7.3: (Fudenberg and Kreps [1995b] If beliefs are strongly asymptotically empirical, and behavior rules satisfy asymptotic myopia with experience time limitations on experimentation, the minimal experience-time experimentation condition at all information sets, and are uniformly non-experimental then if π is not a Nash profile it is unstable; if π is a Nash profile, then it is weakly stable.

It is perhaps not surprising that with the "right" amount of experimentation only Nash equilibria can be reached. Intuitively, the combination of asymptotic myopia and asymptotic empiricism implies that the limit point must be self-confirming equilibria, as in chapter 6. Moreover, at least in games of perfect information the assumption of minimal experience-time experimentation at all information sets implies that every information set is reached infinitely often. Hence, in such games, if play converges, players come to have correct beliefs about play at every information set, and so the limit point must be a Nash equilibrium. In more general games, minimal experience-time experimentation need not imply that all information sets are reached, as in the example above; this is why an additional assumption is needed.

7.4. Learning in Games Compared to the Bandit Problem

The assumptions that give local stability of Nash equilibria and instability of non-Nash equilibria are quite strong. In particular, the uniform non-experimental condition is inconsistent with Bayesian optimization in a multi-armed bandit problem. In this section we consider alternative assumptions giving the same result, and discuss more generally the issue of how learning about an extensive form game differs from learning in a bandit problem.

The classical bandit problem is a simple one-move one-person extensive form game with random payoffs to each action, where the distributions of payoffs for some actions are unknown, and the payoffs to the various actions are distributed independently, so that observing the payoff to one action reveals no information about the distributions governing the payoffs to other choices.[132] It is well known that even in a bandit problem, an impatient player may fail to optimize: if it is believed *a priori* that a particular arm is inferior, it may never be tried, even if, in fact, it is superior. In fact, for any fixed discount factor, experimentation in a bandit problem end in finite time with probability one.[133] However, in the limit as the discount factor goes to one, the amount of time during which experimentation takes place goes to infinity, and the probability of a suboptimal choice goes to zero. In the previous section the basic assumption was that experimentation continues forever. This should be viewed as an effort to capture the limit of optimal play in discounted bandit problems as the discount factor tends towards 1. In the remainder of this section, we will use this limit as motivation for the types of rules that we would like to allow.

---

[132] There is a smaller literature on bandit problems with correlated payoffs. Moreover, one way thinking about learning in extensive-form games is that it corresponds to a bandit problem with a particular and potentially complex form of correlation.

[133] A more detailed discussion of bandit problems can be found in the Appendix to this chapter.

As we indicated, the uniform non-experimental condition is inconsistent with optimal play in either the discounted or undiscounted bandit problem, since the optimal solution will typically involve playing an experimental action with probability 1 at some histories.[134] There are two answers to this problem. One possibility, explored in more detail below, is that the probability of non-experimentation represents a probability of meeting an opponent who is not experimenting in a matching setting. Another possibility is to drop the assumption of uniform non-experimentation altogether. An alternative, proposed by Fudenberg and Kreps, is to modify the definition of stability to include a condition that observed play passes some simple "statistical tests" of exchangeability and independence. The idea is that if the observed histories fail the tests then players should realize that the environment is not after all asymptotically exchangeable and independent. Fudenberg and Kreps then verify that play can converge to Nash equilibrium and satisfy the statistical tests, while play cannot both satisfy the tests and converge to a non-Nash outcome, even when uniform non-experimentation is not required. This formulation does not address how players behave if some player's statistical test fails.[135]

We next examine the minimal experience time condition, which requires experimentation with *all* actions that have been tried infrequently. This is certainly the right strategy in a the classic bandit problem, where the payoffs to the various arms are distributed independently. However, there are several reasons why this might not be optimal in a game. First of all, a player might have several actions each of which lead to the same information sets of all opponents, and which thus yield the same information.

---

[134] It is true that uniform non-experimentation is consistent with ε-optimization in the undiscounted bandit problem, but then so is any fixed and small amount of "trembling" onto other actions. This brings us back to the point made at the end of chapter 6: non-Nash self-confirming outcomes should be viewed as descriptions of what will happen up to some time T for sufficiently small amounts of noise.

[135] Chapter 8 discusses work which does specify how players behave if they detect certain sorts of departures from the assumptions of exchangeability and independence.

Since the payoffs of these actions need not be equal, it makes sense to suppose (and the optimum requires) that the player would experiment only with the action which involved the smallest expected loss.[136] The following example from Fudenberg and Levine [1993] shows how this might happen

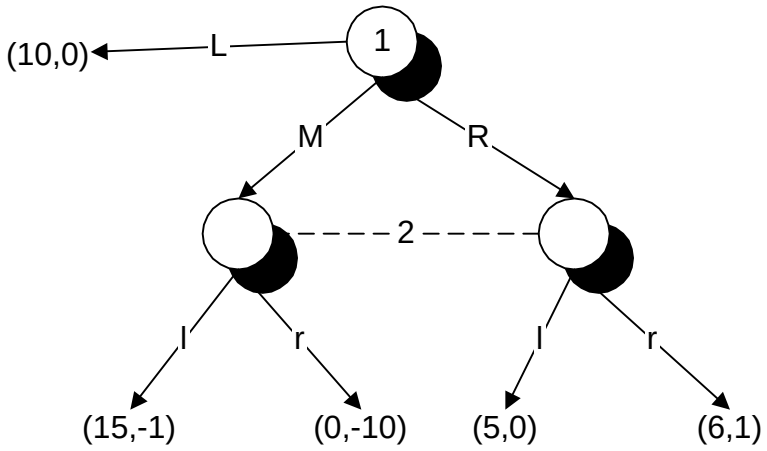


Figure 7.2

Suppose that player 1 assigns a low probability to player 2 playing l. In this case his immediate expected payoff is maximized by playing L himself. Suppose, however, that player 1 is willing to conduct a costly experiment to obtain information about player 2's play. Given player 1's beliefs, the lowest-cost way of obtaining this information is by playing *R*, and indeed, it is quite possible that player 1 will never play M.[137]

[136] But note that such considerations suggest equilibrium refinements in the spirit of Myerson's [1978] properness, because out-of-equilibrium actions tend to be taken as cheaply as possible.

[137] As an aside, we note that this example also shows why optimal experimentation does not yield results in the spirit of forward induction (Kohlberg and Mertens [1986]). Forward induction interprets all deviations from the path of play as attempts to gain in the current round. Since L strictly dominates R, forward induction argues that player 2 will believe that player 1 has played M whenever player 2's information set is reached, and hence that player 2 will play l; this will lead player 1 to play M. In contrast, in our model player 1 deviated from L to gain information that will help him in future rounds, and the cheapest way to do this is to play R. When R is more likely than M, r is optimal for player 2.

There is no very easy modification of the minimal experience time requirement that allows experimentation to be least cost. However, our discussion below of Bayesian learning in a steady state setting takes account of the full optimization process, including the requirement that experiments be chosen with an eye to costs as well as benefits.

A second issue with the minimal experience-time experimentation condition is that we have required it to hold at all information sets. The following example from Fudenberg and Kreps [1995b] shows why this is problematic:

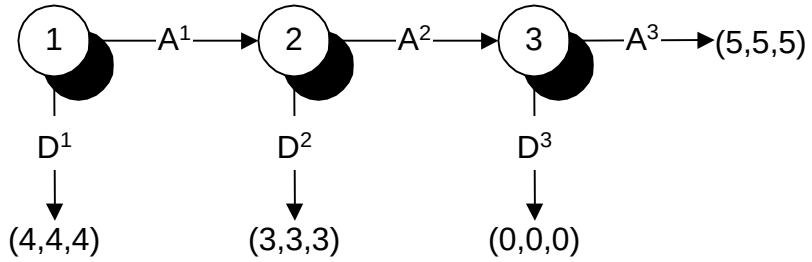


Figure 7.2

Suppose that along some history, player 1 chooses A^1 infinitely often but a vanishing fraction of the time, and that player 2 starts out with the assessment that player 3 is more likely to play D^3 than A^3. Then player 2, in periods when her information set is reached, would see A^2 as a costly but potentially worthwhile experiment. However, the experiment of playing A^2 only pays off if, first of all, player 2 learns that player 3 usually plays A^3, and *in addition* player 1 gives player 2 an opportunity to use that information by playing A^1 again in the not-too-distant future. Since player 2 has few observations on 3's play, she should assign a nonnegligible probability to the event that player 3 usually plays A^3, and so she should expect that she may indeed have something to learn. However, given that player 1 plays A^1 with frequency going to 0, even a very patient, optimizing player 2 might not find it worthwhile to do any experiments with A^2. For

this reason it is not sensible to require the minimal experience time condition at every information set.

For this reason, it is important to note that Nash equilibrium does not require that the minimal experience-time experimentation condition be satisfied at every information set. Instead, it is sufficient that the condition be satisfied at information sets that player *i* feels are "empirically relevant," where loosely speaking information set h^i is empirically relevant given an infinite history if it is reached a "sufficiently large" proportion of the times that it "might have been reached." Formally, Fudenberg and Kreps [1995b] define

Definition 7.4: The behavior rule ρ^i satisfies the *modified minimal experience-time experimentation condition* or *MME* if there is a constant $\beta > 0$, and a non-negative sequence $\nu_t \to 0$ with $t\nu_t$ non-decreasing, and a non-increasing sequence of strictly positive numbers $\delta_k \to 0$ such that for all t and h_t if $a_1^i, a_2^i, \ldots$ is the unique sequence of actions by player *i* that lead to h^i satisfies

$\{\kappa(h^i(a_k^i), h_t) / \kappa(a_k^i, h_t) \geq \delta_{\kappa(a_k^i, h_t)}$ then

$\rho^i(h_t)\left(a \in A(h^i) \middle| \kappa(a, h_t) / \kappa(h^i(a), h_t) \leq \nu_{\kappa(h^i(a), h_t)}\right) \geq \beta.$

The force of this condition can be seen from the following result.

Proposition 7.4: Suppose that player i's behavior satisfies the modified minimum experience-time experimentation condition, and that there is a profile π_* and $\varepsilon > 0$ such that for all infinite histories h in some set Z, and all times t, at every partial history h_t the behavior rules ρ^i assign probability at least ε to every action a for which $\pi_*(a)$ is positive. Then almost surely on Z, every information set that is π_*-relevant to player i will be reached infinitely often.

Roughly speaking, the conclusion of this proposition is that every information set that "matters" to the player is reached infinitely often. Note that the allows there to be probability 1 that player 3's information set to be reached only finitely often in the example of figure 7.2, since the limit profile assigns probability 0 to player 2's information set. (Moreover, it is easy to construct behavior rules that satisfy MME for all players, yet which imply probability 1 that player 3's information set is reached only finitely often.) In contrast, if every player's behavior satisfies minimal experience time experimentation, then as we noted following Proposition 7.2, player 3's information set is reached infinitely often with probability 1. However, we know from Proposition 7.1 that since player 2' s information set is never reached in the limit profile, player 2's beliefs about subsequent play are immaterial.

Fudenberg and Kreps show that the MME condition can be used in place of minimal experience time experimentation to prove results in the spirit of proposition 7.3. In particular, play cannot converge to a non-Nash outcome if beliefs are strongly asymptotically empirical, behavior satisfies MME and is asymptotically myopic with experience-time limitations on experimentation, and either players use statistical tests of independence and exchangeability or behavior play satisfies uniform non-experimentation.

Finally, although we noted that for any given discount factor, experimentation in a classic bandit problem should stop in a finite amount of time, in the setting of extensive-form games there can be histories along which players find it optimal to experiment in a positive fraction of the time, even in the long run. The following example illustrates the complications that may occur if certain unrepresentative samples occur:

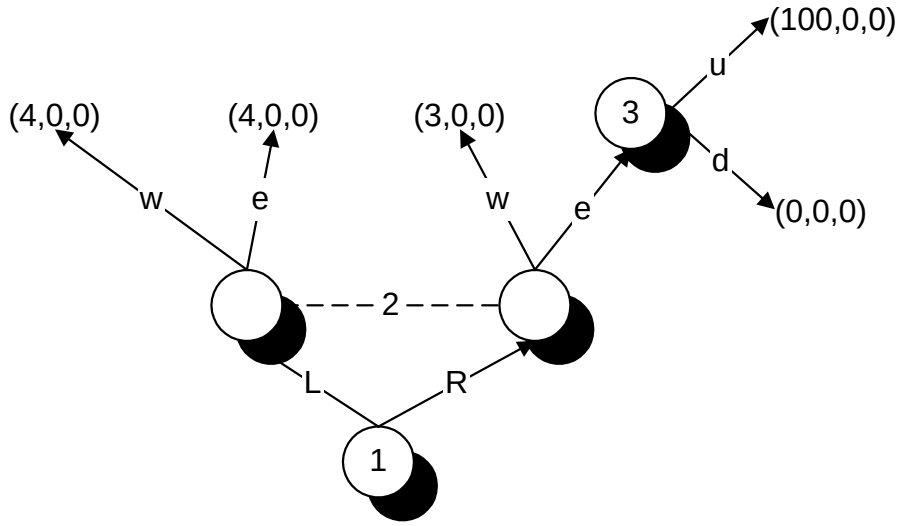


Figure 7.3

Suppose that player 1 has played both L and R many times, and equally frequently. Suppose that it has just happened that when 1 played L, 2 played w half the time, but when 1 played R, 2 has always played w. Since player 1 knows that 2 has an information set, he knows that the actual probability of 2 playing e is about 1/4. Despite this, he has never actually seen player 3 play, and does not know whether 3 is playing u, in which case R would be best, or whether 3 is playing d, in which case L would be best. *A priori* 1 may believe that 3 is likely playing d, in which case from a myopic point of view it would be best to play L. However, despite the large number of observations by player 1, there is still good reason to experiment with R since if 3 is playing u it would be quite lucrative for 1 to play R. This shows that the assumption of asymptotic myopia can be inconsistent with certain unrepresentative samples. Intuitively, though, such samples should have probability 0 in the long run; Fudenberg and Levine [1993] verify this in a closely related setting.

7.5. Steady State Learning

We consider the Bayesian model of learning in which players believe that they face a stationary distribution of opponents' strategies. As we noted above, one possible problem with this is that in the setting a fixed set of players playing a game the assumption is not true, and, particularly if the system does not converge, players may be able to discover this. An alternative explored by Fudenberg and Levine [1993] is to study a model of a population of randomly matched players, with players entering and leaving the population. Because players enter and leave (taking their knowledge with them) this model has a steady state in which the fraction of the population playing particular strategies does in fact remain fixed over time. The goal is to see what happens in this steady state when players are Bayesian optimizers and live for a long time.

The optimal amount of experimentation that takes place in the steady state is complicated. In practice some experiments are more revealing than others, and more patient individuals will be more inclined to experiment than less patient individuals. Moreover, the incentive to experiment depends on how lucky the individual has been with past experiments. What Fudenberg and Levine [1993] show is that when players have sufficiently long lives play resembles that of a self-confirming equilibrium, and if in addition, players are patient enough, it resembles Nash equilibrium.

Specifically, corresponding to each player (except nature) in the stage game is a population consisting of a continuum of players in the dynamic game. In each population, the total mass of players is one. There is a doubly infinite sequence of periods, ..,-1,0,1,.., and each individual player lives T periods. We denote the age of a player by τ. Every period $1/T$ new players enter the ith population, and we make the steady state assumption that there are $1/T$ players in each generation, with $1/T$ players of age T exiting each period.

Every period each player i is randomly and independently matched with one player from each population $i' \neq i$, with the probability of meeting a player i' of age τ equal to its population fraction $1/T$.[138] For example, if $T = 2$, each player is as likely to be matched with a "new" player as an "old" one. Each player i's opponents are drawn independently.

Over his lifetime, each player observes the terminal nodes that are reached in the games he has played, but does not observe the outcomes in games played by others. Thus, each player will observe a sequence of private histories h_τ^i.

The state of the system at a moment of time t is specified by specifying the fraction of the population with each possible type of history $\theta_t^i(h_\tau^i)$. Let ρ^i denote the optimal strategy for a Bayesian player with a non-doctrinaire prior and discount factor δ. For any state θ, it is also useful to define the actual number of people playing the strategy s^i

$$\overline{\theta}^i(s^i) \equiv \sum_{h_\tau^i | \rho^i(h_\tau^i) = s^i} \theta^i(h_\tau^i)$$

With this background , the deterministic dynamic in this state space can be described. We let $f^i(\theta)[h_\tau^i]$ denote the fraction of population i with private history h_τ^i at time $t+1$ when the state at time t was θ; the dynamic in other words is given by $\theta_{t+1} = f(\theta_t)$. New entrants to the population have no experience, so

$$f^i(\theta)[h_0^i] = 1/T .$$

Of the existing population $\theta^i(h_\tau^i)$ with a particular history, the fraction having experience $(h_\tau^i, \rho^i(h_\tau^i), z)$ is the fraction that met opponents playing strategies that led to the terminal node z. Let $\hat{s}^{-i}(s^i, z)$ be the pure strategies for i's opponents that lead to the outcome z.

[138]Boylan [1990] has shown that this deterministic system is the limit of a stochastic finite-population random matching model as the number of players goes to infinity.

$$f_i(\theta)[h_\tau^i, \rho^i(h_\tau^i), z] = \theta^i(h_\tau^i) \sum_{s^i \in \tilde{s}^{-i}(s^i, z)} \prod_{j \neq i} \overline{\theta}^j(s^j)$$

Finally, it is clear that

$$f_i(\theta)[h_\tau^i, \rho^i(h_\tau^i), z] = 0 \text{ if } s^i \neq \rho^i(h_\tau^i).$$

As always, we denote by $\hat{\theta}$ a steady state of the dynamical process *f*. Unlike in previous chapters, we will not examine the convergence of the dynamical process to the steady state, but only the steady state itself. (As in previous chapters, the *existence* of the steady state is not at issue, since the dynamical process is a continuous map from a compact state space to itself, the existence of steady states follows immediately from Brower's fixed point theorem.)

Observe first that steady states in this model need not bear any particular relation to equilibrium of any kind. If players live short lives, they have little opportunity to learn, and basically play against their priors. So the only interesting case is the limit as the length of life $T \to \infty$. Fudenberg and Levine [1993] prove two different results depending on whether or not players are patient: generally with long life steady states approximate heterogeneous self-confirming equilibrium; with patience they resemble Nash equilibria.

The intuition for why long life leads to self-confirming equilibrium has three parts. First, any strategy s^i that is played with positive probability in the steady state must be played by a positive fraction of the population a positive fraction of their life. Second, when the lifetime is large a player who plays a strategy a positive fraction of her life should, by a result of Diaconis and Freedman [1990], have approximately correct beliefs about its consequences. Third, the strategy s^i should maximize the current period expected payoff of most of the players who are playing it. That is, the bulk of players using s^i do so because it is myopically optimal, and not because they are

experimenting This final fact is subtle, because as we saw in the previous section, the optimal experimentation plan is relatively complicated. Combined, these facts imply that playing s^i is an approximate best response to beliefs that are approximately correct along the path of play, that is approximately a self-confirming equilibrium.

In showing that long-life plus patience leads to Nash equilibrium, the order of limits turns out to be quite important: the discount factor must go to one much more slowly than the length of life goes to infinity. It is not currently known whether the conclusion holds for the other order of limits. The intuition of the result is that patient players do enough experimentation to learn the true best responses to the steady state. Note that the fact that in the steady state players do not choose strategies based on calendar time means that the type of incidental correlation of experiments discussed in Fudenberg and Kreps [1995b] is not a problem; in effect in the steady state/random matching model the uniform non-experimental condition is satisfied, providing most players are not actually experimenting.

Note that the obvious argument is that if a player is very patient, and a strategy has some probability of being the best response, the player ought to try it and see. However, we already saw that some strategies may never be tried even though they do have a chance of being a best response, if there is some other strategy that provides the same information at lower cost. Instead, the argument actually used by Fudenberg and Levine [1993] uses the notion that an experiment has an option value: the option is to continue with the experiment if it works. If the steady-state distributions of strategies converge to a limit that is not a Nash equilibrium then there is a strategy being played with appreciable probability that is not optimal against the steady state. This implies that the option value for experimenting with this strategy cannot be converging to zero. On the

other hand, it can be shown that Bayesian optimal play and random matching imply that most option values become small.

7.6. Stochastic Adjustment and Backwards Induction in a Model of "Fast Learning"

This section discusses a model of "fast learning" in extensive games that Noldeke and Samuelson [1993] used to investigate the extent to which learning processes might tend to converge to refinements of Nash equilibrium. Fudenberg and Kreps [1988] identify several factors that suggest that results along these lines may require quite strong assumptions. First of all, beliefs must be correct at the larger class of "sequentially relevant" information sets, instead of the smaller class of relevant ones defined in Section 7.2, and this can require "more experimentation" than is required by MME. For example, in games of perfect information, all information sets are sequentially relevant, and so absent *a priori* restrictions on the payoff functions *every* information set must be reached infinitely often to ensure that only the backwards-induction solution is stable. As shown in the discussion of Figure 7.2, this in turn requires that players experiment even at information sets that are being reached a vanishing fraction of the time, and it is not obvious that even patient players would choose to do this.[139] Second, moving from subgame perfection to sequential equilibrium requires that players come to have common assessments about the relative probability of various nodes within an information set, even if the information set in question is reached with vanishing frequency.

---

[139] Which is not to say that we know that they would not. Indeed, a related and still open question is whether this much experimentation occurs in the steady-state learning model discussed earlier in this chapter in the limit of discount factors tending towards 1. On the other hand, if there is a minimum probability of experimentation as in the model of smooth fictitious play, then this assumption would be satisfied. The issues raised by this possibility are discussed in greater detail at the end of this chapter.

In Noldeke-Samuelson [1993] the Kandori-Mailath-Rob type of analysis is applied to games in which each player moves at most once on any path through the tree. In such games, a player's deviation from expected play cannot signal that he is likely to deviate at a subsequent information set, and so various sorts of refinements coincide. For example, trembling-hand perfection in the strategic form coincides with trembling-hand perfection in the agent-strategic form, and the notion of "rationalizability at reachable nodes" that we defined in Chapter 6, which does not restricts play at information sets that the player was expected to preclude, is equivalent to the stronger notion of sequential rationalizability, which requires "rational" play at every information set.

Noldeke and Samuelson considers anonymous random matching in a finite population with a steady inflow of "mutants" or "replacement players." The analysis will first determine the behavior of the system without these stochastic shocks, and then consider the system in which shocks are present but become vanishingly small. After we have done so, we will explain why the system involves much faster learning than in the models discussed earlier in this chapter.

7.6.1. The Model

Each agent in the model is described by a current strategy and a "conjecture" about the play of the opposing population(s). These conjectures take the form of a single behavior strategy for each population, and so implicitly impose the assumption of independent beliefs we discussed in chapter 6. Further, each agent's strategy is presumed to be a best response to his current conjecture, where agent's goal is to maximize his *ex-ante* expected payoff given his conjecture. In particular, an agent' strategy is allowed to prescribe conditionally dominated actions at information sets that the player's conjecture assigns probability 0. Each period, all agents are randomly

matched to play the game. In particular, the probability that a given agent of player i is matched with a given agent of player j is some fixed number bounded away from 0.

At the end of period, each agent has probability μ of "learning." A learning agent observes the terminal nodes in *every* match this period, and resets his beliefs at the corresponding information sets to equal this period's observation. The agent then adjusts his strategy so that it prescribes a best response to his conjectures at all of his information sets, with the "inertia" assumption that if the agent does not change his actions at information sets where that action is one (possibly of several) best responses to the new conjecture.

This process of belief revision and strategy adjustment, in which players use only their most recent observation and ignore all previous ones, parallels that in the Kandori-Mailath-Rob papers. A new feature here is the assumption that observing play at the information sets that were reached this period has no effect on beliefs about play at the unreached information sets. [140] From a Bayesian perspective, this amounts to supposing that beliefs take the form of a product of independent distribution over play at each information set, so that seeing player 2 shift his response to a given action does not signal that 2 may have changed his response to others; This is a stronger assumption than the independence across *players* that is implicit in the formulation of conjectures as strategy profiles. Note that all agents who learn end the period with the same (and correct) on-path beliefs. Note also that if the agent does not get to "learn," he does not change his beliefs

[140] In this setting it is somewhat trickier to justify the decision rules as being approximately optimal when the system changes only slowly and the agent has a small discount factor, since the most recent observed *outcomes* need not be a sufficient statistic for the entire history of outcomes. However, this is taken care of by the combination of the assumption that learning players observe the outcome in all matches and the independence assumption.. Note moreover that this sort of memoryless learning makes it very hard *a priori* for a mixed- strategy equilibrium to be stable. This did not matter very much in the 2x2 coordination games considered by Kandori, Mailath and Rob, where the mixed equilibrium would clearly be unstable in any sensible dynamic, but it becomes an issue when considering more general extensive-form games.

even if they are inconsistent with the terminal node reached in his own match this period. This strikes us as an odd aspect of the model, but it does not seem important for the results.[141]

The previous paragraph defines the "no mutations" adjustment process $\Gamma(0)$. The state space of this process is the set Θ whose elements specify a strategy and conjecture for each individual agent. To extend this to a process with mutations or replacements, suppose that following an i.i.d. process, each period with probability λ each agent is replaced by another one with an arbitrary conjecture and a strategy that is a best response to the conjecture. These mutations create an ergodic system, denoted $\Gamma(\lambda)$; Samuelson and Noldecke's goal is to characterize the limit of its ergodic distribution μ^{λ} as $\lambda \to 0$.

Two aspects of this system that deserve special emphasis. Note first that the set of mutations or perturbations is somewhat smaller than that considered in Kandori, Mailath, and Rob, since mutants never adopt strictly dominated strategies. For this reason, the transition matrix of $\Gamma(\lambda)$ is not strictly positive, but since all undominated strategies have strictly positive probability it is easy to see that the system is indeed ergodic.[142]

Second, the mutations will be a source of "experiments" with off-path actions. Moreover, since the probability of the event "all agents learn" will in the limit $\lambda \to 0$ be infinitely larger than that of a mutation, the model will generate much more information about off-path play than if each agent only observed the outcomes of their own matches. Consequently, we should expect that "less" experimentation is required to rule out non-Nash outcomes in this model than under the usual observation structure. This effect is

---

[141] In a private communication, Larry Samuelson has recently sketched an argument that all of the asymptotic results of the paper are unchanged if each agent learns the outcome in his own match in every period, provided that agents still set their conjectures about play at each information set equal to their most recent observation there. However, that assumption is less attractive when agents only observe their own matches than if they observe all outcomes, since the agent is trying to learn the aggregate distribution of opponents' play, and he will typically play a different opponent each period.

[142] See the appendix to chapter 5.

strengthened by the assumption that when players learn they revise their conjecture to correspond to their most recent observation, so that a single experiment here can have as much force as an infinite number of them in the model of fictitious play. Indeed, we shall see that the key events to consider in determining the long-run distribution are "a single player *i* experiments, and then all player's revise their conjectures to match the outcome of the experiment before any other players change their actions." For this reason we should expect that convergence to a non-Nash outcome will be less common in this model than in those we discussed earlier in the chapter. This is also why we call the model one of (relatively) "fast learning."

7.6.2. The Deterministic Dynamic

As usual, the method is to work out what happens without mutations first. In this case, the outcome generated by any singleton limit set (steady state) must be the outcome of an independent and unitary self-confirming equilibrium. To see this, note that since each player has some chance of eventually learning, and a player who learns observes play in *all* matches, if play is absorbed by a single outcome all players will eventually learn what that outcome is, and so all players must have correct conjectures at *all* information sets on the corresponding path. Thus the outcome must correspond to a unitary self-confirming equilibrium; the independence is imposed by assumption as we noted above. Conversely, any self confirming equilibrium corresponds to a singleton limit set.

Note that a given self-confirming *outcome* can correspond to many different steady states, since actual play at unreached information sets is arbitrary, and there are only weak restrictions on the conjectures about this off-path play. In particular, if in a steady state θ player *i* could deviate and send play to an unreached subgame, and no other

player's deviation can cause this subgame to be reached, then any different state θ' that differs from θ only in the conjectures of players other than *i* about play in the subgame also self-confirming, and consequently also a steady state. Moreover, there can be steady states in which different agents of a given player, say player 1, disagree about precisely which awful payoff they would get if they gave player 2 the move, so long as in the steady state player 2's information set is never actually reached. Thus, even though the outcome of the steady state must be a unitary self-confirming equilibrium, that outcome can also correspond to a steady state without unitary beliefs.

Due to this huge multiplicity of steady states, the brute-force approach of enumerating all of the steady states of the unperturbed system and then computing minimal order trees is likely to be quite tedious. However, such calculations are not needed, since, as shown below, the large number of steady states makes it so easy for mutations to switch play from one steady state to another that we need only consider transitions that can be caused by a single mutation.

7.6.3. Dynamic with Mutations

We turn now to the case with mutations, so that $\lambda > 0$. We will say that a state is *stochastically stable* if it is contained in the limit of the supports of the ergodic distributions μ^{λ} as $\lambda \to 0$.

Proposition 7.5 (Noldeke-Samuelson): If state θ is stochastically stable so is any other steady state θ' whose basin of attraction (in $\Gamma(0)$) can be reached with a single mutation.

Intuitively, if a single mutation suffices to jump away from θ, the expected time spent in state θ is of order $1/\lambda$, and since θ' is a steady state, it will take at least 1 mutation before this state is left, so that the expected time spent in θ' is at least as large as that spent in θ.

Using this lemma about stable states, Noldecke and Samuelson develop a necessary condition for there to be a stochastically stable outcome; that is, for the limit distribution to be concentrated on states all of which induce the same distribution over terminal nodes. From our remarks above, we see that in order for there to be a stable outcome, there must be a corresponding set of states all of which lead to that outcome, and such that no single mutation leads to a state with a different outcome.

Proposition 7.6 (Noldeke-Samuelson): Consider an extensive-form game in which each player moves at most once on any path of play. Suppose that an outcome is stochastically stable, and that at some stochastically stable state with that outcome player *i* can deviate and send play to some subgame. Then no self-confirming equilibrium of the subgame can give player *i* a higher payoff he received in the stochastically stable outcome.

Sketch of proof: Let *z* be a stochastically stable outcome generated by the stochastically stable set Θ^*. The first step is to check that every state in Θ^* is a steady state and hence self-confirming. (The idea is that non-singleton limit sets of $\Gamma(0)$ must contain states with at least 2 different outcomes.) Suppose that at e outcome z ,there is a player *i* who can take an action *a* that sends play to a subgame *G(a)* that has a self-confirming equilibrium σ that gives the player more than he's getting in *z*. Fix a stochastically stable state θ', and consider the state θ where all players' strategies and conjectures agree with θ' at all information sets outside of *G(a)*, in which player *i* has the same strategy and conjecture as in θ', and such that strategies and conjectures of all players who have an information set in *G(a)* correspond to σ. Since θ' corresponds to an self-confirming equilibrium, so does θ.

Now consider a mutation that makes one agent of player *i* play into this subgame, and then suppose that all player *i'* s learn before any agent of any other player type, and before any further mutations. This sends the system to a new state, whose outcome is

some *z'* that is different than the outcome *z* we started with. Moreover, since play in *G(a)* is a subgame self-confirming equilibrium, the learning mechanism cannot further adjust actions or conjectures in this subgame. Since player *i*'s payoff in this subgame is greater than it had been under the initial outcome *z*, and since at *z* player *i* can force play into the subgame, the learning process starting at *z'* cannot lead back to *z*. Since a single mutation suffices to send the system away from *z*, and at least one mutation will be required to return to *z*, *z* cannot be the unique outcome in the support of the ergodic distribution of the perturbed system.

☑

Corollary 7.1: In a multistage game with observed actions in which each player moves at most once on any path of play, any stochastically stable outcome must be a subgame-perfect equilibrium.[143]

Proof: From Proposition 6.4, in multistage games, every unitary self-confirming equilibrium with independent beliefs has the same outcome as a Nash equilibrium. Thus Proposition 7.6 and the fact that every stochastically stable outcome is self-confirming implies that a stochastically stable implies that a stochastically stable outcome must be a Nash equilibrium outcome with the additional property that no player can deviate and send play to a subgame in where that player gets a higher payoff in some self-confirming

[143] Noldeke and Samuelson assert that this conclusion follows without the restriction to multistage games, but as Larry Samuelson has pointed out to us, their proof is incorrect. However, no couterexample has been found, and it remains an open question whether this restriction is really needed. To see why it might not be, note that while the inconsistent self-confirming equilibrium (A_1, A_2, L_3) in figure 6.1 is a steady state of the unperturbed learning process, it is not locally stable: a single mutation by a player 1 onto D_1 sends the unperturbed dynamic to the Nash equilibrium (D_1, A_2, A_3). This raises the yet-unproved conjecture that all locally stable outcomes must correspond to Nash equilibria

equilibrium. The conclusion then follows from the fact that every subgame-perfect equilibrium of any subgame is self-confirming.

☑

Three aspects of these results deserve emphasis. First, on a technical level, the proof is greatly simplified by the fact that a single mutation suffices to leave the basins of many steady states. Noldeke and Samuelson use the same proof technique in a subsequent paper on learning dynamics in a "screening" model. The reason that this technique is useful in these papers is the assumption that in the "no-noise" learning process, a player who learns observes play in all matches. Thus, the key event in both models is "a single mutation onto a previously unused action, followed by *all* agents of a given player learning." The nature of the learning process means that the single mutation onto a previously unplayed action can have dramatic consequences.

As of this writing, the technical argument has not been extended to other types of learning processes. However, the more second and more general point made by these papers is that dynamics in extensive form games should be expected to be more sensitive to various forms of noise and perturbation than are dynamics in static games with strict equilibria, and we expect that point to hold quite generally.

Third, and relatedly, the sensitivity to perturbations suggests that many games will not have a stochastically stable outcome. This can be seen in the strength of Proposition 6.5, and is illustrated in the following example, which is a 3 player "centipede" game in which each player in succession chooses between G (Go) and S (Stop); if any player chooses S the game ends, and in any case the game ends after player 3's move.

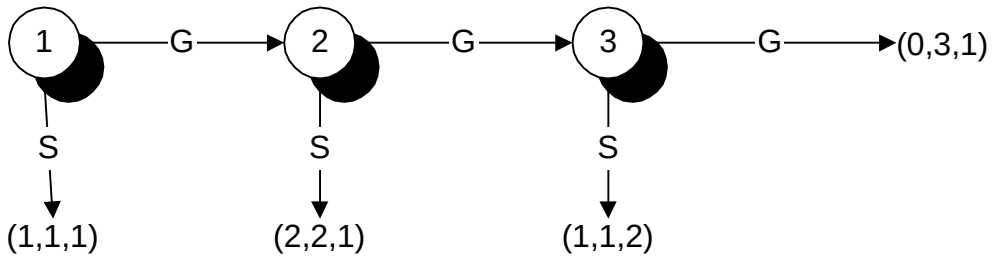


Figure 7.4

The unique subgame perfect equilibrium is (*G*,*S*,*S*); with outcome (*G*,*S*), profile (*G*,*S*,*G*) has the same outcome. This would also need to be in the stochastically stable set if (*G*,*S*,*S*) is. But, at the state corresponding to (*G*,*S*,*G*), where all player 3's play *G*, if all player 2's learn simultaneously, while none of the player 1's do, the state switches to (*G*,*G*,*G*). Suppose that in the following period, all player 1's, learn, and no other players do. This sequence of events, which relies only on the "learn" draws, has positive probability under the unperturbed, no-mutations dynamics, and leads to the steady state (*S*,*G*,*G*) with outcome *S*.

In contrast to this example, suppose the subgame perfect equilibrium gives all players a higher payoff than any other outcome. In this case it is stochastically stable. Noldeke and Samuelson prove a slightly stronger theorem. Consider the outcome of a subgame perfect equilibrium in a game of perfect information. This outcome is the unique stochastically stable outcome if no player has an action that can send play to a subgame in which some terminal node gives that player a higher payoff than he received in equilibrium.

Because it is so unlikely to be satisfied the notion of a single point being stochastically stable is not that useful. One conclusion we can draw from this is to accept

the idea that limit sets may fail to be single points, and conclude that standard refinements are too strong.

7.7. Mutations and Fast Learning in Models of Cheap Talk

This section discusses the application of fast-learning-with-mutations to the cheap-talk 2x2 coordination game we discussed in chapter 3. Recall the structure of the game: There are two stages of play. In the first stage players simultaneously announce messages, which will be treated as signals of their "intended action" L or R; in the second stage they play the coordination game with payoffs

	L	R
L	2,2	-100,0
R	0,-100	1,1

Talk is cheap in the sense that announcing an action has no direct effect at all on the realized payoffs, which depend only on the second-stage choices.

In chapter 3 we observed that ESS is sufficient to eliminate the (R,R) equilibrium provided that there is no "babbling", that is, some message is not sent in equilibrium. By way of contrast, we discussed an argument due to Matsui [1991] based on cyclically stable sets that eliminates the (R,R) equilibrium even with a fixed finite message space. As we noted, Matsui's argument implicitly supposes that players observe and respond to the strategy profile actually used by their opponent, including the parts of the profile that relate to off-path play. Once we recognize the extensive-form nature of a cheap talk game, this is no longer a satisfactory assumption.

In this section we will sketch an argument similar to that of Noldecke and Samuelson that, unlike Matsui , supposes that players can only observe what actually happened in the course of play. Moreover, we will see that the conclusions can depend on whether agents observe only the outcomes in their own matches, or instead can observe the outcomes in all of them, as in Noldeke-Samuelson. Another advantage of these arguments is that they each concern long-run behavior under a single dynamic, and thus sidestep some of the interpretational questions posed by the Gilboa and Matsui [1991] definition of a cyclically stable set.

We begin with an extension of the Noldeke-Samuelson model of the previous section. Since each player moves twice along every path of play, instead of only once as assumed by Noldeke-Samuelson, their model is not immediately applicable to this game. To apply it we will extend their independence assumption to by adding the condition that play by a given player at a given information set is treated as having no information about that player's actions at any other information sets, even those which are successors of the information set in question. With this extension we can show that the limit of the long-run distributions assigns probability 1 to all players choosing the Pareto-optimal action L.

Here is a sketch of the argument, which we have not seen given elsewhere. We will argue first that, starting from any state in which all agents play R in the second stage given the prevailing distribution of messages, the system can move to a state where all agents play L at a "cost" of only 3 mutations. That is, the component in which all agents play R has a modified coradius (see Chapter 5) of at most 3. To see this, let θ^{*R} be the state in which agents believe that all opponents will play R regardless of the first-period message, and all agents choose to both say and play R. If the current outcome is that all agents actually play R, given the prevailing distribution of messages, then all agents must believe that regardless of the message they send their opponent is likely to

play R. Consequently, the state can move to θ^{*R} by a series of single mutations: take each player who is currently saying *L*, and replace his conjectures by those of θ^{*R}. Since these conjectures are consistent with the observed distribution, each such single mutation leads to a new self-confirming equilibrium, and so to a new steady state of the unperturbed adjustment process. Thus the modified cost of this sequence of mutations is only 1, Next, from the state θ^{*R} two mutations are sufficient to shift the state to the basin of the equilibrium component in which all agents play L: suppose that a single agent on each side mutates to the conjecture "all of my opponents will say play L if I said L and will play R if I say R."[144] Suppose moreover that these two mutants are immediately matched with one another, so that they both end up saying L and then playing L, and that this is followed by the event "all agents learn." (Recall that both of these events have positive probability in the unperturbed (no-mutations) dynamic.) Then with the assumption that beliefs are updated separately at each information set, the learning players have the new conjectures "everyone plays L if both messages were L, and plays R otherwise," and so all agents say and do L in the next period. Hence the modified coradius of "all play R" is at most 3.

However, the modifed coradius of the component where all agents play L is proportional to the number of players, and so is much larger if the population is large. In order for "learning" players to start choosing a message that leads to a significant chance of playing R, it must be that *both* messages have a substantial probability of leading to the opponent playing R. (Otherwise, the learning player would choose the message the made it likely his opponent will play L.) This is the key asymmetry between the strategies: plays shifts from R to L if *either* message is likely to result in (L,L), which can occur after

[144] The conjectures about play following "mismatched" messages are unimportant, as will be clear from the argument that follows.

only a single mutation onto an unsent message; while to induce players to choose R, some fraction of the population must mutate.

Note well that this asymmetry depends on the assumption that the event "all players learn at once" is much (infinitely) more likely than any single mutation. This can be seen by considering modifying the model so that agents learn the distribution of outcomes induced by their own chosen strategy, but otherwise keeping the model the same. [145] Here two mutations are not enough to move from θ^{*R} to the basin of the component where all agents play L, and agents will only choose to shift to sending L if a substantial fraction of agents mutates at the same time.

7.8. Experimentation and The Length of the Horizon

Basically the results of this chapter show that we should expect Nash rather than self-confirming equilibrium if there is enough experimentation. On the other hand, we argued, especially in chapter 4, that there is good reason for players to use a rule such as smooth fictitious play which is random, and that there are many reasons to believe that players actually do randomize, including the random utility model of Harsanyi and the empirical research of psychologists. This raises the question of whether and why self-confirming equilibrium should be of interest.

To answer this question, we adopt Binmore and Samuelson's [1993] typology of the short, medium, long and ultra-long run. The short run is a period so short that players have no opportunity to learn, and simply use their priors. In an experimental setting, Stahl and Wilson [1994] have explored models of prior formation that give some

[145] If players only observe their own matches, the assumption that conjectures equal their most recent observation is not very sensible. For example, in the coordination game without cheap talk, the no-mutation process holds constant the numbers of agents playing L and playing R, so that every state is in the support of the limit distribution!

predictive power of play in the first round of an experiment, while falling far short of equilibrium of any kind. Nagel [1993] has examined how the transition from short to medium-run takes place as people begin to best-respond to past play by opponents.

In the medium run, players have an opportunity to learn. In Binmore and Samuelson [1995], the medium run is identified with remaining for a long time near a component of steady states that is not dynamically stable, with the long run being long enough for the system to move away from unstable components and arrive at a stable one. We would prefer to emphasize instead that while experimentation may lead to Nash equilibrium in the long run, self-confirming equilibrium may be a good description of a system in the medium run. The point is that players accumulate data about the on-the-path play of their opponents much more rapidly than they generate data about their off-path play. Consequently, we expect in the medium run that outcomes that are not self-confirming are unstable, but it only makes sense to believe that players will move away from self-confirming equilibrium to full Nash equilibrium (or one of its refinements) over a much longer horizon , so that experimentation will have yielded a substantial amount of data.[146] Unfortunately, we do not know of a way to make a formal distinction between these two horizon lengths.

Finally, in the very long run, we might expect the type of Kandori, Mailath, Rob and Young type of argument to become relevant, as the system spends most of its time near the particular steady state that is stochastically stable. As pointed out by Binmore, Samuelson, and Vaughn [1995], this distinction corresponds to two different orders of limits in evaluating the average behavior of the system: the case where mutations are very rare over the relevant horizon corresponds to the order $\lim_{t\to\infty}\lim_{\varepsilon\to 0}$, where ε is a

[146] This same point has been emphasized by Er'ev and Roth [1994].

measure of the size of the "noise," while the ultra-long run is long enough for the noise to be nonnegligible, so that the appropriate order of limits is $\lim_{\varepsilon\to 0}\lim_{t\to\infty}$.

In interpreting these results, it is important to think of the particular application. For example in studying what happens in an experimental setting, the short- and medium-run seem most relevant, except for a few experiments running very simple games over 50 or more trials, where the long run may be relevant. In talking about rules of thumb, social norms, or customs in an economic setting, we imagine that these have evolved over a very long period of time, and so the long or ultra-long run cases have greater relevance; this is the point of view taken by Young [1993].

Appendix: Review of Bandit Problems

In a multi-armed bandit problem, a single player with discount factor δ must choose from a finite set of "arms" or actions $a \in A$. The action chosen gives rise to a probability distribution over outcomes $\theta \in \Theta$; these outcomes are independently drawn each period from probability distributions $\sigma(a) \in \Delta(\Theta)$, which are unknown. Utility depends on the action and the outcome $u(a,\theta)$. Players beliefs are given by prior distributions $\mu(a)$ over each probability distributions $\sigma(a) \in \Delta(\Theta)$. These priors are independent between actions so that learning the distribution corresponding to one action conveys no information about the distributions corresponding to others; the only way to learn about the distribution generated by a given action a is to play that action. After an action a is chosen in period t, beliefs $\mu_t(a)$ for the corresponding action are updated according to Bayes law. For example, if $\mu_t(a)$ is a continuous density, and a is chosen at time t, with outcome θ then

$$\mu_{t+1}(a)[\sigma] = \frac{\sigma(\theta)\cdot\mu_t(a)[\sigma]}{\int \sigma(\theta)\cdot\mu_t(a)[\sigma']d\sigma'}.$$

Because of the assumed independence, beliefs for actions not chosen are not updated at all.

Our discussion in this Appendix follows Ross [1983]. The problem is ordinarily analyzed by dynamic programming methods; that is, by postulating a value function $v(\mu)$, and using the Bellman relation

$$v(\mu) = \max_{a\in A}(1-\delta)E_\mu u(a,\theta) + \delta E_{\mu',a} v(\mu')$$

which says that the value of particular beliefs are equal to the greatest amount that can be earned by choosing an action that maximizes current expected utility plus the expected value of next period's beliefs. It may easily be shown that v is (weakly) convex in μ.

In the case of a multi-armed bandit problem the solution of the dynamic programming problem was shown by Gittens [1979] to have a particularly simply form. Consider first the simple one-armed bandit problem where the option is to play the one arm, or to drop out and receive a fixed utility U. Let $v(\mu,U)$ be the value function corresponding to this problem. This function may easily be shown to be continuous and non-decreasing in *U*. Let, moreover, $U(\mu)$ be the smallest value of U for which it is optimal to drop out. This is called the Gittens index. Then

$$U(\mu) = \min[U | v(\mu,U) = U]$$

The key fact in solving the multi-armed bandit problem is that the Gittens index for each individual arm can be used to determine which arm to pull. That is, the optimal plan is to compute for the current beliefs, and for each action a, the index $U_a(\mu(a))$, and use the action for which this Gittens index is largest. Of course this result depends very heavily on the assumption that the arms are independent, which means that the Gittens index is not very useful for analyzing extensive form games.

A basic feature of the multi-armed bandit is that there is positive probability of stopping forever on the wrong arm. This is easy to see: Suppose there are two arms, one with a favorable prior and one with an unfavorable prior. Suppose moreover, that the actual draw of σ on the first arm is more favorable than expected, and that the actual draw on the second arm is more favorable than the first arm. Then the first arm will be tried first, and since the only surprise will be that the first arm is more favorable than expected, the second arm will never be tried. That is, as long as the Gittens index on the first arm does not drop below the prior Gittens index on the second arm, the second arm will never be used. But in this example, the second arm is actually more favorable.

On the other hand, we would not expect this phenomenon to be important if the player is patient, that is, if δ is near one. To see that it is not, observe that the strategy of

experimenting with all arms for T periods, then switching to the most favorable arm forever, will, if δ is near one, and T is sufficiently long, give nearly the same expected present value as knowing in advance which is the best arm. Consequently, the optimal policy must also yield approximately the full-information payoff. . This means that for δ near one there cannot be an appreciable chance of stopping on the wrong arm by mistake.

References

Binmore, K. and L. Samuelson [1993]: "Muddling Through: Noisy Equilibrium Selection," University College London.

Binmore, K. and L. Samuelson [1995]: "Evolutionary Drift and Equilibrium Selection," University College London.

Binmore, K., L. Samuelson and K. Vaughn [1995]: "Musical Chairs: Modelling Noisy Evolution," *Games and Economic Behavior*, 11: 1-35.

Boylan, R. [1990]: "Continuous Approximation of Dynamical Systems with Randomly Matched Individuals,," Washington University of St. Louis.

Diaconis, P. and D. Freedman [1990]: "On the Uniform Consistency of Bayes Estimates with Multinomial Probabilities," *The Annals of Statistics*, 18: 1317-1327.

Er'ev, I. and A. Roth [1996]: "On The Need for Low Rationality Cognitive Game Theory: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria," University of Pittsburgh.

Fudenberg, D. and D. K. Levine [1993]: "Steady State Learning and Nash Equilibrium,," *Econometrica*, 61 (May): 547-573.

Fudenberg, D. and D. M. Kreps [1988]: "Learning, Experimentation and Equilibrium in Games," Stanford.

Fudenberg, D. and D. M. Kreps [1995a]: "Learning in Extensive Games, II: Experimentation and Nash Equilibrium," Harvard.

Fudenberg, D. and D. M. Kreps [1995b]: "Learning in Extensive Games, I: Self-Confirming Equilibrium," *Games and Economic Behavior.*

Gilboa, I. and A. Matsui [1991]: "Social Stability and Equilibrium," *Econometrica*, 58: 859-867.

Gittens, J. [1979]: "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society*, Series B: 14, 148-177.

Kohlberg, E. and J. Mertens [1986]: "On the Strategic Stability of Equilibria," *Econometrica*, 54: 1003-1038.

Matsui, A. [1991]: "Cheap-Talk and Cooperation in a Society," *Journal of Economic Theory*, 54: 245-258.

Myerson, R. [1978]: "Refinment of the Nash Equilibrium Concept," *International Journal of Game Theory*, 7: 73-80.

Nagel, R. [1994]: "Experimental Results on Interactive Competitive Guessing," D.P. B-236, Universität Bonn.

Noldeke, G. and L. Samuelson [1993]: "An Evolutionary Analysis of Forward and Backward Induction," *Games and Economic Behavior*, 5: 425-454.

Ross, S. M. [1983]: *Introduction to Stochastic Dynamic Programming*, (New York: Academic Press).

Stahl, D. O. and P. W. Wilson [1994]: "On Players' Models of Other Players: Theory and Experimental Evidence," *Journal of Economic Behavior and Organization*, 25: 309-327.

Young, P. [1993]: "The Evolution of Conventions," *Econometrica*, 61: 57-83.

8. Sophisticated Learning

8.1. Introduction

Throughout our discussion of both evolutionary and learning models, we have emphasized our belief that people are relatively good at learning. So far, however, one deficiency of all the models we have examined is that they are unable to detect simple cycles or other patterns in the data. Either implicitly or explicitly, rules such as fictitious play and its variations, best-response, and stimulus-response are designed to perform well against an i.i.d opponent; none of them attempt to detect cycles or other regularities in the data. In this chapter, we examine learning rules that are more "sophisticated" in the sense that they explicitly attempt to detect patterns.

We consider three ways of modeling the idea that players may attempt to detect patterns in opponents' play. The most traditional starts from a set of opponent strategies, possibly involving complex patterns of play over time, that are *a priori* viewed as possible. Supposing that players have prior beliefs over these strategies leads to a Bayesian model of learning. Such a Bayesian model is equivalent to specifying conditional probabilities over opponents' play conditional on particular events. A related approach is to specify the events that are to be conditioned on, and estimate the conditional probabilities directly. For example, rather than assuming that the probability distributions governing opponents' play are independent of the history, as in fictitious play, we can allow players to believe that the distributions are different in odd and even periods, or that they depend on the actions played in the last period. In this approach, the primitive specification is not a set of opponent strategies that are viewed as possible, but

rather a method of classifying histories into categories within which conditional probabilities are assumed to be constant.

A third approach is to treat as primitive, not the opponents' strategies that are viewed as possible, but the set of the player's own strategies that are viewed as potential best responses. In the computer science literature, these are referred to as *experts* and the goal may be thought of as choosing the expert who makes the best recommendations for a course of play.

We begin in section 8.3 with the Bayesian learning model. Following Kalai and Lehrer [1993], we show that if these beliefs contain a "grain-of-truth" (or a weaker absolute continuity assumption) then play must converge to a Nash equilibrium. However, the "grain-of-truth" assumption is difficult to justify because when opponents act as Bayesian learners, they may well follow strategies that were not *a priori* viewed as possible, in which case the grain of truth assumption will fail. In section 8.4 we explore some of the difficulties with the "grain-of-truth" assumption, including Nachbar's [1995] result providing conditions that make the grain-of-truth impossible. Our conclusion is that it is important to use learning procedures that are robust in the sense that, unlike Bayesian learning, they continue to perform well, even if none of the alternatives viewed as possible actually turn out to be true.

The remainder of the chapter looks at learning procedures that are robust and that attempt to detect at least some patterns. . In section 8.5 we examine procedures proposed in the computer science literature that make it possible to do asymptotically as well as the best expert, even in the worst case. In section 8.6, we show how to extend the fictitious play idea, that players learn about frequencies, to the idea that they learn about conditional frequencies. The main result is that the conditional analog of smooth

fictitious play does about as well asymptotically as if the conditional frequencies are known in advance.

In general we can say little about the relative performance of cautious fictitious play against alternatives such as best response or stimulus response. Despite the fact that cautious fictitious play has better theoretical properties, it may be outperformed by the other models in particular instances. In section 8.7 we show that this is due to the fact that the other procedures may inadvertently be conditioning on parts of the history that are ignored by cautious fictitious play. In fact, we show that given any other procedure we can design a particular conditional cautious fictitious play learning rule by conditioning on the same information used by the procedure in question that regardless of whether the criterion is time average or discounted payoffs, and regardless of the discount factor, never does much worse than the other procedure, and sometimes does considerably better.

In section 8.8 we ask whether sophisticated learning is potentially destabilizing. Sonsino [1994] argues, for example, that sophisticated procedures may result in cycles, where less sophisticated procedures would have converged. In section 8.9 we examine the converse question of whether sophisticated learning procedures can lead to convergence in cases where less sophisticated procedures cycle. Even if cycles can be successfully detected, is it possible that this generates even more complicated cycles that players are unable to detect, or is the system forced to convergence? This depends in an important way on players synchronizing the data they use from the past. Allowing the possibility that patterns more complicated than those contemplated in player's behavior rules may be generated, we observe one implication of this may be that opponent's play may appear to be correlated with the player's own play. Put differently, a player's own choice may contain predictive power about what his opponents are going to do. In

section 8.10, we examine the procedure developed by Foster and Vohra [1995] to take account of this extra information. If players do so, play in the long-run must come to resemble a correlated equilibrium. However, there is still the theoretical possibility that the correlation, which is generated by a time dependence in behavior rules, is still too complicated for players to anticipate; whether this is likely to happen in practice is currently unknown.

One important issue is that once we allow players to condition on histories of play, they may realize that opponents' future play depends on their own current play. If players are not myopic, this raises an important set of issues: first, players may try to manipulate their opponents learning process. This possibility is discussed in section 8.11. Second, players may not be able to correctly infer the causal connection between their own play and their opponents. This possibility was explored in previous chapters examining extensive form games. Finally, even if there is enough experimentation to reveal opponents strategies, it is currently unknown whether or not there are analogs of conditional cautious fictitious play and other robust methods that are applicable in a non-myopic setting.

8.2. Three Paradigms for Conditional Learning

Through most of this chapter we return to the simple setting of a fixed set of myopic players playing a static simultaneous-move game; as usual this should be thought of as a convenient simplification of a model of a larger population. In the stage game the strategies are $s^i \in S^i$, and utilities u^i. This game is repeated over time, and a finite history of play $h_t = (s_1, s_2, \ldots, s_t)$ is a list of how all players have played through period *t*. The play of all players except for player *i* is denoted by h_t^{-i}, an infinite history of play by *h*, and so forth. The set of all finite histories continues to be denoted by *H*.

We examine three behavioral paradigms for detecting patterns. The most traditional is a set of opponents' strategies, possibly involving complex patterns of play over time, that are *a priori* viewed as *plausible*. Adding a prior over these strategies will lead to a Bayesian model of learning. Specifically, let a *model* for player i be a map $m^{-i}(h_t) \in \Delta(S^{-i})$ from histories to correlated strategies in the repeated game for player i's opponents. Bayesian beliefs are then specified by a set of *plausible* models M^i, and a prior over this plausible set.

Instead of focusing on models that are thought to be plausible, we can focus on strategies that are thought to be potential best responses. In the computer science literature, these are referred to as *experts* and the goal may be thought of as choosing the expert who makes the best recommendations for a course of play. Specifically, we define an *expert* for player i to be a map defined on histories $e^i(h_t) \in \Delta(S^i)$; this is simply a strategy in the repeated game.

Given any model, we can consider the corresponding experts which are best responses to the model; given any expert, we can consider the models for which the expert would be a best response. While this is not a one-to-one correspondence between models and experts, nevertheless, there is a close link between the two. In particular we can think of a player's beliefs about the possible dynamics of his opponents play either in the form of a set M^i of *plausible* models, or as a set E^i of *plausible* experts.

Finally, Bayesian priors over models give rise to a set of conditional probabilities of opponents play given the history of all players' play. The essential element is the partitioning of histories into disjoint sets of events. Specifically, we suppose that there is a collection of categories Ψ^i into which observations may be placed. A *classification rule* is a map $\hat{\psi}^i : H \times S^i \to \Psi^i$.[147] The interpretation is that prior to observing s_t^{-i} the

[147] For notational simplicity we limit attention to deterministic classification rules.

player knows h_{t-1}, s_t^i, and must choose a category $\hat{\psi}^i(h_{t-1}, s_t^i)$ based only on this information. Instead of focusing on models, or on experts, we can instead attempt directly to estimate the probability of opponents' play conditional on particular categories.

8.3. The Bayesian Approach to Sophisticated Learning

We begin by considering a learning rule ρ^i generated by playing a best response to a particular prior Γ^i that puts positive probability on each of a countable or finite set of *plausible* models M^i.[148] Our presentation is generally based upon the work of Kalai and Lehrer [1993], although they consider the more general case where players can be either patient or impatient [149]

Given behavior rules ρ^i for each player, there is a well-defined probability distribution $D(\rho)$ over histories, finite and infinite. If there exists a map from histories to Nash equilibria of the stage game $\hat{\sigma}_t(h_{t-1})$ such that $|D(\rho)(s_t|h_{t-1}) - \hat{\sigma}_t(h_{t-1})(s_t)| \to 0$ almost surely with respect to $D(\rho)$, we say that the rules *converge to Nash equilibrium.* Note that our criterion for convergence is the strong one that requires convergence in every period; however, we do not require convergence to a single Nash equilibrium: we allow deterministic movement between several Nash equilibria. Since players are no longer assumed to believe that the world is stationary, there is no reason that they cannot, for example, engage in a cycle between different Nash equilibria.

One desirable feature of Bayesian learning is that if priors are consistent with the true model, beliefs are consistent; that is, they converge to the true model. In the current

[148] Obviously there is no loss of generality in assuming positive weights; the set of plausible models is just the support of the probability distribution. We will comment on the case where the set of models is uncountably infinite below.
[149] We discuss the issue of patience below.

context, this will imply convergence to Nash equilibrium. To be more precise, let $D^i(\rho^i, \Gamma^i)$ be the probability distribution over histories induced by player *i*'s beliefs and the player's own behavioral rule.

Definition 8.1: Beliefs Γ^i are *absolutely continuous with respect to the play path* if there exists some plausible model $m^i \in M^i$ such that $D^i(\rho^i, m^i) = D(\rho)$.

In other words, there should be some plausible model that is observationally equivalent to opponents' actual strategies in the sense that the probability distribution over histories is the same. However, the plausible model may generate different off-path play. In the current context of myopia, this is irrelevant, since players do not care about how their deviation may effect opponents future play; in the context studied by Kalai and Lehrer of patient players, this means that we may have only a self-confirming equilibrium, rather than a Nash equilibrium, since beliefs need be (asymptotically) correct only on the equilibrium path.

Proposition 8.1 [Kalai and Lehrer]: If ρ^i are best responses to beliefs Γ^i that are absolutely continuous with respect to the play path, then they converge to Nash equilibrium.

Proof: This is essentially a result of Blackwell and Dubins [1962]. We will give a sketch of how it follows from the Martingale convergence theorem. Let ω denote the event that m^i and $\tilde{\omega}$ denote the event that it does not occurs. We may write the posterior odds ratio of the event $\tilde{\omega}$ as perceived by player *i*

$$L_t = \frac{D^i(\rho^i,\Gamma^i)(\tilde{\omega}|h_t)}{D^i(\rho^i,\Gamma^i)(\omega|h_t)} = \frac{D^i(\rho^i,\Gamma^i)(s_t|\tilde{\omega},h_{t-1})D^i(\rho^i,\Gamma^i)(\tilde{\omega}|h_{t-1})}{D^i(\rho^i,\Gamma^i)(s_t|\omega,h_{t-1})D^i(\rho^i,\Gamma^i)(\omega|h_{t-1})} = \frac{D^i(\rho^i,\Gamma^i)(s_t|\tilde{\omega},h_{t-1})}{D^i(\rho^i,\Gamma^i)(s_t|\omega,h_{t-1})} L_{t-1}$$

[150].

Under the probability distribution $D^i(\rho^i,m^i) = D(\rho)$

$$E[L_t|h_{t-1}] = \sum_{s_t|D^i(\rho^i,\Gamma^i)(s_t|\omega,h_{t-1}^{-i})>0} \frac{D^i(\rho^i,\Gamma^i)(s_t|\tilde{\omega},h_{t-1})}{D^i(\rho^i,\Gamma^i)(s_t|\omega,h_{t-1})} L_{t-1} D^i(\rho^i,\Gamma^i)(s_t|\omega,h_{t-1}) \leq L_{t-1}$$

which is the standard result that the odds ratio is a supermartingale. Since it is also nonnegative, it follow from the martingale convergence theorem (see, for example, Loeve [1978]) that L_t almost surely converges with respect to $D^i(\rho^i,m^i) = D(\rho)$. This in turn implies that

$$\frac{D^i(\rho^i,\Gamma^i)(s_t|\tilde{\omega},h_{t-1})}{D^i(\rho^i,\Gamma^i)(s_t|\omega,h_{t-1})}$$

almost surely converges to 1 with respect to $D(\rho)$. If the vector $D^i(\rho^i,\Gamma^i)(\cdot|\tilde{\omega},h_{t-1}) - D^i(\rho^i,\Gamma^i)(\cdot|\omega,h_{t-1})$ do not also converge almost surely to zero, then this is impossible; since $\omega,\tilde{\omega}$ are complementary events, is must also be that

$$D^i(\rho^i,\Gamma^i)(\cdot|h_{t-1}) - D^i(\rho^i,\Gamma^i)(\cdot|\omega,h_{t-1}) = D^i(\rho^i,\Gamma^i)(\cdot|h_{t-1}) - D(\rho)(\cdot|h_{t-1})$$

almost surely converges to zero. Since $\rho^i(h_{t-1})$ is a best response to $D^i(\rho^i,\Gamma^i)(\cdot|h_{t-1})$ (recall that players are myopic) it is a best response to $D(\rho)(\cdot|h_{t-1})$. Consequently, looking across players $\rho^i(h_{t-1})$ form a Nash equilibrium.

☑

[150] Since beliefs are absolutely continuous with respect to the play path, the denominator in these expressions has positive probability.

It will not escape the careful reader that the hypothesis reads a lot like a definition of equilibrium: The rules are assumed to be best responses to beliefs that are plausible with respect to the play path generated by those rules. This "fixed point" property makes the hypothesis a difficult one. Moreover, there is also a difficulty in understanding how Bayesian players following deterministic rules can converge to a mixed strategy equilibrium, for example if the game is matching pennies. Before taking up these issues in detail in the next section, we consider an example from Kalai and Lehrer that illustrates some of the limitations of this result.

Example 8.1 [Kalai and Lehrer]: Consider the following two player stage game of "chicken"

	Y	I
Y	0,0	1,2
I	2,1	-1,-1

where the strategies are "yield" (Y) or "insist" (I). This game has two pure strategy equilibria one in which player one yields and two insists, and vice versa and a mixed strategy equilibrium. Following Kalai and Lehrer, we suppose that the plausible set consists of strategies of the form "insist for the first *n* periods (possibly infinite), then yield forever." In addition, we suppose that the prior puts exponentially declining weights on these models, and that insisting forever has positive probability.

However, absolute continuity may not be satisfied in this game. The best response to the prior beliefs is to insist for a finite period of time and then yield if the other player has not done so already, since it becomes increasingly likely that the other player will

never yield. Whether such a path satisfies absolute continuity depends on how different the beliefs of the two players are. If one player is much more pessimistic about his opponent yielding, his best response is to almost surely yield first, and absolute continuity is satisfied. However, if both players have the same beliefs (or nearly so) both will yield at the same time. When this occurs, it is then optimal for both players to stop yielding, a path that has zero probability according to the original beliefs, so absolute continuity is violated.[151] This illustrates the problem in finding sets of plausible models that satisfy the absolute continuity assumption. It also raises the question of whether some more complicated beliefs might satisfy the absolute continuity assumption, an issue which we discuss more generally in the next section.

8.4. Interpreting the Grain of Truth Assumption

As we observed, the problem with interpreting the Kalai-Lehrer result as a favorable result about Bayesian learning lies in the fact that the prior beliefs must satisfy the absolute continuity assumption. However, since absolute continuity is endogenous, finding beliefs in principle requires the same kind of fixed point solution that finding an equilibrium does. One solution to this problem is to interpret this result as a descriptive

[151] The plausible sets in this example are reminiscent of equilibrium play in a "war of attrition" game, in which once a player yields she must yield forever afterwards, so that the players' strategy space reduces to the choice of a time to yield if the "war" is still ongoing. This war of attrition has two pure-strategy equilibrium outcomes, "1 yields at the start and 2 insists" and vice versa. These outcomes correspond to equilibria of the repeated game in which one player always insists and the other always yields, which is why the associated asymmetric prior beliefs satisfy absolute continuity in the repeated game. The war of attrition also has symmetric equilibrium in mixed strategies, which corresponds symmetric priors in the repeated game. However, this mixed equilibrium is not an equilibrium of the repeated game, for precisely the same reason that the associated beliefs do not satisfy absolute continuity: if the opponent's strategy is a randomization over strategies all of which specify that the once the player yields she will continue to do so, then a concession-time strategy will not be a best response. Of course, by definition all three of these equilibria of the war of attrition satisfy absolute continuity in the war of attrition itself, as noted by Kalai and Lehrer, but the war of attrition is not a repeated game.

model, rather than an answer to the question of "how do we get to equilibrium." That is, in this setup, the "equilibrium" allows that initially substantial disagreement exists among players, even though ultimately it disappears. In many ways this interpretation is similar to the model (and result) of Jordan [1991], who examines a full Bayesian Nash equilibrium of a repeated game, where players do not initially know what game they are playing. Ultimately, this too converges to Nash equilibrium of the stage game.

Our interest here, however, is in "learning models," by which we mean that the allowed priors are exogenously specified, without reference to a fixed point problem. Ideally, the priors would have the property t regardless of opponents strategies, the absolute continuity assumption is satisfied, but this is impossible: The space H^{-i} is uncountable, and so any probability distribution on this space must place probability zero on some sequences h^{-i} of play, and if opponents were to actually play one such h^{-i}, the absolute continuity assumption would be violated.

Instead, we will explore the weaker possibility that it is possible to specify a class of priors with the property that if all players pick from this class, the absolute continuity assumption is satisfied. Even this weaker goal can be difficult to achieve. In particular, the condition is less readily satisfied in the infinite horizon than in finite truncations of the game, as the following example shows.

	A	B
A	1,1	0,0
B	0,0	1,1

We suppose that players' beliefs are that the opponent's play is independent of their own, and that priors are "eventually equilibrium." By this, we mean that the plausible sets of strategies are non-contingent strategies of the form $(s_1^{-i}, s_2^{-i}, \ldots, s_t^{-i}, s_t^{-i}, \ldots)$ with an arbitrary beginning, but in which the opponent's play eventually converges to a particular pure strategy. Moreover, all such sequences have positive probability, and only such sequences have positive probability.

As an example of such beliefs, suppose that player 1 believes A is 90% likely in period 1, while player 2 believes B is 90% likely. Moreover, both players beliefs are that if your opponent played $(s_1^{-i}, s_2^{-i}, \ldots, s_{t-1}^{-i})$ in the past, there is only a $(.1)^t$ probability that he will fail to play s_{t-1}^{-i} in period *t*. Then each player always plays the way his opponent did last period, but player 1 initially plays A and player 2 plays B. So play alternates deterministically between (A,B) and (B,A), an event that was thought *a priori* to have probability zero. And the two players never manage to coordinate. However, t the absolute continuity assumption is satisfied with respect to any finite truncation of the game; the problem is that it is not satisfied asymptotically.

Of course it may be argued that players should place positive *a priori* weight on two- cycles. But there is still no guarantee, that this will not result in three- cycles. We must check that when each player optimizes against his prior over the plausible set, the resulting play lies with probability one in the set considered plausible by his opponent. The problem in the example above was that it did not.

Further discussion of this problem is based on Nachbar [1995]. For simplicity, we will limit the discussion to the game of matching pennies

	H	T
H	1,-1	-1,1
T	-1,1	1,-1

Recall that a *model* for player *i* is a map $m^{-i}(h_t) \in \Delta(S^{-i})$ from histories to strategies in the repeated game for player *i*'s opponent and that the set of models viewed as plausible by player *i* is denoted by M^i. If a model puts probability one on a single strategy for player *i*''s opponent, we refer to it (by analogy to a pure strategy) as a pure model. For any pure model m^{-i} we denote by $\tilde{m}^i(m^{-i}) = BR^i(m^{-i})$ the pure model that yields a payoff of 1 in every period against m^{-i}. Following Nachbar we assume that if a pure model $m^{-i} \in M^i$ is viewed as plausible by player *i*, then the pure model $\tilde{m}^i(m^{-i})$ is viewed plausible by player -*i*, that is, $\tilde{m}^i(m^{-i}) \in M^{-i}$.

Suppose that there exist best-responses ρ such that the resulting play with probability 1 is plausible for each player *i*, and such that some pure model m^i has positive probability. Then, by assumption this means the model $\tilde{m}^{-i}(m^i)$ must be viewed as plausible by player *i* and so has positive weight in his prior. By Proposition 8.1, this implies that eventually player *i* must learn player -*i* is playing $\tilde{m}^{-i}(m^i)$. Once player *i* learns this fact, he certainly will not any longer play according to m^i, contradicting the fact that ρ^i is a best-response to *i*'s prior.

The difficulty in this line of argument is that it shows only that best responses cannot be plausible if they put a positive weight on a pure model. However, if the best responses are sufficiently mixed (for example an independent 50-50 coin flip in each period, which is an obvious way to play matching pennies), then they may not put positive weight on any pure model. Consequently the question of whether there exist plausible sets for all players such that the best responses lead to plausible outcomes

remains incompletely answered. Nachbar's argument, which can be generalized to cover many other games and approximate as well as exact optimization, does show that the absolute continuity assumption is a difficult one.

8.5. Choosing Among Experts

The problem with the Bayesian approach is that the true process, which is endogenously determined, may not turn out to be in the set of processes initially considered to be possible, and Bayesian updating can have odd consequences when the support of the prior does not contain the process generating the data. Bayesian updating does minimize a certain measure of logarithmic distance to the "true model", but this may yield little utility, as it leads to a (generically unique) deterministic optimal rule that may yield less than the minmax payoff .

For these reasons, we are interested in learning rules that are robust. Recognizing that even with a very diffuse prior, the process that occurs in the course of playing a game against opponents may not be in the support of the prior, we seek rules that do reasonably well, even if the true process generating the data is different than those initially contemplated. Since our measure of success is the utility achieved by the learning rule, it is convenient at this point to abandon the Bayesian point of view, and instead focus on directly on strategies that are thought to be potential best responses. Adopting the computer science terminology, we will refer to these behavior rules as *experts*, and the goal may be thought of as choosing the expert who makes the best recommendations for a course of play. Specifically, we define an *expert* for player *i* to be a map defined on histories, with $e^i(h_t) \in \Delta(S^i)$; this is simply a strategy in the repeated

game. In place of a plausible set of models, we let E^i denote a set of *plausible* experts. For simplicity we will assume that this is a finite set.

Our goal is to demonstrate that a relatively simple procedure of rating experts by their historical performance does about as well asymptotically as the best expert does. In other words, while none of the experts (or models) necessarily does as well as would be possible if the "true" model or expert were considered *a priori* plausible, there is no need to do worse than the expert who is closest to the "true" expert in the sense of getting the highest time average utility among all experts thought to be *a priori* plausible.

To demonstrate this fact, let us recall the results in Chapter 4 about cautious fictitious play; the results about experts can be derived as a corollary of this basic result. We will employ the version of the result for time-varying utility functions. Define

$$\vec{u}_t^i(e^i) = (1/T)\sum_{\tau=1}^{t} u^i(e^i(h_{t-1}), s_t^{-i})$$

to be the utilities that would be realized if the expert e^i played on behalf of player *i*. Note that while the game played by choosing actions is stationary, the game played by choosing experts is time and history dependent, as utility corresponding to choosing a particular expert will depend upon the action that expert recommends given the history. By analogy with the constriction in Chapter 4, we define a rule $\overline{BR}_e^i(\vec{u}_t^i)$ mapping histories to probability distributions over experts by solving the optimization problem

$$\max_{\vartheta^i} \vartheta^i \cdot \vec{u}_t^i + \lambda v^i(\vartheta^i).$$

where $\vartheta \in \Delta(E^i)$ is a probability distribution over the set of plausible experts, v^i is a smooth function that becomes large at the boundaries of the simplex, and λ is a small positive real number .

We also can define a learning rule in the more ordinary sense by first applying $\overline{BR}_e^i(\vec{u}_t^i)$ then letting the (randomly chosen) expert choose the action; denote this by $\overline{BR}^i(\vec{u}_t^i)$.

In this context, and with this notation, we may define an analog of universal consistency

Definition 8.2: A rule ρ^i (mapping histories to mixed actions) is ε*-universally expert* if for any ρ^{-i}

$\limsup_{T\to\infty} \max_{e^i \in E^i} \vec{u}_t^i(e^i) - \frac{1}{T}\sum_t u^i(\rho_t(h_{t-1})) \le \varepsilon$ almost surely with respect to (ρ^i, ρ^{-i}).[152]

This says that the best expert does no more than ε better than the utility actually received.

With this notation, we may restate Proposition 4.5 as

Proposition 8.2: Suppose that v^i is a smooth, strictly differentiably concave function satisfying the boundary condition that as ϑ^i approaches the boundary of the simplex the slope of v^i becomes infinite. Then for every ε there exists a λ such that the $\overline{BR}^i$ procedure is ε-universally expert.

Notice also that we may equally easily derive Proposition 4.5 from Proposition 8.2: we suppose that the plausible experts each recommend playing a fixed action in every period, and every action is represented by some expert. Thus the best expert gets the payoff of paying the action that is optimal against the time average of play, so a universally expert rule is universally consistent.

---

[152] Note that $u^i(\rho_t(h_{t-1}))$ does not need a time subscript, as the rule $\rho_t(h_{t-1})$ is still by definition a choice of action, not a choice of expert. Equivalently, we could define the learning rule to be a choice of expert, in which case utility would depend on the history. Naturally both ways of computing the utility actually realized yield the same answer.

In the case where $v^i(\sigma^i) = -\sum_{s^i} \sigma^i(s^i)\log(\sigma^i(s^i))$ (which is the entropy defined in chapter 3) the scheme for choosing among experts picks them with a frequency proportional to the exponential of the historical utility. This type of exponential weighting scheme was introduced in computer science by Littlestone and Warmuth [1994], Desantis, Markowski and Wegman [1992], Feder, Mehrav and Gutman [1992] and Vovck [1990]. Vovk [1990] gives a proof of Proposition 8.2 in a special case while the complete theorem is shown by Chung [1994] and Freund and Schapire [1995]. Freund and Schapire [1995] are especially attentive to the rate of convergence. There are also various extensions, such as that of Kivinen and Warmuth [1993] to the case of continuous outcomes. A nice review of this literature can be found in Foster and Vohra [1996].

8.6. Conditional Learning

An alternative way to look for robust rules, is to focus directly on robust methods of estimating the set of conditional probabilities of opponents' play given the history of all players' play. In this view, universal consistency is a criterion for robustness where the probabilities are unconditional. Our discussion follows Fudenberg and Levine [1995].

As we noted above, the essential element of conditioning is the partitioning of histories into disjoint sets of events. We suppose that there is a collection of categories Ψ^i into which observations may be placed. A *classification rule* is a map $\hat{\psi}^i : H \times S^i \to \Psi^i$. Prior to observing s_t^{-i} the player knows h_{t-1}, s_t^i, and must choose a category $\hat{\psi}^i(h_{t-1}, s_t^i)$ based only on this information. For simplicity we focus on the case where there are finitely many categories. In the case of countable many categories, this

method type of classification is known in the non-parametric statistical literature as the method of sieves

Fix a classification rule $\hat{\psi}$. Given a history h_t, we define $n_t^i(\psi)$ to be the total number of times the category ψ has been observed. We define $D_t^{-i}(\psi)$ to be the vector whose components are the frequency with which each strategy profile of *i*'s opponents has appeared when ψ has been observed. For example, the category might correspond to the previous period's play, so that the distribution $D_t^i(s^2)$ is simply the empirical distribution of outcomes conditional on the previous period's play having been s^2.[153] Also, denote the average utility received in the subsample ψ by $u_t^i(\psi)$. For each subsample we can define the difference between the utility that might have been and the utility that actually was as

$$c_t^i(\psi) = \begin{cases} n_t^i(\psi)\left[\max_{s^i} u^i(s^i, D_t^{-i}(\psi)) - u_t^i(\psi)\right] & n_t^i(\psi) > 0 \\ 0 & n_t^i(\psi) = 0 \end{cases}$$

We define the total cost to be $c_t^i = \sum_{\psi \in \Psi^i} c_t^i(\psi)$. Our analog to universal consistency relative to the rule $\hat{\psi}$ for choosing subsamples is that the time average cost c_t^i / t should be small.

Definition 8.3: A behavior rule $\rho^1 : H \to \Sigma^1 = (\Delta(S^1))$ is *ε-universally consistent conditional on $\hat{\psi}$* if for every behavior rule $\rho^2 : H \to \Sigma^2$ $\limsup_{t\to\infty} c_t^i / t \le \varepsilon$ almost surely with respect to the stochastic process induced by ρ.

When the classification rule is fixed, we simply refer to a strategy being *ε-universally conditionally consistent.*

[153] The rules considered by Ayoyagi and Sonisno correspond to categorization by the opponent's play in the "recent" past.

We now restrict attention to rules of the form $\hat{\psi}^i(h_{t-1})$; we discuss the more general case below. Let ρ^i be a learning rule. Given any such rule, we create a conditional analog $\rho^i(\hat{\psi})$ in the following simple way. For any history h_{t-1} we can define another history $h_{t-1}(\hat{\psi})$ to be the sub-history of observations in which the category ψ to which the observation was assigned is the same as $\hat{\psi}(h_{t-1})$. That is, if the history h_{t-1} is assigned to the category ψ, we look only at those periods in which ψ was the assigned category. We then define $\rho^i(\hat{\psi})(h_{t-1}) = \rho^i(h_{t-1}(\hat{\psi}))$, that is, we apply the original rule to the sub-history corresponding to the category ψ. The essential feature of such a conditional rule is that if the original rule is universally consistent, then the extended rule is universally conditionally consistent.

Proposition 8.3: If ρ^i is ε-universally consistent, then $\rho^i(\hat{\psi})$ is ε-conditionally universally consistent.

Proof: We examine the cost in the definition of conditional universal consistency

$$c_t^i / t = \sum_{\psi \in \Psi^i} n_t^i(\psi)\left[\max_{s^i} u^i(s^i, D_t^{-i}(\psi)) - u_t^i(\psi)\right] / t$$

If $n_t^i \to \infty$ then $\max_{s^i} u^i(s^i, D_t^{-i}(\psi)) - u_t^i(\psi) \le \varepsilon$ because ρ^i is ε-universally consistent. On the other hand, if $\lim n_t^i < \infty$ $n_t^i / t \to 0$. So clearly $\limsup c_t^i / t \le \varepsilon$.

☑

8.7. Discounting

So far our discussion of learning has been cast in terms of time averages of utility. This reflects the idea that a "good" learning rule ought to do well in the long run. However, economists generally regard people as impatient, and view discounting as a better model of intertemporal preference. In general we can say little about how well

learning rules do in terms of the discounted present value of the players' payoffs. Early in the game, before there is any data to learn from, the player is essentially guessing what the outcome will be. Regardless of whether "learning" is effective or ineffective, a rule that happens to guess well early on can outperform a rule that guesses poorly. Moreover, in addition to guessing the outcome, the player must guess also which patterns of data are most likely. With a small amount of data, only a relatively small number of conditional probabilities can be estimated. If other players switch strategies every other period, a player who guesses this is likely to be the case will outperform a player who is more focused on the possibility of two-cycles.

As a result of these considerations, we cannot hope to compare two arbitrary learning rules and determine that one rule is "better" from a discounted point of view. What we can hope to do is to compare classes of learning rules broad enough to incorporate various possibilities of "guessing." What we will show in this section is that the class of conditional smooth fictitious play rules has a kind of dominance property. Given an arbitrary rule ρ^i and any $\varepsilon > 0$, we can design a conditional smooth fictitious play rule that never does more than ε worse than ρ^i regardless of the discount factor.

To illustrate this result, considering the example of the best-response learning rule and the Jordan three-player matching pennies game. In this game, player 1 wins if he plays the same action as player 2, player 2 wins if he matches player 3, and player 3 wins by *not* matching player 1. If all players follow a fictitious play, play cycles. However, each player, in a certain sense, waits too long before switching: For example, when player 1 switches from H to T, player 3 does not switch from T to H until player 1 has switched for a sufficiently long time that the average frequency with which he has played H drops to ½. Smooth fictitious play performs similarly. However, a player who plays the best-response rule will switch one period after his opponent does, and as a result will

get a much higher payoff (even in the time average sense) than a player using a smooth fictitious play. The reason that the best response rule does better is that it guesses correctly both that opponents' last period play is a good predictor of this periods play, and that the correlation is positive. If in fact the correlation was negative, so for example the opponents alternated deterministically between H and T, then best response would do considerably worse than a smooth fictitious play.

The basic idea we develop in this section is that it is possible (with a small cost) to have the best of both worlds: use a conditional smooth fictitious play conditioning on the opponents' play last period, with a strongly held prior that this periods play is the same as next period's play. In the short run such a rule does exactly the same thing as best response. In the long run, if the correlation is actually positive as it is in the Jordan example, the rule continues to behave like best response. However, if the correlation is negative, as is the case when the opponent alternates deterministically between heads and tails, eventually the data overwhelms the prior, and the conditional smooth fictitious play begins to match the opponents moves, beating both best-response and even ordinary smooth fictitious play.

To establish the basic result, it is helpful to begin with a simple case. We consider an ordinary smooth fictitious play against a particular guess σ^{-i} about how opponents will play. We first show that for any such guess, we can design a smooth fictitious play whose present value is no more than ε lower than that of the guess, regardless of the discount factor. Since the smooth fictitious play is universally consistent, for discount factors close to one, its present value is not much below that of a best response to the limiting value of the empirical distribution of opponents play, while a particular guess may well be.

Lemma 8.1: For any fixed strategy σ^i and any ε there exists a smooth fictitious play ρ^i such that for any strictly decreasing positive weights β_t summing to one and any ρ_t^{-i}

$$\sum_{t=1}^{\infty}\beta_t u^i(\sigma^i,\rho_t^{-i}) \le \sum_{t=1}^{\infty}\beta_t u^i(\rho_t)+\varepsilon .$$

A proof can be found in Fudenberg and Levine [1995] and is along much the same lines as the proof of Proposition 4.5. The key observation is that careful use of the argument in Proposition 4.5 makes it possible to bound the time average loss uniformly, regardless of the length of the horizon. Since the average present value can be written as a convex combination of time averages over all different possible time horizons, this uniform bound gives the desired result in the discounted case.

The rule that is being outperformed, that of guessing the opponent will always play a single action, is not very interesting. However, let ρ^i be an arbitrary deterministic learning rule,[154] and let $\varepsilon > 0$ be given. Take a set of $\Psi^i = S^i$ to be strategies for player *i*. Define the classification rule $\hat{\psi}^i(h_t) = \rho^i(h_t)$, that is, classify histories according to the way in which ρ^i is going to play. For each s^i choose a smooth fictitious play $\overline{BR}^i(\varepsilon, s^i)$ so that Lemma 8.1 is satisfied with respect to ε, and define a rule

$$\hat{\rho}^i(\rho^i,\varepsilon)(h_t) = \overline{BR}^i(\varepsilon,\hat{\psi}(h_t))(h_t(\hat{\psi}))$$

by applying the appropriate smooth fictitious play to the sub-history of the chosen category. Since we showed that Lemma 8.1 held for even for non-stationary discounting (corresponding to skipping periods when a rule is not used), we have the immediate corollary

[154] The extension to random rules is straightforward.

Proposition 8.4: For any rule ρ^i and any ε there exists a conditional smooth fictitious play $\hat{\rho}^i$ for any discount factor $\delta > 0$ and any ρ_t^{-i}

$$(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u^i(\rho_t^i,\rho_t^{-i}) \le (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u^i(\hat{\rho}_t^i,\rho_t^{-i}) + \varepsilon.$$

This shows that even when discounting is considered, the "extra cost" of using a universally consistent rule can be made arbitrarily small, and moreover the loss can be bounded uniformly over the discount factor. From a normative point of view, this result provides an argument that rational agents "should" use universally consistent rules. Whether this implies that real people will tend to use such rules is a more complicated question, but the result certainly does not make that prediction *less* plausible.

8.8. Does Sophisticated Learning Lead to Complex Dynamics?

Sophisticated learning introduces two new possibilities. One is that systems that were unstable with respect to less sophisticated procedures may be stable when more sophisticated procedures are introduced. We discuss this possibility in the next section. Second, it may be that individual pursuit of more sophisticated procedures leads to less stability in the aggregate.

We consider Aoyagi's [1994] model of conditional but exact fictitious play. In this model, histories are categorized according to the outcome during the most recent L periods of the history, where L is a fixed number. That is, each category corresponds to a sequence of L outcomes, and the rule for assigning histories to categories is to assign a history to the category that corresponds to the most recent L outcomes of the history. As in conditional smooth fictitious play, a separate frequency of opponent's play is tracked for each category. However, the Aoyagi model differs from conditional smooth fictitious play in assuming that following each history, a player plays an exact best response to the

frequency for the category, rather than a smoothed best-response. (This is the only difference with conditional smooth fictitious play.) In this model, Aoyagi shows that strict Nash equilibria are stable, and that in a zero-sum game with a unique equilibrium, the marginal frequencies converge to that equilibrium.

The analysis of mixed equilibria is a difficult one, since players are assumed to use exact fictitious play given the conditional frequencies, rather than a smooth fictitious play. Consequently, near a mixed equilibrium, players are not actually randomizing, but varying their play over time. If players are trying to detect such deterministic variation by their opponent, yet more complicated patterns must be introduced to preserve the equilibrium. By way of contrast, we would expect a smooth conditional fictitious play to be relatively robust: near a mixed equilibrium that is stable with respect to unconditional smooth fictitious play with an initial condition that all categories begin with near equilibrium frequencies we would expect that play is actually random with about the equilibrium probabilities, and so the frequencies in all categories would tend to remain near the equilibrium level. In other words, with a smooth conditional fictitious play, there would be no patterns to detect.

Even if stability properties are preserved by sophisticated learning procedures, it is possible that cycles will be introduced when without such sophisticated procedures there were no cycles. This is particularly true since play in the early period will tend to be relatively random, and this may accidentally establish a pattern or cycle that will then take hold. Although this possibility should hold more generally, it has been studied in the context of models in which players follow relatively unsophisticated procedures until a cycle is detected (or thought to be detected), and then a more sophisticated procedure is introduced. In such a variation on conditional ordinary fictitious play, Aoyagi shows that

the stability of mixed equilibrium is reversed due to constant switching back and forth between sophisticated and unsophisticated procedures.

The issue of convergence to a cycle that would not be possible without a sophisticated procedure has been studied by Sonsino [1994]. Sonsino restricts attention to games with generic payoffs, and which satisfy the condition that every "subgame" that is closed under the best response correspondence contains a pure strategy Nash equilibrium.[155] He assumes that players switch between unsophisticated and sophisticated behavior depending on whether patterns have been identified in the past. Patterns are restricted to be sequences of pure Nash equilibria, which has the unfortunate implication that players must know one another's payoffs, but this assumption is probably not essential. Unsophisticated behavior is similar to that of Sanchiricco [1995] and Hurkens [1994] in that players are assumed to have at least some chance of following the best-response dynamic. Sonsino makes a number of other highly specialized assumptions about the learning procedure and shows that the system converges globally to a cycle through the pure Nash equilibria. If there is enough initial randomness, then there is positive probability that a non-trivial cycle is established.

Unlike the methods discussed above using either experts or conditional smooth fictitious play, Sonsino deals with the exact detection of cycles. That is, either a cycle is "detected" with probability one, or it is not detected at all This creates some complications that are worth noting. One method of detecting cycles is to assume that a cycle ABAC, for example, is detected if it is repeated a sufficient number of times. However, there may be no cycles early in the game, with cycles only emerging after play

[155] Here a "subgame" of a strategic-form game is obtained by restricting each player to some subset of the original strategies. A "subgame" is closed under the best-response correspondence if all best responses to profiles in the set lie in the set; which is the definition of a CURB set (see Chapter 4). however, not all CURB sets are "subgames" in Sonsino's sense, since a CURB set need not be a product set; for example a set consisting of two strict Nash equilibria is a CURB set but not a subgame.

has gone on for some time. We would like players to be able to detect these cycles also. For this reason, it seems reasonable to assume that a cycle is "detected" if it has occurred a sufficient number of times in the recent past. There are complications with this as well: Suppose in particular that the sequence of events ABABABAC repeats three times followed by ABABABA. Let the rule for cycle detection be that if a pattern has repeated three times at the end of the history, it is "recognized." In this example, the pattern ABABABAC has repeated three times at the end of the history, so following ABABABA, the player should expect C. However, the pattern AB has also repeated three times at the end of the history, so following the final A, the player should expect B. In this example, two patterns are "recognizable" and each leads to a different conclusion. Sonsino proposes restrictions on cycle recognition procedures that eliminate this ambiguity.

Note that in conditional fictitious play, either the smooth type discussed in Fudenberg and Levine [1995] or the exact type discussed in Aoyagi [1994], this type of issue need not arise, since these models consider more general rules for classifying histories. . For example, if, following Aoyagi, we categorize histories according to the final L outcomes, and simplify by setting $L = 1$, then following A the frequency is 80% B and 20% C, so in effect this is what is "expected" to happen next. More generally, the model of conditional fictitious play allows for any arbitrary rules for assigning histories to categories, and the play observed in a given category need not be the same each time the category is observed. Thus Sonsino's paper can be viewed as exploring the difficulties that arise with a special sort of assignment rules.

Another issue in exact pattern recognition arises when we have a sequence such as ABCAABCDABCADABCDCCAB in which AB is always followed by C, even though there is not an ABC cycle per se. In such a case, it seems sensible that this pattern

might be recognized. Notice that a conditional (smooth or exact) fictitious play will pick this up, provided that $L \geq 2$.

8.9. Does Sophisticated Learning Lead to Stability?

So far we have examined the possibility that sophisticated learning is destabilizing, We turn now to the question of whether it might be stabilizing, that is whether sophisticated learning enables players to avoid confusion, or whether it simply leads to dynamic processes too complex for them to comprehend.

In this section we consider categorization schemes that are *independent* of the player's own anticipated action; we consider endogenous categorization in the following section. First, a simple example shows that more sophisticated learning need not lead to stability. Consider the Shapley game, a two player three action per player game with payoffs of the form

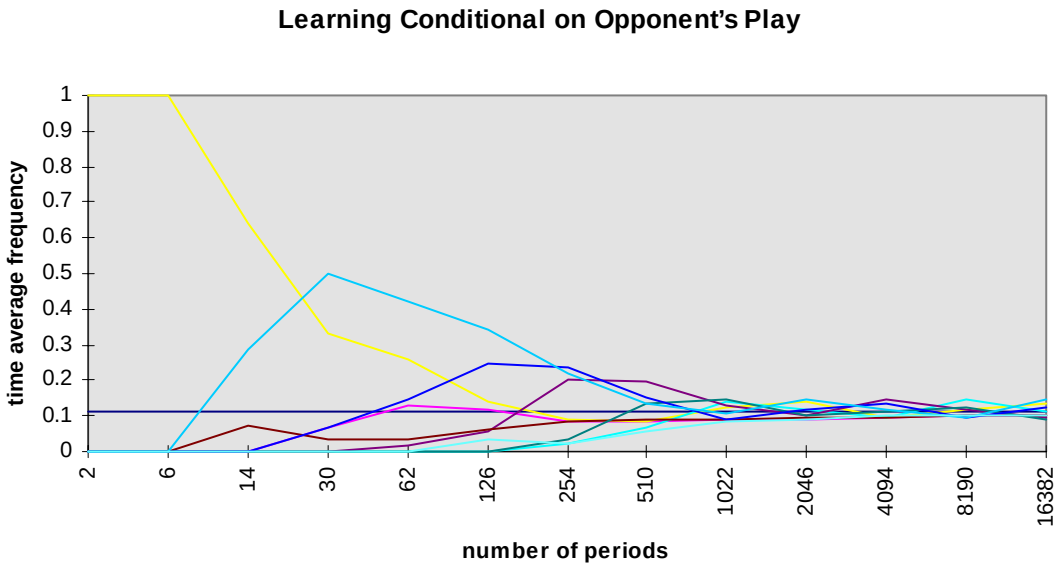
	A	M	B
A	0,0	0,1	1,0
M	1,0	0,0	0,1
B	0,1	1,0	0,0

Suppose first that both players use exponential fictitious play with a single category. We know that play can asymptote approximately to any stable best-response cycle. Recall that in this game, such a cycle begins at A,M. Player 1 then wishes to switch to B. At B,M, player 2 wishes to switch to A, then from B,A to M,A to M,B to A,B then back to the start at A,M. It can be shown that this cycle is asymptotically stable in both the best response and approximate fictitious play dynamic. In the case of interest

to us, approximate fictitious play, the cycles are of ever increasing length. This suggests that players might condition on the profile from last period. Suppose that both players do so. Then it is easy to see that within each of the nine categories, play is simply an approximate fictitious play, and (if the initial condition is the right one in each category), play will simply follow the Shapley cycle within each category. Of course, players might notice this, and introduce even more sophisticated conditional cycle detection, but no matter how complicated the categorization rule they use, as long as they both condition on exactly the same histories, there will still be a Shapley cycle within each category. Suppose, however, the two players are not conditioning on exactly the same histories. This raises the possibility that the players may not be able to "accidentally" correlate their play, as they do when they use exactly the same conditioning procedure. To understand this possibility better, let us consider the case where each player conditions only on his opponent's last action (but not his own) in the Shapley game.[156] The resulting dynamical system has 18 dimensions, since each player must keep track of the number of occurrences of three outcomes for each of three categories corresponding to opponent's last period play. Since it is difficult to analyze such a high dimensional system analytically, Fudenberg and Levine [1995] used a simulation. Each player was assumed to use a smoothing function of the form $v^i(\sigma^i) = -(1/\kappa)\sum_{s^i} \sigma^i(s^i)\log\sigma^i(s^i)$, where $\kappa = 10$, so that within categories players use an exponential fictitious play. The payoffs are those for the Shapley game given above. To initialize the system, each player was endowed with 12 initial observations independent of category. Player 1 is endowed with the initial sample (1,1,10), and player 2 (10,1,1). Given these frequencies, it is optimum

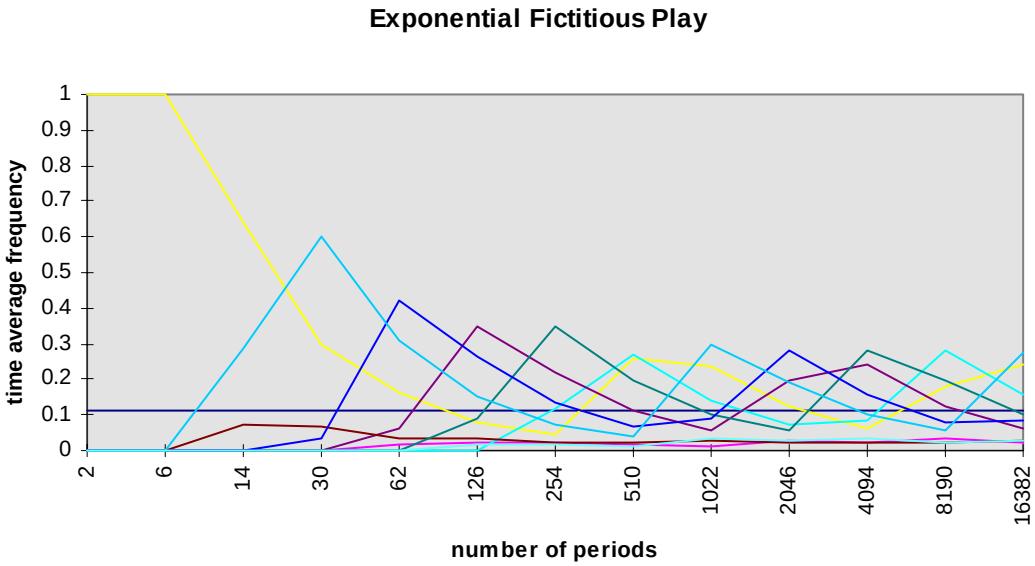
[156] Aoyagi [1994] also considers an example where two players classify observations using different categories. However, in his example players are using exact fictitious play within categories so the player with a more refined category scheme can exploit the player with the less refined scheme. This cannot happen with exponential fictitious play.

for player 1 to play to play A and player 2 to play M, which is an initial condition that starts the Shapley cycle. The graph below reports the time average of the joint distribution of outcomes.



Each line in the graph represents the time average frequency of a single outcome, for example, the line that is nearly 1 for the first 6 periods corresponds to the outcome A,M. The horizontal line represents the common frequency in the unique Nash equilibrium of 1/9. Note that the horizontal axis is measured in logarithmic units, since this is the time scale over which the Shapley cycle occurs. In this simulation the system does not cycle, but has essentially converged to the Nash equilibrium after 1000 periods.

By way of contrast, Fudenberg and Levine also did a simulation in which players did not condition on the history at all. That is, each player uses a single category, and play corresponds to ordinary exponential fictitious play. All the other parameters including the initial conditions were held fixed at the values given above. The results of the simulation are shown in the graph below.



As we expect, the system cycles settling into a relatively stable cycle after about 500 periods. A significant feature is that the frequencies corresponding to the diagonal A,A, M,M and B,B remain close to zero. This is not the case when players condition on each other's previous actions.

In general we do not know to what extent the "noise" introduced by different players conditioning on different histories ultimately causes correlation to break down. If it does, then in the long run learning will lead to Nash equilibrium. Although we have seen that this happens in a particular example, whether it is generally the case remains an open question for future research.[157]

[157] Recent papers by Sanchirico [1995] and Sonsino [1994] have stressed another implication of noise: it can help insure that long-run play ends up in one of the "minimal curb sets" of the game (see chapter 4). This is an interesting fact, but rather different than what we are discussing here, as in the Shapley game the minimal curb set is the entire game.

8.10. Calibration and Correlated Equilibrium

A player may, if he wishes, condition also on his own choice of action. Does this make sense? One implication of the Shapley cycle is that (holding fixed the play of the opposing player) a player get less utility than if he conditioned on his own intended action. If both players condition on the same information, we must ask whether players would continue to unsuccessfully introduce more and more complicated "independent" schemes for detecting cycles or if they would simply condition also on their own anticipated action.

Suppose that players condition on their own anticipated action in the sense that no category can ever be assigned two observations in which the player's own action is different. This has two implications. First, the two players are not conditioning on exactly the same histories since each player can anticipate his own action, but not his opponents. This raises the possibility that players cannot "accidentally" correlate their play leading to Nash equilibrium in much the same way that conditioning only on the opponent's previous action did in the discussion above.[158] Second, the empirical joint distribution of action profiles must come to approximate a correlated equilibrium.[159] This follows immediately since each player, conditional on his own action, is playing an approximate best response to the distribution of opponents play.

The resemblance of the empirical joint distribution of action profiles to a correlated equilibrium is closely related to a point originally made by Foster and Vohra [1993] who consider best responses to beliefs that are *calibrated.* This means that if we

---

[158] Foster and Vohra [1995] examine the case of conditioning only on a player's own anticipated action and give simulation results with convergence to Nash equilibrium very much like those described above in the case where players condition on their opponent's previous action.

[159] However, unless the correlated equilibrium is actually a Nash equilibrium, the empirical joint distribution cannot converge, and must shift from one correlated equilibrium to another.

categorize periods according to forecasts of the opponent's play, the frequency distribution of actual play of the opponent converges to the forecast. For example, we would say that a weather forecaster is calibrated if on those occasions on which a 50% chance of rain was announced, it actually rained 50% of the time. [160]From our perspective, the important feature of calibration of beliefs is that it implies that actions are *calibrated.* By this we mean that each player conditional on his own action, is playing an approximate best response to the distribution of opponents play, the condition that leads directly to correlated equilibrium.

However, even the use of calibrated rules leaves open the possibility that opponents "accidentally" correlate their play. If they did not do so, then we would have the stronger result that play would eventually come to resemble a Nash equilibrium. What we should point out is that calibration guarantees no such thing. The easiest way to make this point is to observe that calibration of actions does not actually require that players condition only on their own actions; it is sufficient that no category can ever be assigned two periods with more than two distinct values of the player's own action. The essential point is that if a joint distribution over two actions and all outcomes has the property that the realized utility is at least that that can be obtained by playing a single best response to the marginal distribution over outcomes, then each action must actually be a best response conditional on that action.[161] In particular, if each player has only two actions, then his strategy is calibrated even if he uses only a single category, for example, exponential fictitious play is calibrated in this case.

---

[160] This notion of calibrated lumps together all histories in which the forecaster's prediction was the same. Refining this by conditioning on all histories leads to a notion very close to the statistical concept of consistency. This connection is examined in detail in Kalai, Lehrer and Smorodinsky [1996].

[161] More generally, it follows from straightforward algebra that if a joint distribution over all actions and all outcomes has the property that the realized utility is at least that that can be obtained by playing a best response to the marginal distribution over outcomes, then no action that has positive probability is a worst response to the distribution of outcomes conditional on that action.

Consider, then, the three player, two action per player game of matching pennies introduced by Jordan [1991]. In this game player 1 wishes to match player 2, who wishes to match player 3, who wishes to avoid matching player 1. Again we can apply the results of Benaim and Hirsch [1994], and, supposing that player use exponential fictitious play,[162] focus on best-response cycles. The key is that in every pure strategy profile exactly one player can gain by deviating. Once he does so, one opponent wishes to switch, and so on in a cyclic fashion. Jordan also shows that this cycle is asymptotically stable under exact fictitious play; Benaim and Hirsch extend this to stochastic smoothed versions. Moreover, since each player has only two actions, it follows that this cycle takes place (approximately) through the set of correlated equilibria. Despite the fact that players are calibrated and play resembles a correlated equilibrium, this cycle is just as disturbing as the Shapley cycle: players observe long sequences of their opponent repeatedly playing the same action, yet fail to anticipate it will happen again. If they (all) introduce schemes conditioning on last period's play, then the cycle simply takes place independently in each category and so forth and so on. Consequently the issue in cycling is not calibration per se, but rather, how the players categorization rules fit together.

We next present Foster and Vohra's result that it is possible to design learning rules that condition on a player's own anticipated action. For simplicity, we suppose that this is the only part of the history to be conditioned on. In other words, we take the categories to be $\Psi^i = S^i$, the players own strategies, and the categorization scheme to be $\hat{\psi}^i(h_{t-1}, s_t^i) = s_t^i$. Formally, we refer to the notion of universal consistency in this case as *calibration*.

---

[162] Implicitly we assume that when there are more than two players, the profile of opponents' actions is treated as a single outcome. This means that each player tracks a joint distribution of opponents' play. Jordan actually supposes that players do fictitious play by keeping a separate distribution for each opponent, assuming that opponents' play is independent. However, in this particular game, there is no distinction between the two procedures, since each player cares only about the play of one opponent.

Definition 8.3: A behavior rule $\rho^1: H \to \Sigma^1$ is ε*-calibrated* if for every behavior rule $\rho^2: H \to \Sigma^2$ $\limsup_{t\to\infty} \sum_{s^i} n_t^i(s^i)\left[\max_{\sigma^i} u^i(\sigma^i, D_t^{-i}(s^i)) - u_t^i(s^i)\right]/t \le \varepsilon$ almost surely with respect to the stochastic process induced by ρ.

Instead of following Foster and Vohra's [1995] proof, we give a simplified construction based on an arbitrary universally consistent rule. Let ρ^i be any ε-universally consistent learning rule. Suppose that player *i* is contemplating playing s^i. Then he ought to try to apply ρ^i to $h_t(s^i)$. The only problem with this is that ρ^i does not put probability one on playing s^i. Suppose, however, that player *i* is contemplating playing σ^i. Then with probability $\sigma^i(s^i)$ he will play s^i, and should therefore play $\rho^i(h_t(s^i))$. Consequently, he will actually wind up playing according to $\sum_{s^i} \sigma^i(s^i)\rho^i(h_t(s^i))$. If in fact

$$\sigma^i = \sum_{s^i} \sigma^i(s^i)\rho^i(h_t(s^i)),$$

then his contemplated play and desired play wind up being the same. Notice that $\sigma^i = \sum_{s^i} \sigma^i(s^i)\rho^i(h_t(s^i))$ is a simple fixed point problem based on a linear map from the non-negative orthant to itself. So it may easily be solved by linear algebra. Denote the solution $\hat{\rho}^i(h_t)$.

Proposition 8.5: If ρ^i is ε-universally consistent, then $\hat{\rho}^i$ is ε-calibrated.

Proof: We examine the asymptotic average cost

$$\begin{aligned}
&\sum_{s^i} n_t^i(s^i)\left[\max_{\sigma^i} u^i(\sigma^i, D_t^{-i}(s^i)) - u_t^i(s^i)\right]/t \\
&= \sum_{s^i}\left(n_t^i(s^i)\left[\max_{\sigma^i} u^i(\sigma^i, D_t^{-i}(s^i))\right] - \sum_{\tau\le t|\hat{\psi}^i_\tau = s^i} u^i(s^i, s_\tau^{-i})\right)/t \\
&= \sum_{s^i}\left(n_t^i(s^i)\left[\max_{\sigma^i} u^i(\sigma^i, D_t^{-i}(s^i))\right]\right)/t - \sum_{\tau=1}^{t} u^i(s_\tau^i, s_\tau^{-i})/t
\end{aligned}$$

By the strong law of large numbers for orthogonal sequences, $\sum_{\tau=1}^{t} u^i(s_\tau^i, s_\tau^{-i})/t$ almost surely has the same in the limit as

$$\begin{aligned}
&\sum_{\tau=1}^{t} u^i(\hat{\rho}_\tau^i, s_\tau^{-i})/t \\
&= \sum_{\tau=1}^{t}\sum_{s^i} u^i(\rho^i(h_\tau(s^i)), s_\tau^{-i})\rho^i(h_\tau(s^i))(s^i)/t \\
&= \sum_{\tau=1}^{t}\sum_{s^i} u^i(\rho^i(h_\tau(s^i), s_\tau^{-i})\rho_\tau^i(s^i)/t
\end{aligned}$$

where we have used the defining equation for $\hat{\rho}^i$. Again applying this strong law of large numbers for orthogonal sequences, this almost surely the same limit as

$$\begin{aligned}
&\sum_{\tau=1}^{t} u^i(\rho^i(h_\tau(s_\tau^i)), s_\tau^{-i})/t \\
&= \sum_{s^i}\sum_{\tau\le t|\hat{\psi}_\tau^i = s^i} u^i(\rho^i(h_\tau(s^i), s_\tau^{-i})/t
\end{aligned}$$

substituting back in the expression for average cost, we find

$$\begin{aligned}
&\sum_{s^i} n_t^i(s^i)\left[\max_{\sigma^i} u^i(\sigma^i, D_t^{-i}(s^i)) - u_t^i(s^i)\right]/t \\
&= \sum_{s^i}\left(n_t^i(s^i)\left[\max_{\sigma^i} u^i(\sigma^i, D_t^{-i}(s^i))\right] - \sum_{\tau\le t|h_\tau^i = s^i} u^i(\rho^i(h_\tau(s^i), s_\tau^{-i})\right)/t \\
&= \sum_{s^i}\left(n_t^i(s^i)/t\right)\left(\left[\max_{\sigma^i} u^i(\sigma^i, D_t^{-i}(s^i))\right] - (1/n_t^i(s^i))\sum_{\tau\le t|h_\tau^i = s^i} u^i(\rho^i(h_\tau(s^i), s_\tau^{-i})\right)
\end{aligned}$$

However, since ρ^i is ε-universally consistent, either

$$\limsup\left(\left[\max_{\sigma^i} u^i(\sigma^i, D_t^{-i}(s^i))\right] - (1/n_t^i(s^i))\sum_{\tau\le t|h_\tau^i = s^i} u^i(\rho^i(h_\tau(s^i), s_\tau^{-i})\right) \le \varepsilon$$

or $n_t^i(s^i)/t \to 0$, which gives the desired result.

☑

We should emphasize that this is not the only learning rule leads to correlated equilibrium in the long run. For example, Hart and Mas-Colell [1996] consider the following rule:

$$\rho^i(h_t)[s^i] = \begin{cases}
(1/\mu)\left[u^i(s^i, D_t^{-i}(s_{t-1}^i)) - u_{t-1}^i(s_{t-1}^i)\right] & s^i \ne s_{t-1}^i, \quad u^i(s^i, D_t^{-i}(s_{t-1}^i)) - u_{t-1}^i(s_{t-1}^i) \ge 0 \\
0 & s^i \ne s_{t-1}^i, \quad u^i(s^i, D_t^{-i}(s_{t-1}^i)) - u_{t-1}^i(s_{t-1}^i) < 0 \\
1 - \sum_{r^i}(1/\mu)\left[u^i(r^i, D_t^{-i}(s_{t-1}^i)) - u_{t-1}^i(s_{t-1}^i)\right] & s^i = s_{t-1}^i
\end{cases}$$

where $\mu > 0$ is a "sufficiently" large constant that $\rho^i(h_t)[s^i_{t-1}] > 0$. Under this rule, the player either plays the same action as last period, or plays another action with probability proportional to how much better that alternative would have done conditional on the history corresponding to the action played last period. Hart and Mas-Colell show using Blackwell's [1956] approachability theorem that if all players use rules of this type, then each rule is calibrated against a class of outcomes that are has probability 1 in this environment.

These rules are not *universally* calibrated; that is, they are not calibrated against all opponents' play. For example, suppose the probability of playing the same action as last period is bounded below by ¾, and that the game is matching pennies. Clearly a clever opponent will always play the opposite of what a player using this rule used last period, since this wins at least ¾ of the time, and so in this case the Hart and Mas-Colell rule loses ¾ of the time. However, since these relatively simple rules work are calibrated against each other, it may be plausible that the actual, rules that people use have this property as well.

8.11. Manipulating Learning Procedures

The focus of attention in this book, as well as in the recent game theory literature, has been on myopic learning procedures, not in the sense that players do not care about the future but in the strategic sense of lacking concern about the consequences of current play for opponents future action. We have justified this by sometimes casual reference to large populations.

This section discusses two related points. First, although the idea of extrapolation between "similar" games suggests that the relevant population may be large even when there are few people playing *precisely* the game in question, there are also situations of

interest in which the relevant population must be viewed as small, so it is of some interest to consider the case of a small population. This raises an important issue: a player may attempt to manipulate his opponent's learning process, and try to "teach" him how to play the game. This issue has been studied extensively in models of "reputation effects, which typically assume Nash equilibrium, but not in the context of learning theory. A second issue has been raised by Ellison [1994], who considers the possibility of contagion effects in the type of large population anonymous random matching model we have used to justify myopia. Under certain conditions, Ellison shows that even in this setting there is a scope for a more rational player to teach his opponents how to play. In particular this is true if the more rational player is sufficiently patient relative to the population size, but not if the population size is large relative to his patience (that is, the order of limits matters).

A Model of Reputation: A simple setting in which to begin to understand teaching an opponent to play the game is to imagine that one player is myopic and follows the type of learning procedure we have discussed in this book, while another player is sophisticated and has a reasonably good understanding that his opponent is using a learning procedure of this type. What happens in this case? This has been studied extensively in the context of equilibrium theory, where models of this type are called "reputational", but not in the context of learning theory. However, much as Kalai and Lehrer [1993] show that the results of Jordan [1991] on equilibrium learning carry over to the case of non-equilibrium learning, so we expect that the lessons of the literature on reputation will carry over also to the case of non-equilibrium learning.

In order to introduce learning into an equilibrium context, it is necessary, as is the case in the Jordan model, to introduce uncertainty on the part of the players about the

game that is being played. That is, while Nash equilibrium and its refinements suppose that players know one another's strategies, they are allowed to have doubts about their opponents' preferences, which in many ways is the same thing as having doubts about their strategies. Suppose, as in many papers on reputation effects, that there are two players, a long-run player and a short-run player. The short-run player is myopic and is the "learner." The long-run player has many different types, the type remaining fixed as the game is repeated, and each type corresponding to different preferences. Consequently, the short-run player wishes to learn the type of long-run player in order to play a best response to the actions of that type. Because of the fact that this is an equilibrium theory, if the short-run player is to have relatively diffuse priors about the strategy of the long-run player, it is important that in equilibrium different types of long-run player really play different strategies. To solve this problem Kreps and Wilson [1982] and Milgrom and Roberts [1982] introduced the idea of committed types with preferences that force them to play a particular strategy, regardless of the particular equilibrium.

The second issue that must be addressed is the long-run consistency of the learning procedure used by the short-run player. Fudenberg and Levine [1992] show that the beliefs of the short-run player converge to something observationally equivalent to the truth at a *uniform* rate, using an argument based on up-crossing numbers of supermartingales. Essentially, the reputational literature introduced the idea of committed types, precisely so that Blackwell and Dubbins absolute continuity assumption would be satisfied.

If we now assume that the long-run player is relatively patient, then Fudenberg and Levine [1989] show that he can get almost as much utility as he could get in the Stackelberg equilibrium of the stage game. The idea is that the long-run player can

guarantee himself at least this much by playing the optimal precommitment strategy forever. The basic argument carries over in a straightforward way to the case of non-equilibrium learning:[163] If the long-run player plays the optimal precommitment strategy forever, the short-run player will eventually learn this and begin to play a best response to it. Since the long-run player is very patient, this means that the average present value received will be nearly that of playing the optimal precommitment strategy with the short-run player playing a best response to it. Moreover, since the short-run player is always playing a best-response to some beliefs about the long-run players strategy, the long-run player cannot really hope to do better than this. The point is, that if your opponent is playing myopically, rather than do the same, you should play as a Stackelberg leader and "teach" him how to play against you.

Teaching in a Large Population: The key ingredient in the reputation model is that the patient (or rational) player can change the behavior of his opponents in a significant way through his own action. It is natural to conjecture that this would not be true in the type of large population, anonymous random matching model we have been using to justify myopic play. However Ellison [1994] has pointed out that this need not be true, due to contagion effects.

This point is best understood in the following example taken from Ellison. Suppose that there is a homogeneous population of N agents playing the following 2x2 pure coordination game with anonymous random matching:

	A	B

[163] Essentially this point is made in Watson [1993] and Watson and Battigalli [1995], who weaken Nash equilibrium to rationalizability.

A	10,10	0,0
B	0,0	1,1

Note that there are two pure Nash equilibria in this game at (10,10) and (1,1). One of these, the Pareto efficient equilibrium at (10,10) is also the Stackelberg equilibrium. In other words, a player who can teach his opponents how to play, would like to teach them to play A.

Suppose first that every player follows the behavior prescribed by exact fictitious play, with prior weights (0,1). Then all players choose B in period 1, and the result is that all players choose B in every period. Suppose next that for some reason, player *N* plays A in period 1, and follows fictitious play in all future periods, while players 1 through *N*-1 continue to follow fictitious play in every period. Then whoever is matched with player *N* in period 1 has weights (1,1) in period 2, and so must play A until at least period 10; call this player 1. Then suppose that player 1 is not matched with player N in period 2, but with some other player 2; this player 2 will also play A at least until period 10. If moreover 1 and 2 are matched with 2 new players 3 and 4 in period 3, and not with themselves or with player N, then there will be 4 players who play A in period 4. At each period until period 10, there is a positive probability that the number of A-players doubles, so that if *N* is small enough there is a positive probability that every player plays A in period 9, so that only A is played from that period on.

Now suppose that player *N* is rational and knows that all other players follow fictitious play. Then player N knows that by playing A in the first period only, and following fictitious play thereafter, there is a nonzero probability that the entire population will move permanently to the Pareto-preferred equilibrium (10,10) in 10

periods.[164] For a small discount factor the short-run cost of inducing this shift might exceed the expected present value of the benefit. Indeed, for any fixed discount factor the expected present value becomes small as the population size grows, since it must take at least log(M) periods to change the play of M other agents. However, changing the order of limits changes this conclusion: for any fixed population there is some time T such that there is a positive chance that all agents will be playing A from period T on if the rational agent plays A in the first period. Thus, the benefit from playing A outweighs its cost if the rational player's discount factor is sufficiently close to 1. Ellison computes that this simple but non-myopic strategy improves on the naïve one in a population of 100 if the discount factor exceeds .67, and that even in a population of 5000 the non-naïve play yields a higher payoff if the discount factor exceeds .94.[165]

Note moreover that *regardless* of the discount factor, the rational player has no incentive for non-naive play if naïve play would yield his preferred equilibrium. Ellison shows that the converse is not true: in general 2x2 coordination games, even when the rational player would prefer the other equilibrium, he cannot steer play in that direction unless the "preferred" equilibrium is also risk-dominant: players defect from a risk-dominated equilibrium too quickly for "contagion" to take hold.

It is also worth noting that the incentive to "teach" opponents in a large population in the example is not robust to noisy play by the players. If players randomize, then the contagion is likely to occur even without the intervention of a rational "teacher," and so

---

[164]If the shift does not occur in the first ten periods, players who have seen A only once will return to playing B in period 11, but there is still a chance that the contagion will resume from the base of players who
saw 2 or more A's in the first 10 periods.

[165] In some sense these calculations may overstate the case, since using a game with less extreme payoff differences would yield less striking numbers. On the other hand, the incentive to "teach" opponents is even greater than in the calculations if players attach greater weight to more recent observations, as with exponential weighting.

the incentive to intervene is reduced. The following example indicates the extent to which contagion is likely to occur with even a "small" amount of noise. Consider the general coordination game (Ellison's example corresponding to a=10)

	A	B
A	*a, a*	0,0
B	0,0	1,1

Here we assume that $a>1$, so that A,A is the Pareto preferred equilibrium. Suppose that players rather than using the usual deterministic fictitious play, use a smooth fictitious play. If the smoothing function is

$$v^i(\sigma^i)=\sum_{s^i}-\sigma^i(s^i)\log\sigma^i(s^i),$$

as we saw in chapter 4, the smoothed best response is

$$\overline{BR}^i(\sigma^{-i})[s^i]\equiv\frac{\exp\left((1/\lambda)u^i(s^i,\sigma^{-i})\right)}{\sum_{r^i}\exp\left((1/\lambda)u^i(r^i,\sigma^{-i})\right)}.$$

If players' play converges to a symmetric deterministic steady state in which each player plays A with probability σ_A, then by a standard extension of the strong law of large numbers, the empirical distributions will converge to the same limit with probability 1.[166] Asymptotic empiricism implies that the assessments converge to this limit value along every path, so that at the steady state

$$\sigma_A=\overline{BR}^i(\sigma_A)=\frac{\exp((1/\lambda)a\sigma_A)}{\left(\exp((1/\lambda)a\sigma_A)+\exp((1/\lambda)\sigma_A)\right)}.$$

This corresponds to the "quantal response equilibrium" of McKelvey and Palfrey [1995]. A calculation shows that for each $a>1$ there is a sufficiently large λ (that is, enough

[166] See, for example, Fudenberg and Kreps [1993]. The only way that this differs from the standard form of the strong law is that the distribution at each date *t* may depend on the history to that date.

noise) that this equation has a unique solution, and that this solution satisfies $\sigma_A > 0.5$, that is, the Pareto preferred action is more likely.

To report on the quantitative significance of noise for the steady state, we measure the size of the noise by $b_A(\lambda) = \overline{BR}^i(\sigma_A = 0)$, that is, the probability that the action *A* is used when the assessment is that the probability of B is one. (In the usual fictitious play, this probability is zero). For each *a* we can calculate the least value of $b_A(\lambda)$ for which there is a unique symmetric steady state, together with the steady state value of σ_A. This is reported in the table below.

a	$b_A(\lambda)$	σ_A
1.1	10.0%	85%
1.3	8.3%	95%
1.5	6.3%	99%
2	4.7%	100%
3	2.9%	100%
4	2.2%	100%
6	1.5%	100%
7	1.2%	100%
10	1.0%	100%

In Ellison's example, even 1% noise is enough to guarantee a unique steady state in which (to the limit of computer precision) 100% of the time A is played. However the table indicates how remarkably strong the contagion effect is: when the Pareto preferred equilibrium is only a 10% improvement (*a*=1.1), a 10% noise ratio leads to a unique equilibrium in which *A* is played 85% of the time. If the system is going to converge to a

favorable steady state anyway, then there is little incentive for intervention by a rational player.

This example shows several things. First of all, the incentive to "teach" opponents is diminished in noisy environments, and it becomes more reasonable for players to behave myopically. This is not to say that the outcome in a noisy model with one rational player is the same as that in the standard fictitious play model without noise, where all players are myopic. Rather, the second point to be drawn from the example is that, once again, small amounts of noise can serve to select between the long-run outcomes of the noiseless model.

References

Aoyagi, M. [1994]: "Evolution of Beliefs and the Nash Equilibrium of a Normal Form Game," Princeton.

Benaim, M. and M. Hirsch [1994]: "Learning Processes, Mixed Equilibria and Dynamical Systems Arising from Repeated Games," UC Berkeley.

Blackwell and Dubins [1962]: *????*.

Blackwell, D. [1956]: "An Analog of the Minmax Theorem for Vector Payoffs," *Pacific Journal of Mathematics*, 6: 1-8.

Chung, T. [1994]: "Approximate Methods for Sequential Decision Making Using Expert Advice," *Proceedings of the 7th Annual ACM Conference on Computational Learning Theory*, 183-189: 1994.

Desantis, A., G. Markowski and M. Wegman [1992]: "Learning Probabilistic Prediction Functions," *Proceedings of the 1988 Workshop of Computational Learning*, 312-328.

Ellison, G. [1994]: "Learning with One Rational Player," MIT.

Feder, M., N. Mehrav and M. Gutman [1992]: "Universal Prediction of Individual Sequences," *IEEE Transactions on Information Theory*, 38: 1258-1270.

Foster, D. and R. Vohra [1993]: "Calibrated Learning and Correlated Equilibrium," Wharton.

Foster, D. and R. Vohra [1995]: "Asymptotic Calibration," Wharton School.

Foster, D. and R. Vohra [1996]: "Regret in the On-line Decision Problem," Wharton School.

Freund, Y. and R. Schapire [1995]: "A Decision Theoretic Generalization of On-Line Learning and an Application to Boosting," *Proceedings of the Second European Conference on Computational Learning*, ?:???.

Fudenberg, D. and D. K. Levine [1989]: "Reputation and Equilibrium Selection in Games with a Patient Player," *Econometrica*, 57 (July): 759-778.

Fudenberg, D. and D. K. Levine [1992]: "Maintaining a Reputation when Strategies are Not Observed," *Review of Economic Studies*, ???.

Fudenberg, D. and D. K. Levine [1995]: "Conditional Universal Consistency," UCLA.

Fudenberg, D. and D. Kreps [1993]: "Learning Mixed Equilibria," *Games and Economic Behavior*, 5: 320-367.

Hart, S. and A. Mas-Collel [1996]: "A Simple Adaptive Procedure Leading to Correlated Equilibrium," mimeo.

Hurkens, S. [1994]: "Learning by Forgetful Players: From Primitive Formations to Persistent Retracts," mimeo.

Jordan, J. [1991]: "Bayesian Learning in Normal Form Games," *Games and Economic Behavior*, 3:60-81.

Kalai, E. and E. Lehrer [1993]: "Rational Learning Leads to Nash Equilibrium," *Econometrica*, 61: 1019-1046.

Kalai, E., E. Lehrer and R. Smorodinsky [1996]: "Calibrated Forecasting and Merging," Northwestern MEDS #1144.

Kivinen, J. and M. Warmuth [1993]: "Using Experts in Predicting Continuous Outcomes," *Computational Learning Theory: EURO COLT*, (: Springer-Verlag), 109-120.

Kreps, D. and R. Wilson [1982]: "Reputation and Imperfect Information," *Journal of Economic Theory*, 50: 253-79.

Littlestone, N. and M. Warmuth [1994]: "The Weighted Majority Algorithm," *Information and Computation*, 108: 212-261.

Loeve, M. [1978]: *Probability Theory II*, (Berlin: Springer Verlag).

McKelvey, R. and T. Palfrey [1995]: "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, Forthcoming.

Milgrom, P. and J. Roberts [1982]: "Predation, Reputation and Entry Deterrence," *Econometrica*, 50: 443-60.

Nachbar, J. [1995]: "Prediction, Optimization and Learning in Repeated Games," *Econometrica*, forthcoming.

Sanchirico, C. [1995]: "Strategic Intent and the Salience of Past Play: A Probabilistic Model of Learning in Games," mimeo.

Sonsino, D. [1994]: "Learning to Learn, Pattern Recognition and Nash Equilibrium," Stanford.

Vovck, V. [1990]: "Aggregating Strategies," *Proceedings of the 3rd Annual Conference on Computational Learning Theory*, 371-383.

Watson, J. [1993]: "A 'Reputation' Refinement without Equilibrium," *Econometrica*, 61: 199-205.

Watson, J. and P. Battigalli [1995]: "On 'Reputation' Refinements with Heterogeneous Beliefs," UC San Diego 95-26.