Plenary talk: 1994 Society for Economic Dynamics and Control Meetings

*Consistency and Cautious Fictitious Play**

Drew Fudenberg and David K. Levine

June 10, 1994

JEL Classification: C72, C73, C63, D83

Keywords: learning, games, consistency, rationality

Abstract: We study a variation of fictitious play, in which the probability of each action is an exponential function of that action's utility against the historical frequency of opponents' play. Regardless of the opponents' strategies, the utility received by an agent using this rule is nearly the best that could be achieved against the historical frequency. Such rules are approximately optimal in i.i.d. environments, and guarantee nearly the minmax regardless of opponents' behavior. Fictitious play shares these properties provided it switches "infrequently" between actions. We also study the long run outcomes when all players use consistent and cautious rules.

Corresponding Author:
David K. Levine FAX: 310-825-3810
Department of Economics
UCLA
Los Angeles, CA 90024

* We thank Larry Blume, Eddie Dekel-Tabak, and Glenn Ellison for helpful comments.

1. Introduction

There are three major views of why we might expect to see equilibrium in a game: the most traditional introspective view has players study the rules closely, and consider their opponents motivation to calculate what strategy they should play. Evolutionary and learning models see equilibrium rather as the outcome of a process in which less than fully rational players grope for optimality over time. Evolutionary models focus on a population of players and the non-modeled idea that the number of players playing actions that have historically been successful will increase over time at the expense of actions that have historically been less successful. Learning models take a more individualistic point of view, focusing on how an individual player might try to deduce from careful observation of opponents' past play how they will play in the future[1]. The focus of this paper is on the issue of individual learning.

There are two types of questions that can be asked about particular learning rules: How well do they do? That is, how much utility do they generate in different environments? Second, what happens in a game if particular learning rules are used? The latter question has been the focus of a number of recent papers, including Fudenberg and Kreps (1993), Fudenberg and Levine (1993), Jordan (1993) and Young (1993), as well as the earlier literature on the process of fictitious play (Brown (1951), Shapley (1964), and so forth).

This paper focuses more on the former question: how well do learning rules do, and what are sensible criteria for evaluating the performance of a learning rule? We propose as *desiderata* for learning rules that they be "safe"- meaning that they guarantee the player at least his minmax payoff- and "consistent", meaning that they should do at least as well as playing the best response to the empirical average of play *if* the opponents'

[1] There are some close connections between dynamics induced by evolutionary and learning processes, that have been explored, for example, by Gaunersdorfer and Hofbauer (1994).

play is given by independent draws from a fixed distribution. We then suggest that behavior rules should be not just consistent, but "universally consistent," meaning that the player should get at least the payoff of playing a best response to the empirical distribution whether or not the environment is in fact i.i.d. Such a universally consistent rule is both consistent and safe.

Standard fictitious play is consistent, but not safe. Our main result is that there is a very simple modification of fictitious play which is universally consistent and so both safe and consistent. We also show that fictitious play itself is consistent provided that it does not alternate "too quickly" between actions.[2] In addition, we investigate the long-run consequences of both players using such rules.

We do not model the internal thought processes of the players, and instead phrase our conditions, assumptions, and results solely in terms of players' behavior. In particular, we will not make separate assumptions about how players update their beliefs on the one hand, and how they use their beliefs on the other. (However, the particular rules we construct can be interpreted as an "almost-best-response" to beliefs of the type used in fictitious play.) Consequently, the object of our analysis is the set of "behavior rules", by which we mean maps from observations to actions.

It should be emphasized that how well a behavior rule performs depends on the environment it is in. For example, consider the game of matching pennies: a single player must guess each period whether nature will choose "Heads" (H) or "Tails" (T). He earns a payoff of 1 if he guesses correctly, -1 otherwise. The rule "always guess H" will perform quite well if the environment is one in which nature always plays H. Of course in other environments, this rule will perform quite badly. Implicitly, behavior rules based on learning attempt to "learn" about the environment they are in, so that in the long-run they

[2] This has been shown independently by Monderer, Samet and Sela (1994).

perform well in a broad class of environments. Generally, this at least includes those environments that converge to long-run equilibrium.

One obvious question is whether there are behavior rules that perform well in the long run against all environments. If performing well means optimization against the true environment it is well known that there can be no such rule. Indeed, Nachbar (1993a,b) has refined this result with a counter-example showing that no rule drawn from a sufficiently rich set of rules can be even approximately optimal against all rules in that set.[3] One way to think about what is going on is to begin with Blackwell and Dubin's (1962) observation that a Bayesian optimizer will perform optimally in the long-run against any environment to which positive probability is associated. The problem shown by Nachbar is that such a Bayesian optimizer may through his behavior generate behavior that did not receive positive weight in his own prior. That is, individuals may easily generate behavior more complicated than that they contemplate as possible for their opponents. In particular, as a practical observation, in learning models in which learning rules fail to converge to an equilibrium, as for example in cobweb cycles, the behavior of agents may seem implausibly stupid.[4]

In this paper we lower our sights somewhat, and look for rules that have sensible properties in all environments even though they are not asymptotically optimal in all environments. We begin with the observation that the world may be more complex than players contemplate in their models, and that players are aware of this. What then is a sensible criterion when a very complicated sequence of Heads and Tails has been observed? It may well be that the environment is generated by a complicated chaotic deterministic model, for example, but to figure this out may be difficult or impossible.

[3] Work in progress by Anderlini and Sabourian (1993) may yield such a result in an evolutionary rather than learning setting.

[4] This critique is applicable to any adaptive expectations model where behavior follows an explosive cobweb, as for example in DeCanio [1979]. It is particularly applicable to models where agents persistently predict that behavior at date *t* will be exactly the same as that at the previous date *t*-1 even when that prediction has proved false in every preceding period.

Hence, from the player’s point of view, it may be sensible to view such a sequence as random, and to at least try to play optimally with respect to the frequencies of heads and tails. Put another way, a player may simply choose to ignore the order in which the observations occur, even though this information is potentially useful. This motivates our desiderata that the behavior rule be universally consistent, in the sense that the rule should (asymptotically) ensure that the player’s realized average payoff is not much less than the payoff from playing the best response to the empirical distribution, uniformly over all possible environments.

If players know they are boundedly rational, they may also wish to allow for the possibility that they are playing against opponents who are cleverer than they are. One way that players might do this is to only use behavior rules that guarantee that their realized payoff is not much lower than their minimax payoff. It is fairly easy to see that any universally consistent rule will be “safe” in this sense, since the best response to the any distribution must be at least the minmax.

The “calibration” result of Foster and Vohra (1994) shows that universally consistent rules exist, but since the proof is existential, it does not indicate the forms that such rules might take. This paper shows that a particular randomized version of fictitious play in which actions are played in proportion to their utility with exponential weights (exponential fictitious play) is universally consistent. Moreover, such a policy can be implemented even in an extensive form game in which opponents strategies are not observed.

Beyond this result, we explore the possible long-run outcomes when all players use rules that guarantee they do at least as well as playing a best response to the empirical frequency distribution. This gives rise to the notion of marginal best-response distributions, which are the only points such a learning process can pass through in the long run. We give some examples and results to show what these types of distributions are like.

2. The Model

We begin by considering a single agent. This agent repeatedly chooses a probability distribution a , called a *mixed action*, over a finite space of actions A, and observes an *outcome* in a finite set Y. If $a \in A, y \in Y$, the agent receives a utility of $u(a, y)$. We use Δ to denote the space of probability distributions over a set. If $a \in \Delta(A)$ and $g \in \Delta(Y)$ then the expected utility is also denoted $u(a, g)$. We say that a is an e *-best response* to g if

$$u(a, g) + e \geq u(\tilde{a}, g)$$

for all alternative mixed actions $\tilde{a}$. If $e = 0$, we refer simply to a *best response.*

Since we are considering a repeated situation, we define a *history* as a sequence of actions and outcomes $h = (a_1, y_1, \ldots, a_t, y_t)$. The number of actions (or outcomes) $t(h)$ is called the *length of the history.* The history truncated by one period is $h - 1 = (a_1, y_1, \ldots, a_{t-1}, y_{t-1})$. It is useful also to define the null history h_0 of zero length. The space of all histories is denoted by H. The *outcome frequency distribution* of a non-null history is the empirical probability distribution over outcomes, and is denoted by $\bar{g}(h)$.

The agent chooses a (mixed) *behavior rule*, which is a map from histories to probability distributions over actions $s : H \to \Delta(A)$. An important example of a behavior rule is that of fictitious play: this requires the existence of a probability distribution over outcomes g_0, called the *prior,* and an integer n_0 called the *prior precision* such that $s(h)$ places weight only on actions that are best responses to

$$\frac{n_0}{n_0 + t(h)} g_0 + \frac{t(h)}{n_0 + t(h)} \bar{g}(h),$$

called the *posterior.* We will be particularly interested in rules which depend on the history only through the empirical distribution of outcomes, that is rules of the form

$a:\Delta(Y)\to\Delta(A)$, with the corresponding behavior rule given by $s(h)\equiv a(\bar{g}(h))$. We call these *stationary* rules.

The agent is also faced with an unknown *environment* which is a rule mapping histories to probability distributions over outcomes $r:H\to\Delta(Y)$. In our applications this environment will correspond to the behavior rules of other players. Notice that the outcome cannot depend on the current action taken by the agent. An important example of an environment is the *i.i.d. environment* which is history independent: $r(h)=r(\tilde{h})$ for all pairs of histories $h,\tilde{h}$.

Our interest is in behavior rules that enable an agent to "learn" about an unknown environment. Our goal is to assess behavior rules by how well they perform. There are two performance criterion of interest: long-run performance, and the rate at which the learning behavior rule converges to this long-run payoff. In this paper we will focus solely on the long-run performance. Consequently our criterion for assessing performance will be the time-average payoff with a "long" time horizon. For any given behavior rule/environment pair, we may define a probability distribution over histories $p(s,r)$. The time average realized payoff $U(s,r)[h]=(1/t(h))\sum_{t=1}^{t(h)}u(a_t,y_t)$ is a random variable with respect to this probability distribution in the obvious way. It is also useful to define the optimized payoff against the empirical history to be $\hat{U}(h)\equiv\max_a u(a,\bar{g}(h))$ with corresponding argmax $\hat{a}(h)$.

Definition 2.1: A behavior rule s is e*-consistent* if there exists a $\bar{T}$ such that for any i.i.d. environment r and for any $T\geq\bar{T}$ there is a subset of histories of length T, H_T, with $p(s,r)[H_T]\geq 1-e$ and for all $h\in H_T$

$$U(s,r)[h]+e\geq\hat{U}(h)$$

A behavior rule is *consistent* if it is ε-consistent for every positive ε.

In other words a behavior rule is ε-consistent if in an i.i.d. environment it does about as well as playing a best response against the empirical distribution, or (equivalently for large *T*) the true probability distribution in that environment. Notice that this a bit stronger than the usual notion of consistency in the statistics literature in that the rate of convergence here is independent of the environment r. However, for the multinomial distribution, is well known that the rate of convergence is uniform, and the following lemma is immediate from Chebychev's inequality.

Proposition 2.1: If s is a fictitious play behavior rule, then it is consistent.

The limitation of consistency is that there is no reason the agent should think that he is facing an i.i.d. environment. Indeed, if a game is played between agents who both use fictitious play, for most initial conditions on beliefs (those that do not begin at an exact equilibrium), the resulting environment will not be i.i.d.. More to the point, a consistent behavior rule can be fooled quite badly by a clever opponent. One additional criterion beyond consistency that seems desirable is that of safety: a player should not get significantly less than his minmax payoff in the long run.

Definition 2.2: A behavior rule s is e-safe if there exists a $\overline{T}$ such that for any environment r and for any $T \geq \overline{T}$ there is a subset of histories of length T, H_T, with $p(s,r)[H_T] \geq 1-e$ and for all $h \in H_T$

$$U(s,r)[h] + e \geq \min_g \max_a u(a,g)\,.$$

A behavior rule is safe if it is ε-safe for every positive ε.

Fictitious play is well known not to be safe. Suppose the game is matching pennies, with the agent trying to match the play of the environment. If fictitious play begins with a prior $g_0 = (\frac{1}{1+\sqrt{2}}, \frac{\sqrt{2}}{1+\sqrt{2}})$ and prior precision $n_0 = 1$, and the environment alternates deterministically between heads and tails, starting with heads, then fictitious play always plays the opposite of the environment, and the agent gets -1, considerably less than the minmax of 0 . Indeed, fictitious play can fail in this way, not only against a clever opponent out to trick the agent, but also in a game in which all players use fictitious play: Fudenberg and Kreps (1993) give an example of this sort.

It is easy to see that fictitious play is not the only vulnerable rule. In particular, deterministic behavior rules can be exploited by an opponent who knows the rule, and chooses in each period an action that minimizes the agent's payoff given the action that the agent will play that period. (For example, if the agent uses a deterministc rule in matching pennies, consider the environment which always picks the exact opposite of the choice made by the agent.) This is essentially the point made by Oakes (1985).

There are obviously behavior rules that are safe: playing the maxmin every period is safe, for example. Unfortunately this does not have the minimal learning property of consistency. An obvious question is whether there is any behavior rule that is both safe and consistent. We will find such a behavior rule below, but first it is useful to define a property that combines both safety and consistency:

Definition 2.3: A behavior rule s is e *-universally consistent* if there exists a $\overline{T}$ for any environment r , and for any $T \geq \overline{T}$ there is a subset of histories of length *T*, H_T , such that $p(s, r)[H_T] \geq 1-e$ and for all $h \in H_T$

$$U(s, r)[h] + e \geq \hat{U}(h) .$$

For small ε, ε-universal consistency means doing more-or-less as well as playing a best response to the historical frequency distribution. In effect, the player ignores all information about the order in which the outcomes occur, and the extent to which they might be correlated with his own play.

There are potentially two problems with playing a best response to the frequency distribution: First, it ignores information about the way the agent's play influences the play of the environment. Suppose for example that the game is the Prisoner's dilemma and the environment is one in which the opponent plays tit-for-tat. Then it is universally consistent to cheat all the time (this is a best response to any frequency distribution) but the opportunity to get a higher payoff by cooperating is being ignored.

However, ignoring causality in this way need not be troubling. In an environment where a large number of players interact anonymously either through market prices or through a random matching procedure, the actions of individual players can have essentially no effect on the future of prices or the population distribution of opponents. In such an environment there is no causality running from the agent's action to future outcomes, so such information is irrelevant. We refer to the learning problem in such an environment as a pure forecasting problem.

Note, though, that even with a pure forecasting problem, an agent who plays a best response to the empirical frequency distribution is ignoring the order in which observations occur: For example, in matching pennies if the environment alternates in a deterministic manner between H and T, a best response to the frequency distribution of 1/2-1/2 yields a payoff of 0. This, however, overlooks the opportunity to do even better by guessing correctly every period and getting a payoff of 1.

One desirable property of universally consistent learning rules is that they are safe.

Proposition 2.2: If s is e-universally consistent it is e-safe and e-consistent.

Proof: This follows immediately from $\max_a u(a, \bar{g}(h)) \geq \min_g \max_a u(a, g)$. ❑

Because no deterministic rule is safe, no deterministic rule can be universally consistent. Moreover, simply adding an arbitrary form of noisy mixing does not make a behavior rule universally consistent. Again in matching pennies, suppose that the agent uses a modified version of fictitious play that assigns probability (1-ε) to the action that is the best response to the agent's posterior, and divides the remaining ε probability equally among the other actions. With the prior we gave earlier, and a "malicious" opponent, the agent still plays the "wrong" action with probability (1-ε) in each period, and so the agent's expected average payoff is only 2ε-1, which is less than the minmax value of 0. Intuitively, if the agent's play is very sensitive to small changes in the empirical average, then there are environments where the empirical average is converging, but the agent's play oscillates in such a way that the agent's realized payoff is lower than the best response to the limit of the empirical averages. Conversely, if the agent not only plays a mixed action, but also varies his mixing probabilities "smoothly" with changes in the empirical average then (since the empirical average adjusts at the inverse of the sample size) the agent's play cannot oscillate wildly from period to period. This is the motivation for our restricting attention to smooth behavior rules in the next section, and also for proposition 4.1, which shows that fictitious play performs well along histories where it exhibits "infrequent switches."

It is easy to give an existence proof showing that for every e there is a behavior rule that is e-universally consistent. The idea originates with Foster and Vohra (1993) who use the same idea to establish a stronger property called calibration, introduced by Dawid (1982). The idea is to consider a hypothetical perverse opponent whose objective is to choose a behavior rule that tricks the agent in the sense that $U(s,r)[h]+e \geq \hat{U}(h)$ will fail. This gives rise to a *T*-period zero sum game where the agent's payoff is $U(s,r)[h]-\hat{U}(h)$ and the opponent's payoff is the negative of this amount. The perverse opponent has a behavior rule that yields him the value of this zero sum game, and by the minmax theorem, the agent has a behavior rule that guarantees him this value. To calculate the value, we know that it is at least what the agent gets from playing any behavior rule against the perverse opponents minmaxing behavior rule. In particular, the agent can in each period play a best response to the conditional distribution (given the history) of the perverse opponent's minmaxing behavior rule. This behavior rule for the agent yields approximately zero in a large sample by the weak law of large numbers. Thus we conclude that the value of the game is at worst $-e$, where $e \to 0$ as $T \to \infty$, which is the desired result. Note, though, that to actually find the desired behavior rule requires solving a dynamic stochastic zero sum game with a very long horizon which is computationally impractical.

However, as we will see below, a very simple randomized and '"smooth" variation of fictitious play has the desired property.

3. Cautious Fictitious Play

In light of our observations in the previous section, we are led to consider behavior rules where the agent’s mixing probabilities depend smoothly on the empirical average. If the stationary behavior rule a is a smooth (twice continuously differentiable) e-best response to the average, we say that it represents e*-cautious fictitious play*. We will show that any such rule can be made e'-universally consistent for any given e' by taking e small enough, *if* there are only two outcomes. (Remember that the outcomes correspond to the profile of opponents’ actions in a game.)

If there are more than two outcomes, then we cannot show that e-cautious fictitious play is e'-universally consistent even for very small e. However, we can show that a particular variation on fictitious play called k*-exponential fictitious play* is e-universally consistent. A k-exponential fictitious play is given by specifying fixed weights $w_a > 0$ and using the stationary rule

$$a(\bar{g})[a] \equiv \frac{w_a \exp\left(ku(a,\bar{g})\right)}{\sum_b w_b \exp\left(ku(b,\bar{g})\right)}.$$

Notice that for fixed weights and k sufficiently large, this scheme assures that the agent is playing an e best response to the historical average so that this is indeed an e-fictitious play.[5]

Proposition 3.1:

(a) For all weights w_a and every e' there exists a k such that k-exponential fictitious play is e'-universally consistent.

[5] This is a special form of the exponential weighting considered in Blume’s (1993), (1994) papers on stochastic adjustment. Blume’s papers consider “myopic” adjustments, in the sense that agents in a large population respond to the *current* distribution of opponents’ actions, and do not use past observations in making their choices.

(b) If there are only two outcomes, then for every e' there exists a e such that every e-cautious fictitious play is e'-universally consistent.

To prove the proposition, we use a method from stochastic approximation theory of approximating a system that involves averaging with a differential equation in virtual time. Fix s, r. The equation of motion for the time average of utility is

$$U[h] = \frac{1}{t(h)}\left[u(a_{t(h)}, y_{t(h)}) + (t(h)-1)U[h-1]\right].$$

Abbreviating $U_t \equiv U[h]$, this may also be written as

$$\begin{aligned} U_t - U_{t-1} &= \frac{1}{t}\left[u(a_t, y_t) - U_{t-1}\right] \\ &= \frac{1}{t}\left\{\left[u(a_t, y_t) - u(a_t, g_{t-1})\right] + \left[u(a_t, g_{t-1}) - U_{t-1}\right]\right\} \qquad (3.1) \\ &= \frac{1}{t}\left\{\left[\frac{\P}{\P g} u(a_t, g_{t-1})\left(y_t - g_{t-1}\right)\right] + \left[u(a_t, g_{t-1}) - U_{t-1}\right]\right\} \end{aligned}$$

where the final line makes use of the fact that the per-period utility function is bilinear.

To find a continuous virtual time approximation, consider a piecewise Lipshitz function $a : \Delta(Y) \to \Re^A$ and a piecewise smooth curve $\tilde{g} : [0, t] \to \Delta(Y)$ in the space of probabilities measures over outcomes. The curve $\tilde{g}$ should be thought of as a continuous time approximation to the time average g_t. Let $F_{a,\tilde{g}}$ be a solution to the differential equation analog of (3.1)

$$\dot{F}_{a,\tilde{g}} = \frac{\P}{\P g} u\left(a(\tilde{g}_t), \tilde{g}_t\right)\dot{\tilde{g}}_t + \left[u\left(a(\tilde{g}_t), \tilde{g}_t\right) - F_{a,\tilde{g}}\right] \qquad (3.2)$$

along this curve. To avoid having to keep track of inessential constants that depend only on the payoffs, we use the order notation. We say that a family of bounded random variables $\tilde{r}(x)$ is of *order x*, written $\tilde{r} = O(x)$ if there are constants B, b independent of x and a such that $\sqrt{E|\tilde{r}(x)|^2} \leq Bx$ if $x \leq b$.

Lemma 3.2: For any smooth a , $d > 0$ and $\hat{a}(g) \in \arg\max_a u(a,g)$, there exists a $\overline{T}$ such that for any $t(h') \geq t(h) \geq \overline{T}$

$$\hat{U}(h') - U[h'] = F_{[\hat{a}-a],\tilde{g}}(t) + O(d) \quad \left(= F_{\hat{a},\tilde{g}}(t) - F_{a,\tilde{g}}(t) + O(d)\right)$$

$$F_{[\hat{a}-a],\tilde{g}}(0) = \hat{U}[h] - U[h]$$

for some piecewise linear curve $\tilde{g}$ connecting $\bar{g}(h)$ and $\bar{g}(h')$ with $t = \log\left(t(h')/t(h)\right)$ and $\left\|\dot{\tilde{g}}\right\| \leq 1$.

Proof: in Appendix A. Note that the conclusion of the lemma uses the fact that solutions to the differential equation (3.2) have the property that $F_{\hat{\alpha},\tilde{\gamma}} - F_{\alpha,\tilde{\gamma}} = F_{\hat{\alpha}-\alpha,\tilde{\gamma}}$. (This follows from the linearity of the payoffs in α.) ❑

From lemma 3.2, the problem of universal consistency is reduced to the study of the differential equation

$$\text{(3.3)} \qquad \dot{F} = \frac{\P}{\P g}\left[u(\hat{a},\tilde{g}) - u(a,\tilde{g})\right]\dot{\tilde{g}} + \left[u(\hat{a},\tilde{g}) - u(a,\tilde{g})\right] - F .$$

In the absence of the first two terms, this differential equation is stable, so that the distance between the optimized and actual payoff tends to be reduced. The second term is of order e , if $a(g)$ is an e -best response to g for all g . If the first term were also small, it would follow that the solution to the differential equation would remain uniformly close to zero, which is the desired conclusion.

The first term is the product of the sensitivity of the payoff loss to the opponents' average play, $\frac{\partial}{\partial\gamma}\left[u(\hat{\alpha},\tilde{\gamma}) - u(\alpha,\tilde{\gamma})\right]$, and the rate at which the average is changing; this rate can be viewed as the extent to which the opponent is trying to trick the agent. Since the exact best response $\hat{\alpha}$ and the smoothed response α may differ significantly, the payoff difference between them when being tricked may be quite large. (The fact that α is an ε best response to $\tilde{\gamma}$ only means that the payoff loss is small against distribution $\tilde{\gamma}$.)

The key idea of the proof is that the agent cannot be “substantially tricked” for a long time, as $\hat\alpha$ and α must be nearly the same, except in regions where several actions are nearly indifferent. To prove this we observe that over sufficiently short curves F does not change very much, and there is an obvious bound:

Lemma 3.3: If F solves (3.3) and a is e-cautious fictitious play then

$$F \le \int_0^t \frac{\P\left[u\left(\hat a(\tilde g_t),\tilde g_t\right) - u\left(a(\tilde g_t),\tilde g_t\right)\right]}{\P g}\dot{\tilde g}_t dt + et + (1-t)F(0) + O(t^2)$$

Proof: To show this we integrate (3.3) term by term. The first term is exact. The second term is bounded using the definition of e-fictitious play. The error in the remaining term is bounded by using the fundamental theorem of calculus and the mean value theorem: for some $0 \le t^* \le t$ the error is $\int_0^t [F(t) - F(0)]dt = \int_0^t \dot F(t^*)t\, dt$. By Lemma 3.2 $\dot{\tilde g}$ can be no greater than one in norm, and all remaining terms in (3.3) are also bounded independent of a. We conclude that $\left\|\dot F(t^*)\right\|$ is bounded independent of a, so that the integral is of order t^2 as desired. Note that the result would be completely trivial if not for the fact that $O(t^2)$ means a uniform bound independent of a.

❑

Lemma 3.4: If a is k-fictitious play (or $\#Y = 2$) then $\int_0^t \frac{\P\left[u\left(\hat a(\tilde g_t),\tilde g_t\right) - u\left(a(\tilde g_t),\tilde g_t\right)\right]}{\P g}\dot{\tilde g}_t dt$ depends on $\tilde g$ only through the endpoints of $\tilde g$.

Proof: Since $\hat a(\tilde g)$ is piecewise constant, $\frac{\P}{\P g}u(\hat a(\tilde g),\tilde g) = \frac{d}{d\tilde g}u(\hat a(\tilde g),\tilde g)$ except on the lower dimensional set of discontinuities of $\hat a(\tilde g)$, so $\int_0^t \frac{\P}{\P g}u(\hat a(\tilde g),\tilde g)\dot{\tilde g}dt$ depends only on the endpoints of $\tilde g$.[6] In the case $\#Y = 2$ the integral $\int_0^t \frac{\P}{\P g}u(a(\tilde g),\tilde g)\dot{\tilde g}dt$ is one-dimensional, so path independence is immediate. In the higher dimensional case, the result

[6] Monderer, Samet and Sela (1994) considering continuous time fictitious play, use a similar argument in the proof of their Theorem B.

will follow provided $D_{\tilde{g}}[\frac{\P}{\P g}u(a(\tilde{g}),\tilde{g})]$ is a symmetric matrix. With exponential weighting

$$D_{\tilde{g}(z)}\left[\frac{\P}{\P g(y)}u(a(\tilde{g}),\tilde{g}))\right]=D_{\tilde{g}(z)}\left[\frac{\P}{\P g(y)}\sum_{y'}\sum_a\frac{w_a\exp(ku(a,\tilde{g}))}{\sum_b w_b\exp(ku(b,\tilde{g}))}u(a,y')\tilde{g}(y')\right]=$$

$$D_{\tilde{g}(z)}\left[\sum_a\frac{w_a\exp(ku(a,\tilde{g}))}{\sum_b w_b\exp(ku(b,\tilde{g}))}u(a,y)\right]=$$

$$-\frac{\left(\sum_b kw_b\exp(ku(b,\tilde{g}))u(b,z)\right)\sum_a w_a\exp(ku(a,\tilde{g}))}{\left[\sum_b w_b\exp(ku(b,\tilde{g}))\right]^2}u(a,y)+\frac{\sum_a ku(a,z)w_a\exp(ku(a,\tilde{g}))u(a,y)}{\sum_b w_b\exp(ku(b,\tilde{g}))}$$

which is certainly symmetric.

❑

Finally, the integral over a straight line between the endpoints can be bounded:

Lemma 3.5: For every d and t there exists an e such that if a is e-fictitious play $\int_0^t\frac{\P\left[u(\hat{a}(\tilde{g}_t),\tilde{g}_t)-u(a(\tilde{g}_t),\tilde{g}_t)\right]}{\P g}\dot{\tilde{g}}_t dt\le d$.

Proof: In Appendix B. The basic idea is illustrated in Figure 1.

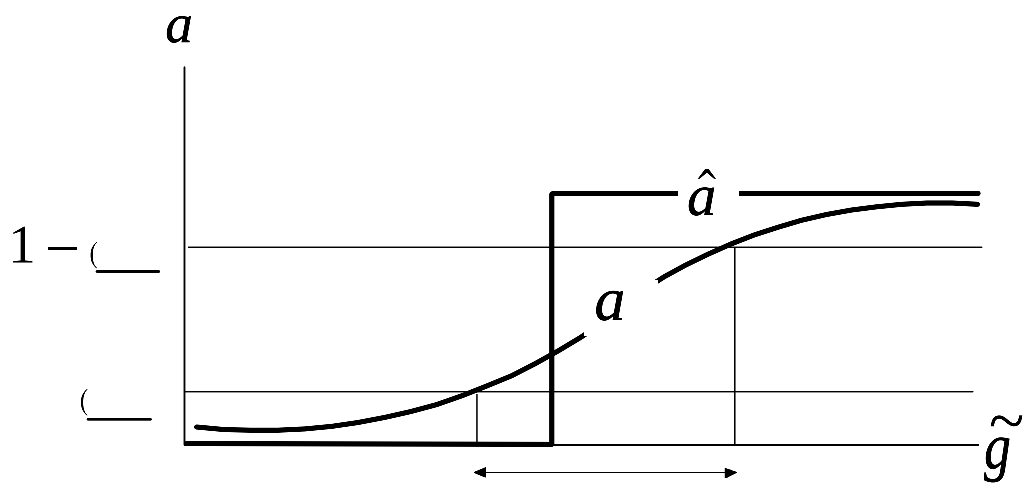

Figure 1

Basically the integral can be broken into two parts: over the region in which $\hat{a}$ and a do not differ by much the integral is small. On the other hand, because a is e-cautious, the size of the region over which $\hat{a}$ and a differ by a great deal can be large only if payoffs are not very sensitive to the outcome, that is $\P\left[u\left(\hat{a}(\tilde{g}_t),\tilde{g}_t\right)-u\left(a(\tilde{g}_t),\tilde{g}_t\right)\right]/\P g$ is small.

❑

Proof of Proposition 3.1: Fix a tolerance e'. First use Lemma 3.5 with $d = e' t / 6$ to pick an $\bar{e}$ with

$$\int_0^t \frac{\P\left[u\left(\hat{a}(\tilde{g}_t),\tilde{g}_t\right)-u\left(a(\tilde{g}_t),\tilde{g}_t\right)\right]}{\P g}\dot{\tilde{g}}_t dt \le e' t / 6$$

along straight lines for all $\bar{e}$-fictitious play. Now use Lemma 3.3 with $e \le \bar{e}, e'/6$ and $O(t) = e$. This gives the long-run bound of F

$$F \le \int_0^t \frac{\P\left[u\left(\hat{a}(\tilde{g}_t),\tilde{g}_t\right)-u\left(a(\tilde{g}_t),\tilde{g}_t\right)\right]}{\P g}\dot{\tilde{g}}_t dt / t + e'/3 .$$

In the two outcome case, let a be any e-cautious fictitious play, and in the general case, let a be a k-exponential fictitious play that is also e-cautious. By virtue of Lemma 3.4, the restriction of the integral to a straight line is not a limitation, so this gives the further bound on F of $F \le e'/2$. Now we may simply use Lemma 3.2 together with Chebychev's inequality to give the desired conclusion.

❑

Since the use of exponential weights may seem somewhat mysterious, it may be useful to look back and see what role they played in the proof. The only use of the exponential weighting was in the proof of Lemma 3.4, where it was used to show that the derivative of the integrand of $\int_0^t \frac{\P}{\P g} u(a(\tilde{g}),\tilde{g})\dot{\tilde{g}}dt$ was symmetric, and consequently that the integral itself is path independent. The remaining parts of the proof show that to a good approximation, the loss from using a in place of $\hat{a}$ is to a good approximation made up of two parts: the loss from the fact that a is only an approximate best-response to the historical average, and the "loss" from being tricked if the actual outcomes are not drawn from the historical average. This latter "loss" may actually be a gain, since the trick may actually favor a over $\hat{a}$, but in any case it is measured by the flow

$$\frac{\P\left[u\left(\hat{a}(\tilde{g}_t),\tilde{g}_t\right)-u\left(a(\tilde{g}_t),\tilde{g}_t\right)\right]}{\P g}\dot{\tilde{g}}\,. \tag{3.4}$$

Notice that the method of proof yields not only an upper bound on the loss, but a lower bound as well. Consequently if the integral of (3.4) is large, the loss will be large. Moreover, if the integral fails to be path independent, there must be closed loops over which it has a non-zero integral: this integral will be positive in one direction and negative in the other. The implication is that a "tricky" opponent can create a continuing loss by moving $\bar{g}$ repeatedly around the loop in the positive direction, and a continuing gain by moving it in the opposite direction. The greater the failure of path independence as measured by the size of this integral, the greater the potential loss or gain.

The conclusion we reach is that if we use the size of the integral (3.4) as a measure of the failure of symmetry, the greater departure from symmetry, the greater the departure from universal consistency. On the other hand, we cannot conclude that a universally consistent strategy dominates an e-fictitious play, since the failure of path independence also guarantees that a "tricky" but "benevolent" opponent could actually provide a higher level of utility than merely the best response to the historical average.

Finally, we should add that the exponential weighting case is the only rule we know of that yields symmetry and so path independence. We do not know whether other such rules exist.

4. Fictitious Play

Before discussing learning in games in which different players are playing particular types of behavior rules, it will be helpful to establish a necessary condition for fictitious play to be consistent.

Definition 4.1: A behavior rule s is e*-consistent against* r if there exists a $\overline{T}$ such that for any $T \geq \overline{T}$ there is a subset of histories of length *T*, H_T, with $p(s,r)[H_T] \geq 1-e$, and $U(s,r)[h]+e \geq \hat{U}(h)$ for all $h \in H_T$.

Given a history *h* we define the *frequency of switches* $h(h)$ to be the fraction of periods *t* in which $a_t \neq a_{t-1}$.

Definition 4.2: A behavior rule s exhibits *infrequent switches* against r if for every e there exists a $\overline{T}$ and for any $T \geq \overline{T}$ there is a subset of histories of length T, H_T, such that $p(s,r)[H_T] \geq 1-e$ and for all $h \in H_T$

$$h(h) \leq e .$$

Proposition 4.1: If s is fictitious play and exhibits infrequent switches against r then for every $e > 0$ it is e-consistent against r.

Remark: This result has been independently obtained by Monderer, Samet, and Sela [1994]. [7]

Proof: Fictitious play plays a best response to the posterior beliefs $\hat{g}(h)$ formed by taking a weighted average of the empirical distribution $\bar{g}(h)$ at the end of the previous period and the prior beliefs γ_0.

[7] To compare their result and ours, note that what we call "infrequent switches" they call "smooth," and that their "belief affirming process" are pairs (σ,ρ) such that each is consistent against the other.

Fix the distribution over histories generated by (σ,ρ), and let h be any history that this distribution assigns positive probability; in the following we will suppress the dependence on h to lighten the notation. Set $T = t(h)$.

Define $\widehat{U}_T \equiv u(\hat{a}(\widehat{g}(h)),\bar{g}(h))$ to be the payoff from playing a best response to posterior beliefs at the end of period T when the opponent's play is given by the empirical distribution, and define $\widehat{\widehat{U}}_T \equiv u(\hat{a}(\widehat{g}(h)),\widehat{g}(h))$, which is the payoff that the player "expects" to get when he believes the distribution of opponents' play is given by $\hat{\gamma}(h)$. Also define $\overline{U}_T(\overline{U}_0)$ recursively by

$$\overline{U}_T = \frac{1}{T}\left(u(\hat{a}(\widehat{g}_T), y_T) + (T-1)\overline{U}_{T-1}\right).$$

This is the expected payoff that would result if the agent's action at each period $t \in \{1,\ldots,T\}$ is a best response to the end-of-period t beliefs $\widehat{g}_t$, averaged with an exogenous "initial utility $\overline{U}_0$. (Of course the agent does not have the information to actually implement this path; we use it only for an upper bound.)

We will show inductively that $U(s, r)[h] \equiv U_T \leq \widehat{U}_T$ and $\widehat{\widehat{U}}_T \leq \overline{U}_T(\widehat{\widehat{U}}_0)$. For $T = 0$, $\widehat{\widehat{U}}_0$ and $\overline{U}_0(\widehat{\widehat{U}}_0)$ are equal by definition, while both U_0 and $\widehat{U}_0$ both equal 0.

Suppose that the inequalities hold at date $T-1$. From the definitions and the linearity of payoffs in the opponent's distribution, we have

$$U_T = \frac{1}{T}\left(u(s_T, y_T) + (T-1)U_{T-1}\right)$$

$$\widehat{U}_T = \frac{1}{T}\left(u(\hat{a}(\widehat{g}_T), y_T)) + (T-1)\widehat{U}_{T-1} + (T-1)[u(\hat{a}(\widehat{g}_T),\bar{g}_{T-1}) - \widehat{U}_{T-1}]\right)$$

$$\widehat{\widehat{U}}_T = \frac{1}{T}\left(u(\hat{a}(\widehat{g}_T), y_T)) + (T-1)\widehat{\widehat{U}}_{T-1} + (T-1)[u(\hat{a}(\widehat{g}_T),\widehat{g}_{T-1}) - \widehat{U}_{T-1}]\right)$$

Moreover, $u(\hat{a}(\widehat{g}_T),\widehat{g}_{T-1}) - \widehat{\widehat{U}}_{T-1} < 0$ from the definition of $\widehat{U}$. This plus the inductive hypothesis that $\hat{\hat{U}}_{T-1} \leq \overline{U}_{T-1}(\hat{\hat{U}}_0)$ establishes $\widehat{\widehat{U}}_T \leq \overline{U}_T(\widehat{\widehat{U}}_0)$. To establish $U_T \leq \widehat{U}_T$, observe

$$
\begin{aligned}
\widehat{U}_T &= \frac{1}{T}\Big(u(\hat{a}(\widehat{g}_T), y_T) + (T-1)\widehat{U}_{T-1} + (T-1)(u(\hat{a}(\widehat{g}_T), \bar{g}_{T-1}) - \widehat{U}_{T-1}\Big) \\
&= \frac{1}{T}\Big(u(\hat{a}(\widehat{g}_T), y_T) + (T-1)\widehat{U}_{T-1} + (T-1)(u(\hat{a}(\widehat{g}_T), \bar{g}_{T-1}) - u(\hat{a}(\widehat{g}_{T-1}), \bar{g}_{T-1}))\Big) \\
&\geq \frac{1}{T}\Big(u(\hat{a}(\widehat{g}_{T-1}), y_T) + (T-1)\widehat{U}_{T-1} + (T-1)(u(\hat{a}(\widehat{g}_{T-1}), \bar{g}_{T-1}) - u(\hat{a}(\widehat{g}_{T-1}), \bar{g}_{T-1}))\Big) \\
&= \frac{1}{T}\Big(u(\hat{a}(\widehat{g}_{T-1}), y_T) + (T-1)\widehat{U}_{T-1}\Big) = \frac{1}{T}\Big(u(s_T, y_T) + (T-1)\widehat{U}_{T-1}\Big)
\end{aligned}
$$

so that the desired inequality follows directly from the inductive hypothesis.

Finally, since $\hat{a}(\widehat{g}_{T-1}) = \hat{a}(\widehat{g}_T)$ except in periods when switches occur, we know that $\lim_{T\to\infty}|U_T - \overline{U}_T(0)| = 0$ along any history with infrequent switches. Moreover, it is clear that $\lim_{T\to\infty}|\overline{U}_T(0) - \overline{U}_T(\widehat{\widehat{U}}_0)|$ and $\lim_{T\to\infty}|\widehat{U}_T - \widehat{\widehat{U}}_T| = 0$, since the averaging asymptotically eliminates the initial utility difference. Consequently $\lim_{T\to\infty}|U_T - \overline{U}_T(\widehat{\widehat{U}}_0)| = 0$. Since we just showed $\widehat{\widehat{U}}_T \leq \overline{U}_T(\widehat{\widehat{U}}_0)$ we must have $\lim_{T\to\infty} U_T - \widehat{\widehat{U}}_T \geq 0$. A similar argument using $U_T \leq \widehat{U}_T$ and $\lim_{T\to\infty}|\widehat{U}_T - \widehat{\widehat{U}}_T| = 0$ shows $\lim_{T\to\infty} U_T - \widehat{\widehat{U}}_T \leq 0$, so that $\lim_{T\to\infty}|U_T - \widehat{\widehat{U}}_T| = \lim_{T\to\infty}|U_T - \widehat{U}_T| = 0$. This yields the conclusion of the proposition.

□

Obviously any behavior rule that is asymptotically the same as fictitious play has the same property.

5. Learning in Games

We turn now to a setting where a number of agents play each other in a game. We will assume that there are N agents $i = 1,2,\ldots,N$, and that each has an action space A_i. The space of outcomes for agent *i* is simply that actions taken by opponents $Y_i = \times_{j \neq i} A_j$, and the payoff function is u_i.

We suppose that every agent is playing a universally consistent learning rule. Denote distributions over action profiles by m, with corresponding marginals over $\times_{j \neq i} A_j$ denoted by m_{-i}. It is convenient also to denote the expected utility from the distribution as $u_i(m)$. Then to a good approximation in the long run each agent will be getting at least the same utility as he could get by playing a best response to the marginal empirical distribution of opponents' play. This motivates the following definition.

Definition 5.1: A (correlated) distribution m has *the marginal best-response property* if for each agent i $\max_{a_i} u_i(a_i, m_{-i}) \leq u_i(m)$. The marginal best-response property is called *exact* if $\max_{a_i} u_i(a_i, m_{-i}) = u_i(m)$.

A *behavior profile b* specifies a behavior rule for each player *i*. Given behavior profile *b*, we can compute the resulting probability distribution *p(b)* over outcomes, and hence obtain probability distributions over the empirical distributions from period 1 to *T* for any *T*. Denote these empirical distributions by $\bar{\mu}^T$, and let ν^T denote the probability distribution over the $\bar{\mu}^T$. In general there is no reason to expect that the ν^T will converge, but since the space of measures on a compact set is compact (in the topology of weak convergence) we know the sequence will have accumulation points. Note moreover that these accumulation point need not be a degenerate measures: for example, the long-run empirical distribution might take on one of two values, depending on the realization of play in the first period. However, if every player is using an ε-universally consistent rule, then except on a set of histories with probability ε the long-run empirical distribution must have (within ε) the marginal best response property. Passing to the limit yields the following proposition.

Proposition 5.1: Consider a sequence of ε-universally consistent behavior rules with ε converging monotonically to 0, and let $T(\varepsilon) \to \infty$ be such that each $T(\varepsilon)$ is greater than the $\overline{T}(\varepsilon)$ in the definition of universal consistency. For each ε, let ν_ε denote the probability distribution over empirical distributions from periods 1 to $T(\varepsilon)$ induced by the associated ε-universally consistent behavior profile. If ν^* is any accumulation point (in the topology of weak convergence) of the ν_ε, then ν^* assigns probability 1 to distributions with the marginal best-response property.

In other words, if agents are universally consistent, in the long run we will see the empirical time average distribution over profiles move only within the set of distributions having the marginal best-response property. This suggests that it is of interest to understand how big and what the set of distributions having the marginal best-response property looks like. Of even greater importance is to understand the utilities that can arise from these distributions. We refer to such utility vectors as *marginal best-response points*. Moreover, if players actually use cautious fictitious play, and not some other universally consistent behavior rule, then the exact same method of proof that establishes that cautious fictitious play is universally consistent shows that players can do no more than e better than playing a best response to the historical frequency. In this case Proposition 5.1 can be strengthened from marginal best-response to exact marginal best response. In other words, we may view the set of marginal best-response points as the set of asymptotic possibilities when players play *some* universally consistent behavior rules, and the set of exact marginal best-response points as the set when they play cautious fictitious play.

It is immediate from the definition that the set of correlated equilibria are a subset of the set of distributions with the marginal best-response property, while the set of Nash equilibria are a subset of the set of distributions with the strict marginal best-response property. We shall see below that the converses of these results are false.

In zero sum games we have a very quick result: since each player is getting at least the minmax, any distribution with the marginal best-response property must give each player at least (and so exactly) the value of the game. This in turn implies the opponent must be playing a minmaxing behavior rule. In other words

Proposition 5.2: If m is has the marginal best-response property in a zero sum game, then $u_i(m) = \min_{a_{-i}} \max_{a_i} u_i(a_i, a_{-i})$, and (m_1, m_2) is a Nash equilibrium.

Note however that this result cannot be strengthened to show that m is actually a Nash equilibrium, that is, that play is independent. (This is true in 2x2 games.) Consider the following "Rock, Scissors and Paper" game

$$\begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

The value of this zero sum game is 0, and the unique equilibrium point is (1/3,1/3,1/3). Consider on the other hand the distribution over profiles given by

$$\begin{pmatrix} 1/9 & 0 & 2/9 \\ 0 & 2/9 & 1/9 \\ 2/9 & 1/9 & 0 \end{pmatrix}$$

It is easily checked that both marginals are (1/3,1/3,1/3), and since the matrix is symmetric, both players get an expected payoff of zero. In other words, this distribution has the exact marginal best-response property, but is not a Nash equilibrium.

Another interesting case to consider is the non-zero sum Shapley game

$$\begin{pmatrix} 0,0 & 0,1 & 1,0 \\ 1,0 & 0,0 & 0,1 \\ 0,1 & 1,0 & 0,0 \end{pmatrix}$$

It has been shown by Shapley [1964] (see also Gaunersdorfer and Hofbauer [1994]) that in this game fictitious play cycles ever more slowly through (UM,DM,DL,ML,MR,UR). Because switching between profiles drops in frequency to zero, the condition of Proposition 4.1 is satisfied, and fictitious play is consistent in this example. We conclude from Proposition 5.1 that when T is large, to a good approximation the empirical time average distribution of profiles (which never puts any weight on the diagonal) is always a distribution with the exact marginal best-response property. Obviously in this example there are many distributions with this property. Note moreover, that this shows that the set of distributions that have the exact marginal best-response property are not a subset of the set of correlated equilibria, as it is known from Foster and Vohra (1993) that in the Shapley game utility remains bounded away from that at any correlated equilibrium.

This leaves the question of whether there are actually correlated equilibria that are not exact marginal best-responses. The following Battle-of-the-Sexes example shows there are:

$$\begin{pmatrix} 1,2 & 0,0 \\ 0,0 & 2,1 \end{pmatrix}$$

The distribution

$$\begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

clearly a correlated equilibrium, indeed, it is a public randomization over Nash equilibria. Given the marginal, player 1 prefers to play D and player 2 prefers to play L. Each receives an expected utility against the marginal of 1 against the correlated equilibrium payoff of 1.5.

One crucial question is whether there are broad classes of games in which the marginal best-response property imposes no restrictions on payoffs, that is, that the set of marginal best-response points are the entire socially feasible individually rational set.

Consider the generic case in which no pair of profiles yield exactly the same utility for all players. In this case extremal points in the socially feasible set can be achieved by only one distribution that places all weight on a single profile. This implies that extremal points are marginal best-response points only if they are Nash equilibrium payoff vectors. Combining this with the obvious fact that the set of marginal best-response point is closed yields the following proposition.

Proposition 5.2: Suppose that no pair of profiles yields exactly the same utility for all players. Then an extremal point that is not a Nash equilibrium is contained in an open set that has no marginal best-response points.

In other words, the set of marginal best-response points is bounded away from the extremal points.

6. Incomplete Observation

We wish to conclude by considering settings such as extensive form games and moral hazard models, in which the player does not actually observe the outcome y, but only a noisy signal that may depend on his own action. A useful example to have in mind is a two-period prisoner's dilemma. If the agent chooses cheat in the first period he will never learn how his opponent will respond to cooperation in the first period. We know, for example, from Fudenberg and Levine [1993] and Fudenberg and Kreps [1993] that in such models learning rules that do not experiment frequently may fail to learn a best response. However, cautious fictitious play experiments infinitely often, so it seems plausible that it could be modified to perform in a universally consistent manner, even with imperfect information.

We will consider the extreme case of the least amount of information that might be available to an agent about the outcome: we assume that the agent does not observe *y* but only his own utility *u*. Notice that exponential fictitious play requires only historical average utilities, and not actual observations of *y*. This motivates the definition of a k -*exponential fictitious play with respect to the utility rule* $\bar{U}^a(h)$ as

$$\alpha(h)[a] \equiv \frac{w_a \exp\left(k\bar{U}^a(h))\right)}{\sum_b w_b \exp\left(k\bar{U}^b(h))\right)}.$$

Now if $\bar{U}^a(h)$ is asymptotically the same as $u(a,\bar{\gamma}(h))$, this rule will have properties identical to those of k -exponential fictitious play.

Consider a long period over which two actions are played with (approximately) fixed positive probabilities. Since the probabilities of the actions fixed and positive the frequency of outcomes conditional on either of the two actions is approximately the same over this period. Notice that this would not be the case if the action probabilities are time dependent: time dependent outcome frequencies can then cause the conditional frequencies to differ between the two actions.

Since each action has the same conditional frequency of outcomes, the only issue is the appropriate assignment of weights to the observations. If we update utility by weighting observations in inverse proportion to the likelihood that the action is taken, then asymptotically the utility average corresponding to each action is based on the same underlying frequency. In other words, if we use the updating rule

$$\overline{U}^a(h) = \begin{cases} \overline{U}^a(h-1) & a_T \neq a \\ \frac{1}{T}\left(\frac{1}{a(h-1)[a]}u(a, y_T) + (T - \frac{1}{a(h-1)[a]})\overline{U}^a(h-1)\right) & a_T = a \end{cases}$$

then universal consistency is achieved despite the fact that only the agent's own utility is observed.

7. Appendix A: Proof of Lemma 3.2

Lemma 3.2 follows directly from the linearity of the differential equation and Lemmas A.1 and A.2 below.

Lemma A.1: If a is smooth, then any $d > 0$, there exists $\overline{T}$ such that for any $t(h') \geq t(h) \geq \overline{T}$

$$\begin{aligned} U[h'] &= F_{a\tilde{g}}(t) + O(d) \\ F_{a\tilde{g}}(0) &= U[h] \end{aligned} ,$$

for some piecewise linear curve $\tilde{g}$ connecting $\bar{g}(h)$ and $\bar{g}(h')$ with $t = \log\left(t(h')/t(h)\right)$ and $\left\|\dot{\tilde{g}}\right\| \leq 1$.

*Proof***:** A standard weak law of large numbers calculation using Chebychev's inequality shows that

(A.1) $$\frac{\sum_{t=T}^{T+lT}(y_t - E(y_t|h-1))}{lT} = O(1/\sqrt{lT}).$$

Similarly for payoffs we have

(A.2) $$\begin{aligned} O(1/\sqrt{lT}) &= \frac{(\sum_{t=T}^{T+lT} u(a_t, y_t) - u(a_t, E(y_t|h-1)))}{lT} \\ &= \frac{\sum_{t=T}^{T+lT} u(a_t, y_t) - u(a_T, \sum_{t=T}^{T+lT} E(y_t|h-1))}{lT} + O(\|Da\| l) \\ &= \frac{\sum_{t=T}^{T+lT} u(a_t, y_t) - u(a_T, \sum_{t=T}^{T+lT} y_t)}{lT} + O(\|Da\| l) + O(1/\sqrt{lT}) \end{aligned}$$

where the second line follows from Taylor's theorem, and the final line by making use of (A.1). Let $\bar{g}(h_T^{T+lT})$ denote the empirical distribution of outcomes between T and $T + lT$. We may rearrange (A.2) as

(A.3)
$$\frac{\sum_{t=T}^{T+lT} u(a_t, y_t)}{lT} = u(a_T, \bar g(h_T^{T+lT})) + O(\|Da\| l) + O(1/\sqrt{lT})$$

Next, we turn to the movement of the empirical distribution itself. We have

$$\bar g(h_{T+lT}) = \frac{T}{T+lT}\bar g(h_T) + \frac{lT}{T+lT}\bar g(h_T^{T+lT})$$

or

(A.4)
$$\begin{aligned}\bar g(h_T^{T+lT}) &= \frac{1+l}{l}\bar g(h_{T+lT}) - \frac{1}{l}\bar g(h_T)) \\ &= \frac{1+l}{l}(\bar g(h_{T+lT}) - \bar g(h_T)) + \bar g(h_T)\end{aligned}$$

Combining (A.3) and (A.4) we have

$$\begin{aligned}U_{T+lT} - U_T &= \frac{l}{1+l}\left(\frac{\sum_{t=T}^{T+lT} u(a_t, y_t)}{lT} - U_T\right) \\ &= \frac{l}{1+l}(u(a_T, \bar g(h_T^{\mathrm{T}+lT})) - U_T) + O(1/\sqrt{lT}) + O(\|Da\| l) \\ &= \frac{l}{1+l}\left[u\left(a_T, \frac{1+l}{l}\bar g(h_{T+lT})\right) - u\left(a_T, \frac{1+l}{l}\bar g(h_T)\right)\right] + u(a_T, \bar g(h_T)) - U_T) \\ &\qquad + O(1/\sqrt{lT}) + O(\|Da\| l) \\ &= \left[u(a_T, \bar g(h_{T+lT})) - u(a_T, \bar g(h_T))\right] + \frac{l}{1+l}(u(a_T, \bar g(h_T)) - U_T) + O(1/\sqrt{lT}) + O(\|Da\| l) \\ &= \left[u(a_T, \bar g(h_{T+lT})) - u(a_T, \bar g(h_T))\right] + l\,(u(a_T, \bar g(h_T)) - U_T) + O(1/\sqrt{lT}) + O(\|Da\| l)\end{aligned}$$

Taking $1/\sqrt{l\overline{T}} \le d, \|Da\| l \le d$ and observing that $d\log(T+lT)/dl = 1$ then yields the desired conclusion.

□

Lemma A.2: If $\hat a(g) \in \arg\max_a u(a, g)$, then

$$\hat U(h') - \hat U(h) = F_{\hat a, \tilde g}(t),$$

for every piecewise linear curve $\tilde g$ connecting $\bar g(h)$ and $\bar g(h')$ with $t = \log(t(h')/t(h))$.

Proof: Follows from the fact that $\hat{a}$ is locally constant and changes only at points of indifference to $\tilde{g}$. This is the virtual time analog of Proposition 4.1, and the interested reader may wish to refer to the proof of that proposition in the text. See also Monderer, Samet and Sela (1994) Theorem B.

□

8. Appendix B: Proof of Lemma 3.5

Lemma 3.5: For every d and t there exists an e such that if a is e-fictitious play

$$\int_0^t \frac{\P\left[u\left(\hat a(\tilde g_t),\tilde g_t\right) - u\left(a(\tilde g_t),\tilde g_t\right)\right]}{\P g}\dot{\tilde g}_t dt \le d$$

Proof: By Lemma 3.4 it suffices to consider lines of the form $\tilde g(t) = \tilde g_0 + tg*$. Define $(\quad (\tilde g) \equiv \hat a(\tilde g) - a(\tilde g)$, and let Φ be the payoff matrix with elements $\Phi_{ay} = u(a,y)$. Then we must evaluate

$$\begin{aligned} &\int_0^t \frac{\P\left[u\left(\hat a(\tilde g_t),\tilde g_t\right) - u\left(a(\tilde g_t),\tilde g_t\right)\right]}{\P g}\dot{\tilde g}_t dt \\ &= \int_0^t \left(\hat a(\tilde g_t) - a(\tilde g_t)\right)\frac{\P\Phi\tilde g_t}{\P g}\dot{\tilde g}_t dt \quad . \\ &= \int_0^t \mathrm{a}(\tilde g_0 + tg*)\Phi g * dt \end{aligned}$$

Since $\sum_a (\quad _a(g) = 0$, for any a we may write

$$(\quad (\tilde g)\Phi g' = (\quad _{-a}(\tilde g)\Phi^a g'$$

where $(\quad _{-a}$ is the vector of all components except a and $\Phi^a_{by} = \Phi_{by} - \Phi_{ay}$. Note that $\Phi^a_{b\cdot}\tilde g$ is the payoff difference between a and b against $\tilde g$. Consequently we may write the integral as

$$\begin{aligned} &\int_0^t \mathrm{a}(\tilde g_0 + tg*)\Phi g * dt \\ &= \int_0^t \mathrm{a}_{-a}(\tilde g_0 + tg*)\Phi^a g * dt \\ &= \int_0^t \sum_{b \ne a} \mathrm{a}_b(\tilde g_0 + tg*)\Phi^a_{b\cdot} g * dt \end{aligned}$$

It is clearly sufficient, therefore, to show that term-by-term that

$$|\int_0^t (\quad _b(\tilde g_0 + tg*)\Phi^a_{b\cdot} g * dt| \le d$$

Because $\tilde g$ is restricted to varying along a straight line, the set on which a particular action is a best response is a (connected) subinterval. Consequently we may break the integral up into an integral over subintervals along which one action remains a best response. Since there are at most as many such subintervals as there are actions, it suffices to prove the desired bound separately in each such subinterval. Let *a* be the best response over some such subinterval: for $\tilde{g}$ in this subinterval either *b* is a best response also, in which case $\Phi^{a}_{b\cdot}\tilde{g} = 0$, or $(\;\; _{b}(\tilde{g}) = a\; _{a}(\tilde{g}) - 0 \geq 0$ and $\Phi^{a}_{b\cdot}\tilde{g} \geq 0$. Consequently $(\;\; _{b}(\tilde{g})\Phi^{a}_{b\cdot}\tilde{g} \geq 0$.

Suppose that there are two points g, g' for which *a* is a best response, and such that $|(\;\; _{b}(g)|, |(\;\; _{b}(g')| \geq (\underline{\quad}$. If a is an e-cautious fictitious play then $e \geq (\;\; (\tilde{g})\Phi\tilde{g} = (\;\; _{-a}(\tilde{g})\Phi^{a}\tilde{g}$, and it follows for all *b* that $e \geq (\;\; _{b}(\tilde{g})\Phi^{a}_{b\cdot}\tilde{g} \geq 0$. Since this bound holds for both $\tilde{g} = g, g'$, and noting that $g = \tilde{g}_{0} + tg^{*}, g' = \tilde{g}_{0} + t'g^{*}$ we may conclude that $|\Phi^{a}_{b\cdot}g^{*}| \leq e/(\underline{\quad}|t - t'|$. In other words, if *b* is played with significant probability over a long subinterval where *a* is a best response, it must yield essentially the same payoff against g^{*} as *a*.

Let s be the total length of time over the subinterval of length t_{a} in which $(\;\; (\tilde{g}_{t}) \geq (\underline{\quad}$. Then

$$\begin{aligned} &|\int_{0}^{t} \mathrm{a}_{b}(\tilde{g}_{0} + tg^{*})\Phi^{a}_{b}g^{*}\,dt| \\ &\leq (t_{a} - s)\underline{\mathrm{a}}\left\|\Phi^{a}\right\| + se/\underline{\mathrm{a}}s \\ &\leq t\,\underline{\mathrm{a}}\left\|\Phi^{a}\right\| + e/\underline{\mathrm{a}} \end{aligned}$$

The desired result now follows by choosing $(\underline{\quad} \leq d/t\left\|\Phi^{a}\right\|$ and $e \leq d(\underline{\quad}$.

□

9. References

Anderlini, Luca and Hamid Sabourian, 1993, Cooperation and Effective Computability, (Cambridge University).

Blackwell, D. and L. Dubins, 1962, Merging of Opinions with Increasing Information, Annals of Mathematical Statistics, 38, 882-886.

Blume, L., 1993, The Statistical Mechanics of Strategic Interaction, Games and Economic Behavior, 4, 387-424.

Blume, L., 1994, How Noise Matters, (Cornell University).

Dawid, A.P., 1982, The Well-Calibrated Bayesian, Journal of the American Statistical Association, 77: 605-613.

DeCanio, Stephen J., 1979, Rational Expectations and Learning from Experience, Quarterly Journal of Economics, XCIII, 47-58.

Foster, D.P. and R.V. Vohra, 1993, Calibrated Learning and Correlated Equilibrium, (Ohio State).

Fudenberg, Drew and David Kreps, 1993, Learning Mixed Equilibria, Games and Economic Behavior, 5.

Fudenberg, Drew and David K. Levine, 1993, Steady State Learning and Nash Equilibrium, Econometrica, 61, 547-573.

Gaunersdorfer, Andrea and Josef Hofbauer, 1994, Fictitious Play, Shapley Polygons and the Replicator Equation, (Universitat Wien).

Jordan, James S., 1993, Three Problems in Learning Mixed-Behavior rule Nash Equilibria, Games and Economic Behavior, 5, 368-386.

Monderer, Dov, Dov Samet and Aner Sela, 1994, Belief Affirming in Learning Processes, (Technion, Haifa).

Nachbar, John H., 1994a, On Learning and Optimization in Supergames, (Washington University of St. Louis).

Nachbar, John H., 1994b, On Computing a Best-Response in a Discounted Supergame, (Washington University of St. Louis).

Oakes, D., 1985, Self-Calibrating Priors Do Not Exist, Journal of the American Statistical Association, 80, 339.

Shapley, Lloyd, 1964, Some Topics in Two Person Games, Annals of Mathematical Studies, 5, 1-28.

Young, H. Peyton, 1993, The Evolution of Conventions, Econometrica, 61, 57-84.