Conditional Universal Consistency[1]

June 30, 1995

revised August 13, 1997

Drew Fudenberg

David K. Levine[2]

Abstract: Each period, a player must choose an action without knowing the outcome that will be chosen by "Nature," according to an unknown and possibly history-dependent stochastic rule. We discuss have a class of procedures that assign observations to categories, and prescribe a simple randomized variation of fictitious play within each category. These procedures are "conditionally consistent," in the sense of yielding almost as high a time-average payoff as could be obtained if the player chose knowing the conditional distributions of actions given categories. Moreover given *any* alternative procedure, there is a conditionally consistent procedure whose performance is no more than epsilon worse regardless of the discount factor. Cycles can persist if all players classify histories in the same way; however in an example, where players classify histories differently, the system converges to a Nash equilibrium. We also argue that in the long run the time-average of play should resemble a correlated equilibrium.

---

[1] The authors are grateful for financial support from NSF grants SBR-9223320, SBR-9223175, SBR-9409180 and the UCLA Academic Senate. This paper benefited from conversations with Glen Ellison and Dean Foster. We would like to thank participants at the Second International Conference on Economic Theory at Carlos III, Madrid for helpful comments.

[2]Departments of Economics Harvard and UCLA.

1. Introduction

In this paper we consider the decision problem of a single player who must choose an action without knowing an outcome chosen either by Nature or by other players. The distribution of these outcomes is governed by an "outcome function" that maps histories (sequences of past actions and outcomes) into current outcomes, but the player does not know the form that this function takes. Thus the player's task is to learn the outcome function through repeated observations.

We have a designed a class of procedures that are a simple randomized variation on fictitious play. We show that the player obtains almost as high a time-average payoff as he could if he divided the sample into categories, and knew in advance both the frequencies within the categories, and which category the current period's outcome was going to be drawn from. We call this property "conditional consistency." We show that this is possible if (i) the categories are chosen with respect to information available at the time decisions are taken, and (ii) the number of categories used grows sufficiently slowly. Moreover, given any learning procedures, we can find a procedure of this type that does no more than epsilon worse regardless of the discount factor; in this sense, there is no need to look outside this class of procedures.

To focus thoughts, it is useful to consider the game "matching pennies," in which the player announces H or T, the outcome is H or T, and the player's payoff is 1 if he guesses correctly, and 0 otherwise. If the player knew what that period's outcome was going to be at the beginning of each period, he could guess correctly all of the time, with

average payoff of 1. Obviously there is no procedure that can guarantee that the player does about this well regardless of the rule that generates the outcomes, since the player's minimax payoff is 0. However, there are procedures that guarantee that the player's average payoff is at least as high as if the player knew the asymptotic frequency of outcome H over the entire infinite horizon, uniformly over all rules for generating outcomes. Blackwell [1956] and Hannan [1957] were the first to design learning rules with this property, which we call "universal consistency."

We do not have time to present a complete review of the many universally consistent rules here, but we should stress one important property these rules share: All universally consistent rules must prescribe randomized play in at least some instances. (Some intuition for this fact can be gained by considering the minimax theorem, and noting that "guessing the frequencies" is a kind of zero-sum game. Alternatively, note that any deterministic procedure can be "defeated" by the outcome function that "predicts" the action the player is about to play and chooses the action that will minimize the player's payoff.) In particular, "classical" fictitious play is not universally consistent. However, Fudenberg and Levine [1995] show that there are simple "smoothed" or randomized versions of fictitious play that do satisfy universal consistency.

The starting point for this paper is the fact that universal consistency is a fairly weak criterion, since it requires only that the players "get the time averages right," and allows them to ignore even simple patterns in the data. For example, if the outcome in matching pennies alternates deterministically HTHT... between H and T, universal consistency only requires the player win half the time, while there are simple universally consistent rules that can win all the time against this pattern of outcomes. The weakness

of universal consistency leads us to develop to the concept of *universal conditional consistency*, which is the topic of this paper. Under this criterion, we divide the player's data into subsamples, and ask if the player's payoff is as high as if he knew the frequencies of the subsamples and was told each period which subsample that period's outcome was going to be drawn from. For example, we could divide the sample into all the observations in which the previous outcome had been H, and all those in which the previous observation had been T (plus the first observation). In the example of deterministic alternation, this would mean that the outcome in the first subsample would always be T, and in the second subsample it would always be H. Thus, when faced with this simple deterministic pattern, the only way the player could do as well as he could by knowing the subsample frequencies in advance would be to win most of the time.

Arbitrary methods of dividing the sample into subsamples do not make sense. We focus on *classification rules* which assign outcomes to categories based only on information available prior to the observation being made, namely the past history of actions and outcomes, the action chosen immediately prior to the outcome in question, and calendar time. Whether it is in fact possible to be universally conditionally consistent with respect to such a scheme of dividing the sample depends on the number of categories into which outcomes can be assigned. For example, we could assign the outcome in each time period a different category, in which case universal conditional consistency would be the same thing as optimization against the actual realized sequence of outcomes, which we already observed is impossible.

Given a categorization rule, at any moment there is frequency data for certain categories. For each such category, we can compute the best response(s) to the historical

frequency of opponent's play in that category, as in the process of fictitious play. We can also consider slightly randomized or "smoothed" fictitious play as we did in our [1995] paper. We define a conditional smooth fictitious play as a particular randomization between smoothed fictitious plays corresponding to the different categories for which observations are available. Our main results are that these conditional smooth fictitious plays have two important properties: First, they are universally conditionally consistent. This shows that they have good asymptotic properties. But more striking, given any learning rule, with or without good asymptotic properties, we can find a conditional smooth fictitious play that does almost as well regardless of the discount factor; that is, over any short- medium- or long- run, and regardless of the rule generating the observations. In other words, we do not have to give up much short-term performance to assure good asymptotic performance, even in the worst case.

We also discuss the issue of which categorization rules are sensible. Most obvious examples of categorization rules are finite ones. For example, if a player suspects that the outcome may be following a deterministic two-cycle, he can use two categories according to whether the period is odd or even. Or, if the player is concerned that the outcome may be generated by a first-order Markov process, he could use the previous period's outcome to define categories. While we have no formal results on "optimal categories, " we do have some observations to make about category selection. Most important is the observation that, if it turns out a player's own play is a good prediction of the outcome, he can make use of this information to be "calibrated" in the sense of incorporating any information implicit in the player's own intended play. If a player chooses to use a calibrated rule, he needed not actually condition on how own play, provided that it does

not turn out to be a good prediction of the outcome. Nevertheless, if all players in a game use calibrated rules, then the time-average of play must converge to the set of correlated equilibria. Since we argue that there is no need to give up asymptotic performance to achieve short-term goals, this is, in our view, a strong argument in favor of the idea that the long-run time averages should lie near the set of correlated equilibria.

This paper is closely connected to several lines of research. Foremost is the literature on universal consistency, beginning with Blackwell [1956] and Hannan [1957], and continuing through Banos [1968], Megiddo [1980], Fudenberg and Levine [1995] and Auer, Cesa-Bianchi, Freund and Schapire [1995]. This work is closely connected to work in the computer science literature by Vovk [1990], Desantis, Markowski and Wegman [1992], Feder, Mehrav and Gutman [1992], Kivinen and Warmuth [1993], Chung [1994], Littlestone and Warmuth [1994] and Freund and Schapire [1996] on "worst case analysis."

A second related line is in the papers by Aoyagi [1994] and Sansino [1994] on extending the model of fictitious play to allow players to try to detect cycles; we discuss these papers in Section 5.

The third related line of research is the papers on calibrated beliefs by Foster and Vohra [1993,1994]. They consider the problem of a forecaster who each period must issue a forecast, that is, a probability distribution over an exogenous set of outcomes such as the weather. The forecaster is said to be calibrated if in the asymptotic limit, the empirical frequencies in periods where each particular probability distribution is forecast exactly match the forecast probabilities. That is, exact calibration requires that (asymptotically) it should have rained in exactly 1/3 of the periods in which the forecast

probability of rain was 1/3, and so on. Foster and Vohra give an algorithm that ensures that beliefs are approximately calibrated, based on tracking the historical frequencies for each (of finitely many) possible probability announcements.

The type of calibration considered by Foster and Vohra is relatively naïve. Kalai, Lehrer and Smorodinsky [1996] use Dawid's [1982] stronger notion of calibration with respect to checking rules, which can be viewed as a generalization of our conditioning scheme. They show that calibration with respect to all checking rules is equivalent to the merging of opinions, and since the merging of opinions with respect to all models is impossible, so is calibration with respect to all checking rules. One interpretation of our results is that calibration is possible provided that the set of checking rules is sufficiently small.

Foster and Vohra go on to consider play in games when each player's behavior rule is to play a best response to calibrated beliefs. This guarantees that each action is a best response to the frequency distribution of outcomes in those periods in which it was chosen, which is exactly the requirement (in this case) of universal conditional consistency. Moreover, as Foster and Vohra show, it assures that in the long run if all players play this way, the time average of play converges to a correlated equilibrium.

Foster and Vohra's results provide an argument in favor of the idea that in the long run the time-average of play should resemble a correlated equilibrium. The results here generalize[3] those of Foster and Vohra, and provide a stronger argument in several

[3] They are not a strict generalization, because we deal only with games, and not with other types of forecasting problems.

respects. First, the behavior rules we construct are simpler[4], and hence more easily interpretable and more straightforward to implement. Moreover, we strengthen their results by considering the case of discounted payoffs as well as time averaging. Finally, we do not require that players actually condition on their own actions, but only that they are prepared to so if evidence suggests so doing would be beneficial.

2. The Model

Each period t = 1, 2,..., the player chooses an action *a* from a finite set of actions *A*, then observes an outcome $y \in Y$, a finite set. The utility from the action *a* when the outcome is *y* is denoted by $u(a, y)$. The space of probability distributions over a (finite) set *S* is denoted by $\Delta(S)$. Mixed actions are denoted by $\alpha \in \Delta(A)$, and mixed outcomes by $\gamma \in \Delta(Y)$. We will write $u(\alpha, \gamma) = \sum_{a,y} u(a, y)\alpha(a)\gamma(y)$ for the expected utility to mixed actions and outcomes. It is convenient to define $\overline{U} = \max_{a,y}|u(a, y)|$ to be the greatest difference in utility between any action/outcome pair.

A strategy for the player can depend only on the information available to him when he moves, namely the past values of his own play and the outcome. A history of play for the player is denoted by $h_t = (a_1, y_1, \ldots, a_t, y_t)$, the space of all histories of play is denoted by *H*. A (behavior) strategy for the player is a map $\sigma: H \to \Delta(A)$, while an outcome function is a map $\rho: H \to \Delta(Y)$. Each strategy-outcome function pair induces a stochastic process over action/outcome pairs, where given the history h_{t-1} the conditional probability of a_t, y_t is $\sigma(h_{t-1})[a_t]\rho(h_{t-1})[y_t]$. In other words, the player and nature play

---

[4] Fudenberg and Levine [1996] and Hart and Mas-Colell [1996] also provide simpler algorithms.

independently, and must base their play only on the history of actions and outcomes. In some interpretations, the outcomes may be chosen by other players rather than by nature.

We are also interested in classifying observations into subsamples. We suppose that there is given *a priori* a countable collection of categories Ψ into which observations may be placed. A *classification rule* is a map $\hat{\psi}: H \times A \to \Psi$.[5] The interpretation is that prior to observing y_t the player knows h_{t-1}, a_t, and must choose a category $\hat{\psi}(h_{t-1}, a_t)$ based only on this information.

Fix a classification rule $\hat{\psi}$. Given a history of actions and outcomes h_t, we define $n_t(\psi)$ to be the number of times that the category ψ has been observed, and $P_t(\psi)$ to be the vector whose components are the number of times that each outcome has occurred in the subsample corresponding to ψ. Thus $P_t(\psi) / n_t(\psi) \in \Delta(Y)$ represents the empirical distribution of outcomes conditional on the category ψ. For example, the category might correspond to the previous period's outcome, so that the distribution in question is simply the empirical distribution of outcomes $P_t(y) / n_t(y)$ conditional on the previous period's outcome having been *y*.[6] It will also be convenient to define $e(y)$ to be the unit vector corresponding to the outcome y.

Notice that although there are countably many categories, for a fixed horizon *t* there are only finitely many possible histories. It follows that there is a minimal finite subset $\hat{\Psi}_t \in \Psi$ to which all observations taken through time *t* must belong. We refer to

---

[5] For notational simplicity we limit attention to deterministic classification rules. The use of random classification rules would lead to notional changes only, provided that Assumption 1 below is satisfied almost surely.

this subset as the set of effective categories, and let m_t denote the number of effective categories. Our basic assumption is that the number of effective categories does not grow too rapidly.

Assumption 1: $\lim_{t\to\infty} m_t / t = 0$.

We consider two different criteria for the performance of strategies. First we consider the time average utility the player receives, compared to how much he can get *ex post* if he is restricted to strategies that are measurable with respect to the same classification rule. If we define the most utility that is possible against the mixed outcome γ as $V(\gamma) \equiv \max_a u(a,\gamma)$, the most utility per period that can be achieved *ex post* using a single action in the subsample ψ is

$$V\left(\frac{P_t(\psi)}{n_t(\psi)}\right)$$

Denote the total utility received in the subsample ψ by $u_t(\psi) = \sum_{\tau\in\psi} u(a(\tau)$. For each subsample we define the difference between the utility that might have been and the utility that actually was as

$$c_t(\psi) = \begin{cases} n_t(\psi)V\left(\frac{P_t(\psi)}{n_t(\psi)}\right) - u_t(\psi) & n_t(\psi) > 0 \\ 0 & n_t(\psi) = 0 \end{cases}$$

We define the total cost to be $c_t = \sum_{\psi\in\hat{\Psi}_t} c_t(\psi)$.

Our criterion for successful long run learning relative to the rule $\hat{\psi}$ for choosing subsamples is that the time average cost c_t / t should be small.

---

[6] The rules considered by Ayoyagi and Sonsino correspond to categorization by the opponent's play in the

Definition 1: A strategy $\sigma: H \to \Delta(A)$ is ε*-universally consistent conditional on* $\hat{\psi}$ if for every outcome function ρ, $\limsup_{t\to\infty} c_t / t \le \varepsilon$ almost surely with respect to the stochastic process induced by σ, ρ.

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

As with universal consistency, universal conditional consistency measures the rule's effectiveness in optimizing against an exogenous distribution of opponents' play. However, it does *not* require that players experiment in order to learn how their opponent's play varies with their own, nor does it require that the rule do a good job of influencing their opponents' play. Thus the condition seems most natural in settings of anonymous random matching. To illustrate this point, consider a repeated Prisoner's Dilemma game, with actions "C" (cooperate) and "D" (defect). Since "D" is a strictly dominant strategy, it maximizes player 1's expected payoff in the current period after any history, so the rule "always play rule D" is universally conditionally consistent with respect to *any* categorization scheme. Of course, if player 2's strategy is Tit-for-Tat, then player 1's time average payoff would be much higher if he always played C.

Second, we wish to consider a discounted criterion. Since strategies may work well or poorly in the short run depending upon how well the player guesses ρ, we cannot hope for a criterion like universal consistency to be satisfied in the first few periods. However, we can compare two particular strategies. We will allow time varying discount

---

"recent" past.

factors, so we will say that $\{\beta_t\}$ is a *sequence of discount factors* if for each *t*, β_t is a strictly positive number, the sequence $\{\beta_t\}$ is strictly decreasing and $\sum_{t=1}^{\infty}\beta_t = 1$.

Definition 2: A strategy $\hat{\sigma}$ is ε *-as good as* a strategy σ if for all sequences of discount factors $\{\beta_t\}$, all histories *h*, and al outcome functions ρ,

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

(Note that this condition is not probabilistic; it must hold over all histories.)

3. Conditional Smooth Fictitious Play

We begin by introducing a smoothing of the best response function which takes on interior values (that is, strictly prefers mixed strategies) and is continuous. As noted in the introduction, all universally consistent rules must prescribe randomized play in at least some instances. Some intuition for this fact can be gained by considering the minimax theorem, and noting that "guessing the frequencies" is a kind of zero-sum game. Alternatively, note that any deterministic procedure can be "defeated" by the outcome function that "predicts" the action the player is about to play and chooses the action that will minimize the player's payoff. In particular, "classical" fictitious play is not universally consistent. Further discussion of why randomization is critical to universal consistency can be found in Fudenberg and Levine [1995]; see Fudenberg and Levine [1997] for descriptive/behavioral motivations for considering randomized learning rules.

Let $v:\text{int}(\Delta(A)) \to \Re$ be any smooth strictly differentiably concave function satisfying the boundary condition that as γ approaches the boundary of $\Delta(A)$ $\|Dv\| \to \infty$. Define $\|v\| = \sup_{\alpha}|v(\alpha)|$, and suppose that $\|v\| \le \overline{U}$. For example, the

function $v(\alpha) = -(1/\kappa)\sum_a \alpha(a)\log\alpha(a)$, discussed more extensively below, satisfies these properties. For any such function the function

$$V^v(\gamma) = \max_\alpha\left(u(\alpha,\gamma) + v(\alpha)\right)$$

will be smooth and convex and satisfy $\left|V(\gamma) - V^v(\gamma)\right| \le \|v\|$. Note the convenient fact that $\left|V^v\right| \le 2\overline{U}$. In other words, if *v* is small, V^v is a smooth approximation of *V*. We also define $\alpha^v = \arg\max_\alpha(u(\alpha,\gamma) + v(\alpha))$. This is unique since *v* is strictly concave, and by the boundary condition it is interior to the simplex.

There are several useful properties of V^v, α^v which we now summarize. First, since *v* is strictly *differentiably* concave, the implicit function theorem implies that α^v is itself smooth. Moreover, by the envelope theorem $DV^v(\gamma) = u(\alpha^v(\gamma),\cdot)$. This implies that V^v is smooth, and since the domain is compact that the second derivative is bounded. We can apply Taylor's theorem to conclude that there is a constant B^v such that

$$\left|V^v(\gamma') - V^v(\gamma) - DV^v(\gamma)(\gamma' - \gamma)\right| \le B^v\left|\gamma' - \gamma\right|^2.$$

When there is only a single category, the historical frequencies in that category are the same as the overall historical frequencies, given by P_{t-1} / n_{t-1}. The procedure of playing at time *t* a best response to these frequencies is called *fictitious play*. Of greater interest in the current context is the procedure of *smooth fictitious play* that is given by $\alpha^v\left(P_{t-1} / n_{t-1}\right)$.

It may be useful to consider a specific example. Suppose as above that

$$v(\alpha) = -(1/\kappa)\sum_a \alpha(a)\log\alpha(a).$$

Then

$$u(\alpha,\gamma)+v(\alpha)=u(\alpha,\gamma)-(1/\kappa)\sum_a \alpha(a)\log\alpha(a),$$

and the first order condition for a maximum is

$$u(a,\gamma)-(1/\kappa)(\log\hat{\alpha}^v(a)+1)-\lambda=0.$$

where λ is the Lagrange multiplier corresponding to the constraint that the probabilities $\alpha(a)$ must sum to one. We may easily solve these equations to find

$$\alpha^v(a)=\frac{\exp\kappa(u(a,\gamma)}{\sum_{a'}\exp\kappa(u(a',\gamma)}.$$

Substituting P_{t-1}/n_{t-1}, we find

$$\alpha^v(a)=\frac{\exp\kappa(u(a,P_{t-1}/n_{t-1})}{\sum_{a'}\exp\kappa(u(a',P_{t-1}/n_{t-1})}$$

This is exactly the logistic fictitious play introduced in Fudenberg and Levine [1995].[7] In the case of a single category, this method of play is ε-universally consistent, where $\varepsilon\to 0$ as $\kappa\to 0$.[8]

Because P_0 is null, we are left with flexibility on how design a fictitious play in the initial period. The procedure we shall use is to assume that in each category there is an initial "prior" sample $P_0(\psi)$, which may contain fractional observations. The size of the prior sample is defined as $n_0(\psi)\equiv\sum_y P_0(\psi)(y)$. While we do not require that the size of the prior be the same for every category, we do assume that there is a uniform bound.

Assumption 2: $\sup_{\psi\in\Psi} n_0(\psi)<\infty$.

We also define the *augmented sample,* $\tilde{P}_t(\psi) = P_t(\psi) + P_0(\psi)$, with corresponding sample size $\tilde{n}_t(\psi) = n_t(\psi) + n_0(\psi)$.

In the general case of multiple categories, there is a different smooth fictitious play

$$\alpha^{\nu}\left(\frac{\tilde{P}_{t-1}(\psi)}{\tilde{n}_{t-1}(\psi)}\right)$$

corresponding to the subsample for each category. If the categorization rule picks a single category at time *t*, so that $\hat{\psi}(h_{t-1}, a)$ does not depend on *a* but only on the history h_{t-1}, then we may define a *categorical smooth fictitious play* by

$$\sigma^{\nu}(h_{t-1}) = \begin{cases} \alpha^{\nu}\left(\frac{\tilde{P}_{t-1}(\hat{\psi}(h_{t-1}, a))}{\tilde{n}_{t-1}(\hat{\psi}(h_{t-1}, a))}\right) & t > 1 \\ \alpha^{\nu}\left(\frac{\tilde{P}_{t-1}(\psi_0)}{\tilde{n}_{t-1}(\psi_0)}\right) & t = 1 \end{cases}.$$

In other words, the strategy should be based only on the subsample corresponding to the category currently selected. In the case where the category is endogenous, in the sense that it depends on the action chosen, there is a kind of fixed point problem: the category determines the probability distribution over actions by a smooth fictitious play, but on the other hand, the probability distribution over actions determines which category is actually picked. We refer to this as the *calibrated case*.

In the calibrated case, we define categorical smooth fictitious play by solving this fixed-point problem. Set $\tilde{P}^a_{t-1} = \tilde{P}_{t-1}(\hat{\psi}(h_{t-1}, a)), \tilde{n}^a_{t-1} = \tilde{n}_{t-1}(\hat{\psi}(h_{t-1}, a))$ to be the cell counts

---

[7] We originally called this "κ -exponential fictitious play."

and totals corresponding to each different action. Define the matrix of different smooth fictitious plays corresponding to different actions (and implicitly, different subsamples)

$$A^{\nu}(P_{t-1}, n_{t-1}) = \left\{ \alpha^{\nu}\left(\frac{\widetilde{P}^{a}_{t-1}}{\widetilde{n}^{a}_{t-1}} \right) \right\}_{a \in A} .$$

Notice that by assumption on *v*, the α^{ν}'s are strictly positive, and hence so is the matrix A^{ν}. It follows from the Perron-Frobenius theorem (see, for example, Gantmacher [1959]) that the equation $\alpha = \alpha A^{\nu}(P_{t-1}, n_{t-1})$ has a unique solution. This unique solution is denoted $\sigma^{\nu}(h_{t-1})$, and is called *categorical smooth fictitious play.*

We can now state our main results, deferring the proofs to section 6 below.

Proposition 1: The categorical smooth fictitious play σ^{ν} is $2\|v\| + \frac{B^{\nu}}{1 + \inf_{\psi \in \Psi} n_0(\psi)}$ universally conditionally consistent.

Notice that as $\|v\| \to 0$, we must have the maximum norm of the second derivative of V^{ν}, which we denote by B^{ν}, converging to infinity. Consequently, the smaller we take $\|v\|$, the larger we must take the number of prior observations $\inf_{\psi \in \Psi} n_0(\psi)$ if the approximation error in the definition of universal conditional consistency is to be reduced.

Proposition 2: For any pure[9] strategy s and any ε there exists a categorical smooth fictitious play σ^{ν} that is ε-as good as *s*.

---

[9] The assumption that *s* is pure simplifies the notation; the theorem can easily be extended to incorporate mixed strategies as well.

In other words, σ^{ν} is almost as good as *s* and is also universally consistent (by proposition 1). Note that it is easy to construct a rule that is universally consistent and does as well as *s* for a fixed per-period discount factor δ: simply play *s* until enough time has gone by that nothing much matters given the δ, then switch to something universally consistent. The reason that this result is not trivial is that the definition of "as good" requires that σ^{ν} yield about as high a present value for *any* sequence of decreasing weights on successive time periods. For this reason, the rules that we construct in the proof of proposition 2 have the property that the amount of time needed to achieve a given degree of universal consistency is fixed. Thus, if *s* is not nearly universally consistent, there will be discount factors δ sufficiently close to 1 and outcome functions for which σ^{ν} will substantially outperform *s* in discounted present value. In other words, the uniformly small sacrifice ε provides real insurance against bad outcome functions.

To illustrate proposition 2, consider the particular rule of playing a best response to the previous period's outcome and Jordan's [1993] 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; the same is true for a smooth fictitious play. 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 (in the time average sense, even) than a player using a smooth fictitious play.

The reason that best response does better is that it guesses correctly both that opponents' last period play is a good predictor of this period's play, and that the correlation is positive. If in fact the correlation was negative, so for example the opponent alternated deterministically between H and T, then best response would do considerably worse than a smooth fictitious play. The basic idea of Proposition 2 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 period's 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.

4. Use of Categorization Rules

So far, we have held fixed the particular categorization rules. We now examine the issue of which categorization rules make sense. Most obvious examples of categorization rules are finite ones. For example, if a player suspects that *y* may be following a two-cycle, he can use two categories according to whether the period is odd or even. He could also use a number of categories corresponding to the number of

outcomes, and categorize current outcomes according to the outcome in the previous period. This type of scheme is considered by Aoyagi [1994] and Sonsino [1995]. This method has the additional advantage of doing well when the opponents play follows cycles that are growing in length over time. This can be the case when both player use exponential fictitious play: when the same outcome occurs many times in a row, prediction based solely upon the last period will err only on the infrequent occasions when there is a switch from one outcome to another.

To understand why particular categorization rules may or may not be useful, we analyze play between a fixed pair of in a specific example. As we noted earlier, our consistency concepts are more compelling with anonymous random matching than in a setting where the same agents face each other repeatedly, so the following should be taken as a possible suggestion of what one might find in richer models, and not as a self-contained analysis. Consider then the Shapley game, a two player, three action per player game with payoffs of the form

	L	M	R
U	ε,ε	0,1	1,0
M	1,0	ε,ε	0,1
D	0,1	1,0	ε,ε

where $\varepsilon > 0$. We consider first, as did Shapley, the case where $\varepsilon = 0$. Suppose that both players use a single category, and given the history in that category they use logistic fictitious play, for example. From results of Benaim and Hirsch [1994] we know that

play can asymptote approximately to any stable best-response cycle. In this game, such a cycle begins at L,,M. Player 1 then wishes to switch to D. At D,M, player 2 wishes to switch to L, then from D,L to M,L to M,R to U,Rthen back to the start at U,,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.

This is reminiscent of Aoyagi's observation that players who play exact best responses to their beliefs and asymptotically condition in the same way as each other will end up following the dynamics of a fictitious play in each category separately. Intuitively, the use of a common categorization scheme is much like a public randomization between initial conditions of the fictitious play.

Why should a player wish to use a calibrated rule? This is most easily illustrated in the case $\varepsilon > 0$. In this case we have

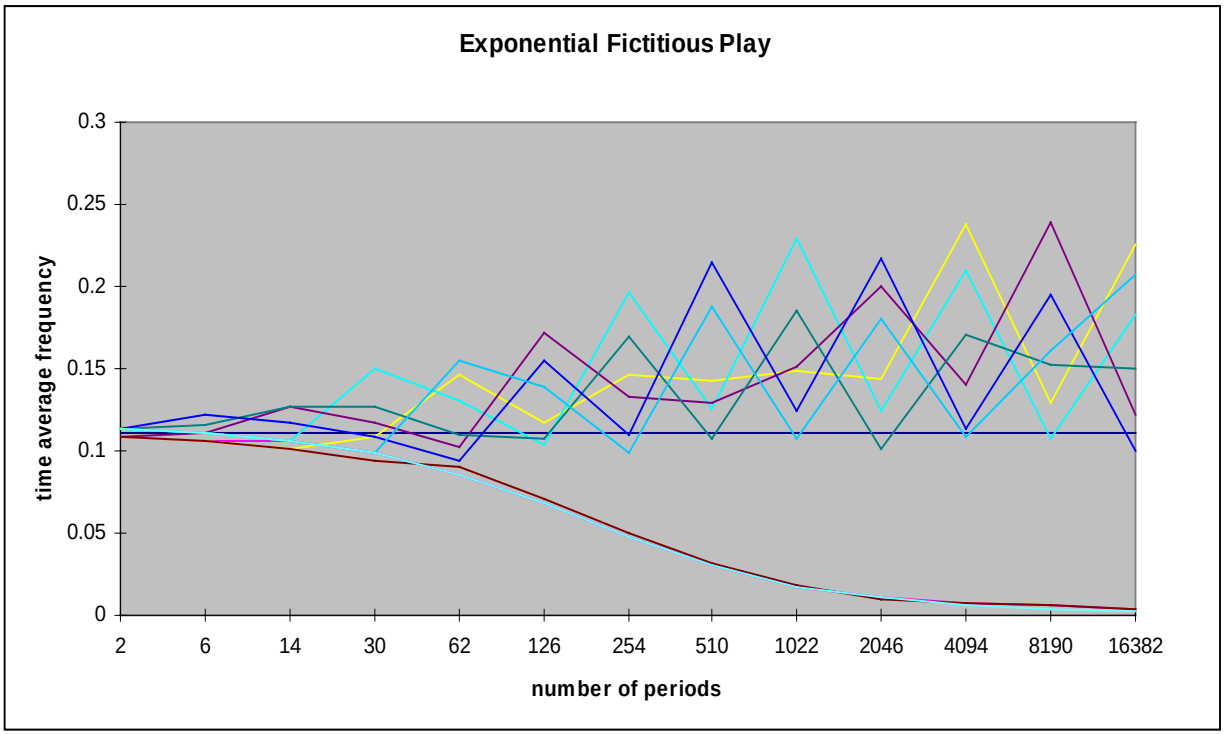
Proposition 3: The Shapley game with $\varepsilon > 0$ there is a unique correlated equilibrium ; it is the Nash equilibrium of equal weight on all outcomes.

Proof: Consider the first row, and suppose that the equilibrium weights are $\rho_{LL}, \rho_{LM}, \rho_{DM}$. If these are not all zero, then U must be a best response for player 1 against these weights. However, D is a strictly better response unless either the weights are all equal, or $\rho_{DM} > \rho_{UM}$. In a similar way, examining the final column, and making a similar argument for player 2, we see that either all weights in this row are equal (possibly to zero) or that $\rho_{MR} > \rho_{DM}$. Continuing around the Shapley cycle in this way, we see that we get a contradiction unless all the weights are equal, which corresponds to the unique Nash equilibrium.

❑

On the other hand, while we have not been able to give an algebraic proof that the Shapley cycle remains locally stable when $\varepsilon > 0$, simulations[10] with ordinary and exponential fictitious play for $\varepsilon < 1/3$ always give rise to cycles even for initial conditions that begin near the Nash equilibrium. A typical Monte Carlo simulation of the time average of the joint distribution of play is shown below, where $\varepsilon = 1/4$ and players use smoothing function of the form $v(\alpha) = -(1/\kappa)\sum_a \alpha(a)\log\alpha(a)$, where $\kappa = 40$ is shown below.

[10] The simulations were done using Excel 7.0 on a Dell XPS 100c with a bug-free Pentium chip. The spreadsheets used for the simulations are available electronically on the World Wide Web at http://levine.sscnet.ucla.edu/.



As can be seen the initial condition is nearly Nash. The three lines going to zero are the frequency of the three diagonal outcomes: asymptotically, the diagonal receives practically no weight. This is true of all of the simulations.

We conclude from proposition 3 that the cycle does not contain any correlated equilibria. It follows from the definition of correlated equilibrium that at each moment of time that one of the two players has received less utility in time average than he could have received if he had conditioned on his own action. However, this player could have guaranteed a better result by using a calibrated strategy.

It is worth pointing out that if instead of using a calibrated strategy, players condition on their opponent's previous action, play in simulations converges to the Nash

equilibrium.[11] However, as we already argued above, if players condition on both their opponent's and their own previous period play, then we simply get a more complicated variation of the Shapley cycle. The key is that if the long-run average play is not converging to a correlated equilibrium, then players will benefit from using a calibrated rule.

5. Calibrated Play and Correlated Equilibrium

In the previous section we argued that either play converges to a correlated equilibrium, or players have an incentive to use calibrated rules. On the other hand, if players all use universally calibrated rules, the long run average of play must in fact converge to a correlated equilibrium.

Suppose in fact that players do indeed use calibrated rules in the sense that they condition on their own anticipated action. By this we mean that no category can ever be assigned two observations in which the player's own action is different. In this case, it is immediate that the empirical joint distribution of action profiles must come to approximate a correlated equilibrium, 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

---

[11] 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.

categorize periods according to forecasts of opponents' play, the frequency distribution of actual opponents' play converges to the forecast. This implies that actions are *calibrated* in the sense 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 on their own actions; it is sufficient that they condition on their own pairs of actions. By this we mean that no category can ever be assigned two periods with more than two distinct values of the players 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.[12] 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.

---

[12] 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 is a worst response to the distribution of outcomes conditional on that action.

Consider once again the three player, two action per player game of matching pennies introduced by Jordan [1993]. 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,[13] 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, whether or not players use categorization rules in lock step.

Finally, we should observe that the long-run behavior of the exponential fictitious play is quite different from that of calibrated play: in the Shapley example with $0 < \varepsilon < 1/3$ calibrated play converges globally to the unique correlated equilibrium

---

[13] 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 in his example by keeping a separate distribution for

(which places weight 1/9, 1/9, 1/9 on the diagonal). In contrast, exponential fictitious play does not converge, and asymptotically places almost no weight on the diagonal.

6. A Bound for Universal Conditional Consistency

We turn now to proving the two main results. It is convenient to begin by imagining that, in a sense to be made precise, the player gets to play against his prior as well as the actual sample. More precisely, each category has a *prior* utility

$$u_0(\psi) = n_0(\psi)u(\alpha^\nu(P_0(\psi)/n_0(\psi)), P_0(\psi)/n_0(\psi))$$

corresponding to the utility that would have been received had the player played a "smoothed best-response" to repeated observation of the prior sample. We can also define a *utility with prior* for the entire course of play $\tilde{u}_t(\psi) \equiv u_t(\psi) + u_0(\psi)$ that includes the prior utility. If we compute the cost using not only the prior utility, but also replacing the utility that could have been achieved V with the smooth version V^ν, we get the *smooth cost with prior*

$$\tilde{c}_t(\psi) = \tilde{n}_t(\psi)V^\nu\left(\frac{\tilde{P}_t(\psi)}{\tilde{n}_t(\psi)}\right) - \tilde{u}_t(\psi).$$

There is of course also the *total smooth cost with prior* $\tilde{c}_t = \sum_{\psi\in\hat{\Psi}_t}\tilde{c}_t(\psi)$

The basic idea in what follows is that V^ν is very close to V, and in a large sample, the prior does not matter very much, so it is sufficient to study the smooth cost with prior. Indeed, define the *incremental (smooth) cost (with prior)*

$$\tilde{g}_t^\nu(a_t, y_t) = \tilde{g}^\nu(a_t, y_t|h_{t-1}) = \tilde{c}_t^\nu - \tilde{c}_{t-1}^\nu.$$

---

each opponent, and assuming their 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.

Our two propositions follow from a lemma that gives a basic bound on the incremental cost.

Lemma 1: (a) $|\tilde{g}_t^{\nu}(a, y)| \leq 4\overline{U}$

(b) $\tilde{g}_t^{\nu}(\sigma^{\nu}(h_{t-1}), \gamma_t) \leq \|\nu\| + \dfrac{B^{\nu}}{1 + \inf_{\psi \in \Psi} n_0(\psi)}$.

Remark 1: Note that the bound (b) is by far the most significant; the bound (a) simply says that $\tilde{g}_t^{\nu}$ is a bounded random variable, which makes it easy to use the strong law of large number. The bound (b) is what enables us to reach conclusions about universal consistency.

Remark 2: Since $\tilde{g}_t^{\nu}(\sigma(h_{t-1}), \gamma) \leq \|\nu\| + B^{\nu} / (1 + \inf_{\psi \in \Psi} n_0(\psi))$, then the same bound must also be satisfied by the maxmin strategy for the zero-sum game $-\tilde{g}_t^{\nu}(a, y)$. If $\hat{\sigma}^{\nu}$ is such a strategy, we refer to it as a *myopic* play. It follows that Lemma 1 (and Proposition 1) remain true if the strategy of smooth categorical fictitious play is replaced with myopic play.

Proof of Lemma 1:

We begin by writing out $\tilde{g}_t^{\nu}$. Let $\psi_t^a = \hat{\psi}(h_{t-1}, a)$ be the category at time *t* chosen by the classification function if *a* is played then. Abbreviate $\tilde{n}_{t-1}^a = \tilde{n}_{t-1}(\psi_t^a)$ and $\tilde{P}_{t-1}^a = \tilde{P}_{t-1}(\psi_t^a)$. With this notation, (a) follows from the following calculation

$$
\begin{aligned}
&|\tilde{g}_t^{\nu}(a,y)| \\
&=\left|(\tilde{n}_{t-1}^{a}+1)V^{\nu}\left(\frac{\tilde{P}_{t-1}(\psi_t^a)+e(y)}{\tilde{n}_{t-1}^{a}+1}\right)-u(a,y)-\tilde{n}_{t-1}^{a}V^{\nu}\left(\frac{\tilde{P}_{t-1}(\psi_t^a)}{\tilde{n}_{t-1}^{a}}\right)\right| \\
&\le|(\tilde{n}_{t-1}^{a}+1)\left(V^{\nu}\left(\frac{\tilde{P}_{t-1}(\psi_t^a)+e(y)}{\tilde{n}_{t-1}^{a}+1}\right)-V^{\nu}\left(\frac{\tilde{P}_{t-1}(\psi_t^a)}{\tilde{n}_{t-1}^{a}}\right)\right)+|V^{\nu}\left(\frac{\tilde{P}_{t-1}(\psi_t^a)}{\tilde{n}_{t-1}^{a}}\right)-u(a,y)| \\
&\le|(\tilde{n}_{t-1}^{a}+1)\sup_{\gamma}\left|DV^{\nu}(\gamma)\right|\left|\frac{\tilde{n}_{t-1}^{a}e(y)-P_{t-1}(\psi_t^a)}{(\tilde{n}_{t-1}^{a}+1)n_{t-1}^{a}}\right|+|V^{\nu}\left(\frac{\tilde{P}_{t-1}(\psi_t^a)}{\tilde{n}_{t-1}^{a}}\right)-u(a,y)| \\
&\le\sup_{\gamma}\left|DV^{\nu}(\gamma)\right|+|V^{\nu}\left(\frac{\tilde{P}_{t-1}(\psi_t^a)}{\tilde{n}_{t-1}^{a}}\right)-u(a,y)|\le 4\overline{U}
\end{aligned}
$$

to prove (b)

$$
\tilde{g}_t^{\nu}(a,\gamma_t)\le(\tilde{n}_{t-1}^{a}+1)\sum\nolimits_{y}\left[V^{\nu}\left(\frac{\tilde{P}_{t-1}^{a}+e(y)}{\tilde{n}_{t-1}^{a}+1}\right)-V^{\nu}(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}})\right]\gamma_t(y)-u(a,\gamma_t)+V^{\nu}(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}).
$$

Moreover, since V^{ν} is smooth we may apply Taylor's theorem as discussed above to conclude

$$
\begin{aligned}
\tilde{g}_t^{\nu}(a,\gamma_t)&\le\sum\nolimits_{y}u(\alpha_{\nu}(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}},\cdot)\left[e(y)-\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}\right]\gamma_t(y)-u(a,\gamma_t)+V^{\nu}(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}})+\frac{B^{\nu}}{\tilde{n}_{t-1}^{a}+1} \\
&\le u(\alpha_{\nu}(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}),\gamma_t)-u(\alpha_{\nu}(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}),\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}})-u(a,\gamma_t)+V^{\nu}(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}})+\frac{B^{\nu}}{1+\inf_{\psi\in\Psi}n_0(\psi)}
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
V^{\nu}\left(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}\right)&=u\left(\alpha^{\nu}\left(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}\right),\left(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}\right)\right)+v\left(\alpha^{\nu}\left(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}\right)\right) \\
&\le u\left(\alpha^{\nu}\left(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}\right),\left(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}\right)\right)+\|v\|
\end{aligned}
$$

yielding

$$
\begin{aligned}
\tilde{g}_t^{\nu}(a,\gamma_t)&\le u(\alpha_{\nu}(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}),\gamma_t)-u(a,\gamma_t)+\|v\|+\frac{B^{\nu}}{1+\inf_{\psi\in\Psi}n_0(\psi)} \\
&=-u(a-\alpha_{\nu}(\frac{\tilde{P}_{t-1}^{a}}{\tilde{n}_{t-1}^{a}}),\gamma_t)+\|v\|+\frac{B^{\nu}}{1+\inf_{\psi\in\Psi}n_0(\psi)}
\end{aligned}.
$$

We may rewrite this in terms of the matrix $A^{\nu}(\tilde{P}_{t-1}, \tilde{n}_{t-1})$ as

$$\tilde{g}_t^{\nu}(\alpha, \gamma_t) \le -u(\alpha\left(I - A^{\nu}(\tilde{P}_{t-1}, \tilde{n}_{t-1})\right), \gamma_t) + \|v\| + \frac{B^{\nu}}{1 + \inf_{\psi \in \Psi} n_0(\psi)}$$

Since $\hat{\alpha}^{\nu}\left(\tilde{P}_{t-1}, \tilde{n}_{t-1}\right)$ solves $\alpha\left(I - A^{\nu}(\tilde{P}_{t-1}, \tilde{n}_{t-1}\right) = 0$, the conclusion follows.

❑

proof of Proposition 1: Observe first that

$$\begin{aligned} \limsup_{T\to\infty} \tilde{c}_T^{\nu} / T &= \limsup_{T\to\infty} \left(\sum_{t=1}^{t} \left(\tilde{c}_t^{\nu} - \tilde{c}_{t-1}^{\nu}\right)\right) / T \\ &= \limsup_{T\to\infty} \left(\sum_{t=1}^{T} \tilde{g}_t^{\nu}\right) / T \end{aligned}$$

By Lemma 1(a) $\tilde{g}_{\tau}^{\nu}$ are uniformly bounded random variables. In addition by Lemma 1(b)

$$E[\tilde{g}_t^{\nu} | h_{t-1}] = \tilde{g}_t^{\nu}(\sigma^{\nu}(h_{t-1}), \rho(h_{t-1})) \le \|v\| + B^{\nu} / (1 + \inf_{\psi \in \Psi} n_0(\psi)) .$$

It follows directly from the strong law of large numbers for orthogonal sequences[14] othat

$$\limsup_{T\to\infty} \tilde{c}_T^{\nu} / T \le \|v\| + B^{\nu} / (1 + \inf_{\psi \in \Psi} n_0(\psi)) .$$

Next leting $\tilde{c}_T$ denote the cost where the actual utility V is used in place of the smoothed utility V^{ν} (but keeping the prior observations) we observe that since $\left\|V^{\nu} - V\right\| \le \|v\|$

$$\limsup_{T\to\infty} \left|\tilde{c}_T^{\nu} - \tilde{c}_T\right| / T \le \|v\|,$$

from which

$$\limsup_{T\to\infty} \tilde{c}_T / T \le 2\|v\| + B^{\nu} / (1 + \inf_{\psi \in \Psi} n_0(\psi)) .$$

Finally, observe that

$$\left|\frac{1}{T}\sum_{t=1}^{T}u(a_t,y_t)-\frac{1}{T+\sum_{\psi\in\hat{\Psi}_T}n_0(\psi)}\left(\sum_{t=1}^{T}u(a_t,y_t)+\sum_{\psi\in\hat{\Psi}_T}P_0(\psi)\right)\right|$$
$$\leq\left|\frac{\sum_{\psi\in\hat{\Psi}_T}n_0(\psi)}{T+\sum_{\psi\in\hat{\Psi}_T}n_0(\psi)}\frac{1}{T}\sum_{t=1}^{T}u(a_t,y_t)\right|+\left|\frac{1}{T+\sum_{\psi\in\hat{\Psi}_T}n_0(\psi)}\left(\sum_{\psi\in\hat{\Psi}_T}P_0(\psi)\right)\right|$$
$$\leq\left|\frac{m_T\sup_{\psi\in\Psi}n_0(\psi)}{T+m_T\sup_{\psi\in\Psi}n_0(\psi)}\frac{1}{T}\sum_{t=1}^{T}u(a_t,y_t)\right|+\left|\frac{m_T\sup_{\psi\in\Psi}n_0(\psi)}{T+m_T\sup_{\psi\in\Psi}n_0(\psi)}\right|\to 0$$

since by Assumption 1 $m_T/T\to 0$ while by Assumption 2 $\sup_{\psi\in\Psi}n_0(\psi)<\infty$. This implies that $\left|\tilde{u}_T/T-u_T/T\right|\to 0$ and that

$$\frac{1}{T}\left|\sum_{\psi\in\hat{\Psi}_T}\tilde{n}_T(\psi)V\left(\frac{\tilde{P}_T(\psi)}{\tilde{n}_T(\psi)}\right)-n_T(\psi)V\left(\frac{P_T(\psi)}{\tilde{n}_T(\psi)}\right)\right|\to 0$$

since the difference in the objective functions is going to zero. So we conclude $\limsup_{T\to\infty}\left|\tilde{c}_T-c_T\right|/T=0$ giving the desired result.

☑

Turning to Proposition 2, we observe that it follows from the following Lemma dealing with the non-history contingent case.

Lemma 2: For any strategy σ of playing a constant action α and any ε there exists a smooth fictitious play σ^{ν} that is ε-as good as σ.

Before proving the Lemma, we discuss how it can be used to prove Proposition 2. The rule that is being outperformed in Lemma 2, that of guessing the opponent will always play a single action, is not a terribly interesting rule. However, let s be an arbitrary deterministic learning rule,[15] and let $\varepsilon>0$ be given. Take a set of $\Psi=A$ to be

---

[14] Recall that a sequence of r.v.'s y_t is orthogonal if $Ey_ty_k=0, t\neq k$.

[15] The extension to random rules is straightforward.

set of actions. Define the classification rule $\hat{\psi}(h_t) = s(h_t)$, that is, classify histories according to the way in which s is going to play. For each a choose a smooth fictitious play $\sigma(\varepsilon, a)$ so that Lemma 2 is satisfied with respect to ε, and define a rule

$$\hat{\sigma}(s,\varepsilon)(h_t) = \sigma(\varepsilon, \hat{\psi}(h_t))(h_t(\hat{\psi}))$$

by applying the appropriate smooth fictitious play to the sub-history of the chosen category. Lemma 2 immediately implies that $\hat{\sigma}$ is the rule that gives Proposition 2. Note that the fact that Lemma 2 holds for arbitrary non-stationary discount sequences is important, since even if we use stationary discounting in Proposition 2, when we look at sub-histories to apply Lemma 2, the discount factor will no longer be stationary, since the sequence of sub-histories involves skipping periods during which a particular category was not used.

Proof of Lemma 2: We may assume without loss of generality that α is a best-response to an open subset of opponent correlated strategies, since if not, α is weakly dominated by a strategy that does satisfy this property, and we may simply replace α with this dominating strategy in the calculations below. Consequently, there exists a prior such that α is a best response to P_0 / n_0. From Lemma 1(b)

$$\tilde{g}_t^v = (t+n_0)V^v\left(\frac{\tilde{P}_t}{\tilde{n}_t}\right) - u(\sigma^v(h_{t-1}), \rho(h_{t-1})) - (t+n_0-1)V^v\left(\frac{\tilde{P}_{t-1}}{\tilde{n}_{t-1}}\right) \le \|v\| + \frac{B^v}{t+n_0}.$$

We may now compute

$$\begin{aligned}
&u(\alpha, \rho(h_{t-1}))) - u(u(\sigma^v(h_{t-1}), \rho(h_{t-1})))\\
&= \tilde{g}_t^v + u(\alpha, \rho(h_{t-1})) - \left((t+n_0)V^v\left(\frac{\tilde{P}_t}{\tilde{n}_t}\right) - (t+n_0-1)V^v\left(\frac{\tilde{P}_{t-1}}{\tilde{n}_{t-1}}\right)\right)\\
&\le \|v\| + \frac{B^v}{t+n_0} + u(\alpha, \rho_t) - \left((t+n_0)V^v\left(\frac{\tilde{P}_t}{\tilde{n}_t}\right) - (t+n_0-1)V^v\left(\frac{\tilde{P}_{t-1}}{\tilde{n}_{t-1}}\right)\right)
\end{aligned}$$

Adding up over periods, we have

$$\begin{aligned}
&\textstyle\sum_{t=1}^{T} u(\alpha,\rho(h_{t-1})) - u(\sigma^{v}(h_{t-1}),\rho(h_{t-1})) \\
&\le T\|v\| + \frac{TB^{v}}{t+n_0} + \textstyle\sum_{t=1}^{T}\left(u(\alpha,\rho(h_{t-1})) - \left((t+n_0)V^{v}\left(\frac{\tilde{P}_t}{\tilde{n}_t}\right) - (t+n_0-1)V^{v}\left(\frac{\tilde{P}_{t-1}}{\tilde{n}_{t-1}}\right)\right)\right) \\
&= T\|v\| + \frac{TB^{v}}{t+n_0} + \textstyle\sum_{t=1}^{T} u(\alpha,\rho(h_{t-1})) - \left((T+n_0)V^{v}\left(\frac{\tilde{P}_T}{\tilde{n}_T}\right) - n_0 V^{v}\left(\frac{\tilde{P}_0}{\tilde{n}_0}\right)\right)
\end{aligned}$$

Noting that $(T+n_0)\left(\frac{\tilde{P}_t}{\tilde{n}_t}\right) = T\left(\frac{P_t}{\tilde{n}_t}\right) + n_0\left(\frac{P_0}{n_0}\right)$ and $\tilde{P}_0 / \tilde{n}_0 = P_0 / n_0$

$$\textstyle\sum_{t=1}^{T} u(\alpha,\rho(h_{t-1})) - u(\sigma^{v}(h_{t-1}),\rho(h_{t-1})) \le T\|v\| + \frac{TB^{v}}{t+n_0} + \sum_{t=1}^{T} u(\alpha,\rho(h_{t-1})) - TV^{v}\left(\left(\frac{P_t}{n_t}\right)\right)$$

We also have

$$u(\alpha,\left(\frac{P_T}{n_T}\right)) \le \max_a u(a,\left(\frac{P_T}{n_T}\right)),$$

and since α is by assumption a best response to P_0 / n_0, it must also be that

$$u(\alpha,\left(\frac{P_T}{n_T}\right)) \le \max_a u(a,\left(\frac{P_T}{n_T}\right)) \le V^{v}(\left(\frac{P_T}{n_T}\right)) + \|v\|.$$

Applying this inequality give

$$\textstyle\sum_{t=1}^{T} u(\alpha,\rho(h_{t-1})) - u(\sigma^{v}(h_{t-1}),\rho(h_{t-1})) \le 2T\|v\| + \frac{TB^{v}}{t+n_0}.$$

Finally, choosing v, n_0 appropriately, we get

$$\textstyle\sum_{t=1}^{T} u(\alpha,\rho(h_{t-1})) - u(\sigma^{v}(h_{t-1}),\rho(h_{t-1})) \le T\varepsilon$$

for all *T*.

Making use of the fact that average present value may be written as a convex combination of time averages over shorter time intervals we immediately can convert this into a result for discounting;

$$\sum_{t=1}^{\infty}\beta_t\left(u(\alpha,\rho(h_{t-1}))-u(\sigma^{\nu}(h_{t-1}),\rho(h_{t-1}))\right)$$
$$=\sum_{T=1}^{\infty}(\beta_{t-1}-\beta_t)\sum_{t=1}^{T}\left(u(\alpha,\rho(h_{t-1}))-u(\sigma^{\nu}(h_{t-1}),\rho(h_{t-1}))\right)$$
$$\leq \varepsilon\sum_{T=1}^{\infty}(\beta_{t-1}-\beta_t)\sum_{t=1}^{T}t=\varepsilon$$

☑

7. Conclusion

Given any particular learning rule we have shown how to construct a rule that does almost as well regardless of the discount factor and does as well asymptotically as if the conditional probabilities of a sufficiently small number of events was known in advance. This is true regardless of the process generating the outcomes. These rules have the interpretation of being a convex combination of smooth fictitious plays corresponding to the different available categories. If strategies that successfully condition on the most recently chosen actions are used by all players, then in the long the time average of the frequency of strategy profiles must remain near the set of correlated equilibria. Moreover, it is not essential that players actually condition on their own most recently chosen action, it is enough that they have the option of doing so, and use a categorization procedure with the property that if has been historically advantageous to condition on own past action, then this type of categorization will be introduced. This suggests that in the long run the play of sophisticated players should resemble a correlated equilibrium.

The convergence-to-correlated- equilibrium s result cannot in general be strengthened to eliminate cycles or give convergence to a Nash equilibrium, no matter how sophisticated the learning procedure employed by players are, since they might "accidentally" coordinate their learning procedures. In an example, noise, in the form of lack of coordination between players' learning procedures, does lead to a Nash

equilibrium. Whether this type of noise, or noise from a large population of players playing independently, might generally lead to Nash equilibrium remains an open question.

References

Aoyagi, M.[1994]: “Evolution of Beliefs and the Nash Equilibrium of Normal Form Games,” mimeo, U. Pittsburg.

Auer, P., N. Cesa-Bianchi, Y. Freund and R. Schapire [1995]: “Gambling in a rigged casino: The adversarial multi-armed bandit problem,” *36th Annual IEEE Symposium on Foundations of Computer Science*.

Banos, A. [1968]: “On Pseudo-Games,” *Annals of Mathematical Statistics*, 39, 1932-1945.

Benaim, M. and M. Hirsch [1994]: “Learning Processes, Mixed Equilibria and Dynamical Systems Arising from Repeated Games,” mimeo.

Blackwell, D. [1956]: “Controlled Random Walks,” *Proceedings International Congress of Mathematicians*, III 336-338, Amsterdam, North Holland.

Chung, T.H. [1994]: “Approximate Methods for Sequential Decision Making Using Expert Advice,” *Proceedings of the 7th Annual ACM Conference on Computational Learning Theory*, 183-189, 1994.

Dawid, A. [1982]: “The Well-Calibrated Bayesian,” *Journal of the American Statistical Association*, 77, 605-613.

Dawid, A. [1985]: “The Impossibility of Inductive Inference,” *Journal of the American Statistical Association*, 80, 340-341.

Desantis, A., G. Markowski and M. Wegman [1992]: "Learning Probabilistic Prediction Functions," *Proceedings of the 1988 Workshop of Computational Learning*, 312-328.

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 [1996]: "Regret in the On-line Decision Problem," Wharton School.

Foster, D. and R. Vohra [1993]: "Calibrated Learning and Correlated Equilibrium," mimeo Wharton.

Foster, D. and R. Vohra [1994]: "Asymptotic Calibration," mimeo, Wharton.

Foster D. and R. Vohra [1995]: results presented at the Conference on Learning in games held at Carlos III University, Madrid

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. Kreps [1993]: "Learning Mixed Equilibria," *Games and Economic Behavior* 5, 320-367.

Gantmacher, F. [1959]: *Matrix Theory vol. 2*, New York: Chelsea.

Fudenberg, D. and D. Levine [1993], "Self Confirming Equilibrium", *Econometrica*, 61: 523-545.

Fudenberg, D. and D. Levine [1995] "Universal Consistency and Cautious Fictitious Play," *Journal of Economic Dynamics and Control*,

Fudenberg, D. and D. Levine [1996], "An Easier Way to Calibrate," UCLA.

Fudenberg, D. and D. Levine [1997] "*The Theory of Learning in Games,"* MIT Press, Cambridge, MA.

Hannan, J. [1957]: "Approximation to Bayes Risk in Repeated Plays," in M. Dresher, A.W. Tucker and P. Wolfe, eds. *Contributions to the Theory of Games*, volume 3, 97-139, Princeton University Press.Jordan, J. [1993]: "Three problems in Learning Mixed-Strategy Nash Equilibria," *Games and Economic Behavior* 5, 368-386.

Hart, S. and A. Mas-Colell [1996], "A Simple Adaptive Procedure Leading to Correlated Equilibrium," Hebrew University of Jerusalem.

Kalai, E., E. Lehrer and R. Smorodinsky [1996]: "Calibrated Forecasting and Merging," Northwestern MEDS #1144

Kanivokski, Y. and P. Young [1994]: "Learning Dynamics in Games with Stochastic Perturbations," mimeo.

Kivinen J. and M. Warmuth [1993]: "Using Experts in Predicting Continuous Outcomes," *Computational Learning Theory: EURO COLT*, 109-120, Springer-Verlag.

Littlestone, N. and M. Warmuth [1994]: "The Weighted Majority Algorithm," *Information and Computation*, 108, 212-261 (also appeared in the 30th FOCS conference, 1989).

Megiddo, N. [1980]: "On Repeated Games with Incomplete Information Played with Non-Bayesian Players," *International Journal of Game Theory*, 9, 157-167.

Oakes, David [1985] "Self-Calibrating Priors Do Not Exist," *Journal of the American Statistical Association*, 80.

Sanchirico, C. [1995] “A Probabilistic Model of Learning in Games,” mimeo, Columbia University.

Sonsino, D, [1995] “ Learning to Learn, Pattern Recognition, and Nash Equilibrium,” mimeo, Stanford GSB.

Shapley, L. [1964]: “Some Topics in Two-Person Games,” in *Advances in Game Theory*, (M. Drescher, L.S. Shapley, and A.W. Tucker, eds.) Princeton, NJ: Princeton University Press.

Vovck, V. [1990]: “Aggregating Strategies,” *Proceedings of the 3rd Annual Conference on Computational Learning Theory*, 371-383.