CONVERGENCE OF BEHAVIOUR RULES IN ITERATED MATRIX GAMES

by

JONATHAN PATRICK

B. Arts & Sc., McMaster University, 1997

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Mathematics
Institute of Applied Mathematics

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1999

© Jonathan Patrick, 1999

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Mathematics
The University of British Columbia
Vancouver, Canada

Date 13/10/99

Abstract

This master's thesis reports on a foray into Game Theory, focusing solely on the two-person (not necessarily zero-sum) game. Primarily, I am interested in the convergence properties of different behaviour rules and how one might proceed to introduce some form of learning into the strategies of the players involved in the game. Therefore, I begin with the introduction of some key equilibrium sets – namely the set of Nash Equilibria (NE), the set of correlated equilibria (CE) and the marginal best-response set (MBR). I briefly discuss the relationship between these three sets before moving on to describe some desirable properties of behaviour rules. From there, I introduce six behaviour rules (four from the literature, two original) that attempt to incorporate some form of learning into the game. The four from the literature are Fictitious Play, Exponential Fictitious play, Regrets 1 and Regrets 2. I have named the two original behaviour rules Past Response and Modified Regrets. I then move on to describe the convergence properties of each.

This thesis was originally motivated by a talk given by Andreu Mas-Collel on the properties of the two Regrets-based behaviour rules. Thus, a fair amount of time is spent reworking the convergence proofs of both Regrets1 and Regrets2 as they were developed by Mas-Collel and Sergiu Hart. I then suggest an alternative proof of the Regrets1 convergence properties. I close off the paper with some numerical results from three games – a zero-sum game, a game developed by Lloyd Shapley (called the Shapley game) and a game called Battle of the Buddies. They are designed to give some numerical confirmation of the convergence theorems stated earlier in the paper as well as some indication as to where further study might be useful.

Table of Contents

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Loewen, for his unceasing optimism and constant encouragement – not to mention his patience and willingness to give of his time. Thanks is also due to Dr. Puterman from the Management Science department for his willingness to help out in any way and Dr. Anstee for being my second reader. Equally worthy of thanks are David Urminsky and Marc Mailhot (my housemates) who had to put up with my sometimes less than congenial self on those days when the thesis writing got the better of me. (Dave also saved me from taking a bat to a few computers by providing much needed computer knowledge.)

Chapter 1
Introduction: Playing the Game

A mathematical game is an attempt to model the interaction between individuals or organizations. A game may have any number of "players", representing either individuals or groups. In this thesis, we will restrict ourselves to the most obviously practical and common case of a two player game. This is not a limiting restriction as most non-cooperative games can be reduced to two players simply by having each player view all the others as one unit. Thus each player responds to the outcome of all the other players' actions rather than looking at each action individually. We will refer to this second player, N, as "nature" or the "environment". The first player will be denoted by the letter M. We will, more often than not, take the position of player M with player N's point of view usually following analogously.

1.1 The one-shot game

In order to illustrate the basic setup, consider the simple game called "matching pennies". In this game M tries to match the choice of heads or tails made by N, while N tries to prevent M from doing so. Each receives a numerical pay-off of $+1$ for a success and -1, for a failure. Thus if N chooses heads and M chooses tails then M receives a pay-off of -1 (penalty) and N receives a pay-off of $+1$ (reward). In the event that M plays strategy i, (in this game either heads, h, or tails, t), and N plays strategy j (again either heads or

tails), the payoff (sometimes called utility) for player M is represented by $u_M(i,j)$ and for player N by $u_N(i,j)$. Thus $u_M(h,t) = -1$ and $u_N(h,t) = 1$. Of course, if M knows nature's choice, *in advance of making his own decision*, then M will win every time ... leading to a rather uninteresting game. Even partial knowledge of N's action will confer an advantage on M. The effects of what information is available to each player and how best to use that information is a key issue in game theory. Here we address the simplest case where both players act independently and in ignorance of the other's action. I will denote the set of possible strategies (or actions) available to player M by S_M, and the set of possible actions available to nature by S_N.

In the matching pennies game, both M and N have only two possible actions, h and t. Their payoffs therefore can be represented in a very simple matrix form where the columns represent the strategies of nature and the rows represent the strategies of M, i.e., the top right hand entry of the matrix refers to the event that M chooses heads and N chooses tails. The entries in the matrix refer to the payoff received in each case. Thus, the payoff matrices of players M and N respectively are

$$u_M = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad u_N = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

It is easy to see that in any two person game where player M has m strategies, $|S_M| = m$, and player N has n strategies, $|S_N| = n$, we can represent their respective payoffs for each possible pair of strategies in two $m \times n$ matrices. Such a game will be called an $m \times n$ *matrix game* and is defined by its two payoff matrices.

The Matching Pennies game is an example of what is called a zero-sum game. Such a game is characterized by the fact that $u_M = -u_N$. Much of the theory is simplified in this sub-class of games.

1.2 Mixed Strategies: Best Response and Minimax

There are obviously many ways for a player in a game to choose an action. For instance, in the "matching pennies" game, M could choose to play heads, or to play tails, or, using a random number generator, to pick heads with probability p and tails with probability $1-p$, for any preselected p obeying $0 \leq p \leq 1$. In more mathematical terms, this amounts to M placing a probability measure, μ, on the set of possible strategies, S_M. N, naturally, must do likewise, placing a probability measure, ν, on the set of strategies, S_N. (Thus the case where M plays heads with certainty is simply the measure that places all its weight on heads and none on tails. Probability measures that place all their mass on a single action are called *pure strategies*; general measures are called *mixed strategies.*) I will denote the set of probability measures on a set S by $P[S]$. Once M and N have chosen their measures $\mu \in P[S_M]$ and $\nu \in P[S_N]$ respectively, player M's expected payoff is

$$u_M(\mu, \nu) = \sum u_M(i,j)\mu(i)\nu(j), \tag{1.1}$$

where the summation is taken over all $i \in S_M$ and all $j \in S_N$.

A natural question concerns how player M decides on a measure considering that the measure ν of player N is unknown. One method is to make an (educated) guess as to how N is likely to act and then to respond in a manner that maximizes one's own expected payoff based on that guess. This is formalized in the following definition.

Definition: We say that $\hat{\mu}$ is an *ϵ-best response to ν* for M if

$$u_M(\hat{\mu}, \nu) \geq \sup_{\mu \in P[S_M]} u_M(\mu, \nu) - \epsilon.$$

If $\epsilon = 0$ then we refer to $\hat{\mu}$ simply as a best response to ν.

Among the best-response measures, there is, necessarily, a measure that places all its weight on one strategy. This follows from the fact that $u_M(\mu, \nu)$ is a linear function in

μ. That is, it can be written as,

$$u_M(\mu,\nu) = \sum_{i \in S_M} \left(\sum_{j \in S_N} u_M(i,j)\nu(j) \right) \mu(i)$$

Thus, one best-response measure, $\hat{\mu}$, places all its weight on the strategy $i \in S_M$ which has the greatest coefficient, $\sum_{j \in S_N} u_M(i,j)\nu(j)$. It should be fairly obvious that the set of best-response measures to $\nu, B[\nu] = argmax_{\mu \in P[S_M]} u_M(\mu,\nu)$, is never empty (since $|S_M|$ is finite) and may consist of more than one measure. For the purposes of this thesis, in cases where there is more than one best response we will use the arbitrary rule which chooses the best-response measure which places all its weight on the action with the smallest subscript. That is, we pick $\mu = \delta_k$ where

$$k = \min\{m : \delta_m \in B[\nu]\}.$$

The key question concerning the origins of this educated guess (since the actual measure, ν, of player N is unknown) will be dealt with later.

Of course, one need not proceed by always playing a best-response to some estimate. A more cautious player might want to maximize the worst-case outcome. Specifically, this means that player M would choose μ in order to maximize:

$$\min_{\nu \in P[S_N]} u_M(\mu,\nu).$$

The von Neumann minimax theorem asserts that, for a zero-sum matrix game, each player can insure that his pay-off does not dip below a certain value, v. In mathematical terms, this means that there exists a $\hat{\mu} \in P[S_M]$ such that

$$u_M(\hat{\mu},\nu) \geq v \qquad \forall \nu \in P[S_N].$$

Similarly, there exists a measure $\hat{\nu} \in P[S_N]$ such that

$$u_N(\mu,\hat{\nu}) \geq -v \qquad \forall \mu \in P[S_N].$$

In a zero-sum game, this implies that there exists $(\hat{\mu}, \hat{\nu}) \in P[S_M] \times P[S_N]$ and a scalar v such that

$$u_M(\mu, \hat{\nu}) \leq v \leq u_M(\hat{\mu}, \nu) \qquad \forall (\mu, \nu) \in P[S_M] \times P[S_N]$$

([1]). Other perhaps more interesting methods of choosing a measure will be discussed later on once more complexity has been added to the game.

The above discussion might lead one to question whether there might exist a pair of measures (one from each player) such that neither player would benefit from unilaterally changing measure (that is, by playing differently while assuming that the other player's measure remains the same). The following definition formalizes this concept.

Definition: A pair of measures $\hat{\mu} \in P[S_M]$ and $\hat{\nu} \in P[S_N]$ is a *Nash equilibrium combination* iff no single player can unilaterally increase his (expected) payoff by choosing a different measure. Thus $(\hat{\mu}, \hat{\nu})$ is a Nash equilibrium iff both:

$$u_M(\hat{\mu}, \hat{\nu}) \geq \max_{\mu \in P[S_M]} u_M(\mu, \hat{\nu}) \quad \text{and} \quad u_N(\hat{\mu}, \hat{\nu}) \geq \max_{\nu \in P[S_N]} u_N(\hat{\mu}, \nu).$$

I will denote the set of Nash equilibria by NE. This is a (possibly empty) subset of the Cartesian Product space, $P[S_M] \times P[S_N]$. Note that if $(\hat{\mu}, \hat{\nu}) \in$ NE, then $\hat{\mu}$ must be a best response to $\hat{\nu}$ in the sense defined above. Similarly, $\hat{\nu}$ must also be a best response to $\hat{\mu}$.

In the matching pennies game, if $\hat{\nu} = [0.5, 0.5]$, then

$$u_M(\mu, \hat{\nu}) = 0 = u_N(\mu, \hat{\nu}) \qquad \text{for all } \mu \in P[S_M]$$

This includes the two pure strategies, $\mu_h = [1, 0]$ and $\mu_t = [0, 1]$. To find a NE pair, $(\hat{\mu}, \hat{\nu})$, involving the above $\hat{\nu}$, we must choose $\hat{\mu}$ to ensure that

$$0 = u_N(\hat{\mu}, \hat{\nu}) \geq \max_{\nu \in P[S_N]} u_N(\hat{\mu}, \nu)$$

This is equivalent to finding $\hat{\mu}$ such that

$$\begin{aligned} 0 &\geq \max_{\nu \in P[S_N]} [(\hat{\mu}(2) - \hat{\mu}(1))\nu(1) + (\hat{\mu}(1) - \hat{\mu}(2))\nu(2)] \\ &= \max [\hat{\mu}(2) - \hat{\mu}(1), \hat{\mu}(1) - \hat{\mu}(2)] . \end{aligned}$$

Thus for $(\hat{\mu}, \hat{\nu}) \in$ NE, we are forced to choose $\hat{\mu}(1) = \hat{\mu}(2)$, i.e., $\hat{\mu} = [0.5, 0.5]$. Such a choice ensures that $(\hat{\mu}, \hat{\nu}) \in$ NE.

1.3 Co-ordinated Strategies (Measures on the Product Space)

So far, we have only considered placing separate measures on the two players' strategy sets. A natural extension is to place a measure, α, on the product space $S_M \times S_N$ instead. Thus, under this measure, each *pair* of strategies $(i, j) \in S_M \times S_N$ is given a certain probability or weight. We will call such a measure a *correlated measure.* The expected value of the payoff function for player M (and similarly for player N), imposed by the measure α, is then defined by:

$$E^{\alpha}(u_M) = \sum_{(i,j) \in S_M \times S_N} u_M(i, j)\alpha(i, j)$$

where $\alpha(i, j)$ is the weight placed on the pair of strategies (i, j) by the measure α.
I will use the short-hand notation, $u_M(\alpha)$ for $E^{\alpha}(u_M)$ whenever possible.

All this may also be written in matrix notation. The correlated measure, α, can be written as a $|S_M| \times |S_N|$ matrix A, with nonnegative entries, $\alpha(i, j)$, that sum up to one. In the case where α is a product measure, writing μ and ν as row vectors gives the matrix representation $A = \mu^T \nu$. A correlated measure thus represents a product measure if and only if its matrix representation has rank 1. We have already seen that player M's (and player N's) utility function can be represented in matrix form as u_M (u_N). Thus $u_M(\alpha) = tr(A^T u_M)$.

One might wonder why one would want to study correlated measures seeing as we are here only interested in non-cooperative games. After all a correlated measure seems to represent a form of cooperation between players. It will become clearer later as to why this is not necessarily so but an initial hint comes from the following fact. It should be evident that

$$\max_{\alpha} u_M(\alpha) \geq \max_{\mu,\nu} u_M(\mu,\nu)$$

since the maximum on the right hand side is taken over a subset of the measures that are available on the left. Thus it seems possible that both players might gain more by considering product measures, *even if they are not interested in cooperating.*

Once one has a correlated measure, it is a simple procedure to deduce the marginal measures for each player. Under a correlated measure α, the marginal measure for player M is given by:

$$\alpha_M(i) = \sum_{j \in S_N} \alpha(i,j), \quad \forall i \in S_M$$

The marginal for player N, α_N, follows analogously.

However, a measure $\alpha \in P[S_M \times S_N]$ may not derive from the product of two measures, $\mu \in P[S_M]$ and $\nu \in P[S_N]$. The following example makes this clear. Consider a correlated measure that places the following probabilities on the nine pairs of strategies in a 3×3 game:

$$A = \begin{bmatrix} .20 & .30 & .00 \\ .00 & .20 & .05 \\ .15 & .00 & .10 \end{bmatrix}$$

The marginal measure for player M is $(.5, .25, .25)$ and for player N is $(.35, .5, .15)$, but there are no marginals that, when multiplied together, will produce the measure, α, given above. An easy way to ascertain this is to note that the rank of the above matrix is 3.

The introduction of correlated measures leads to a redefinition of the concepts of *best-response* and *equilibrium*:

Definition: A (correlated) measure α has the *marginal best-response property* if,

$$\max_{\mu \in P[S_M]} u_M(\mu, \alpha_N) \leq u_M(\alpha) \quad \text{and} \quad \max_{\nu \in P[S_N]} u_N(\alpha_M, \nu) \leq u_N(\alpha).$$

The marginal best-response property is called *exact* if the above two equations are actually equalities ([4]). I will denote the set of measures with the marginal best-response property (not necessarily exact) by MBR.

Definition: A correlated measure α is called a *correlated equilibrium* if the following two equations are satisfied:

$$u_M(\alpha) \geq \sum_{i \in S_M, j \in S_N} u_M(\psi(i), j)\alpha(i,j) \quad \forall \psi : S_M \to S_M$$
$$u_N(\alpha) \geq \sum_{i \in S_M, j \in S_N} u_N(i, \phi(j))\alpha(i,j) \quad \forall \phi : S_N \to S_N.$$

I will denote by CE, the set of correlated equilibria.

In words, a correlated equilibrium is a probability distribution or measure that assigns probability to all possible combinations of players' strategies in such a way that no one player can, by himself, increase his expected gain by redistributing the measure. One way of visualizing this is in terms of the payoff matrix. Suppose we hold the probability of each entry in the matrix A fixed. Then a correlated measure is in CE if and only if player M (for instance) cannot increase his expected payoff by relabelling the rows in any way. In the matching pennies game, the possibilities for relabelling are 1) to assign tails to the first row and heads to the second, 2) to assign tails to both or 3) to assign heads to both.

1.4 Relationships between Equilibrium Concepts

The set of correlated measures with the marginal best-response property, MBR, contains but need not equal CE. To see this, recall that the best response set to any $\nu \in P[S_N]$ includes a pure strategy. Hence the best response set to α_N includes a pure strategy, which we will assume places all its mass on $k \in S_M$. Thus, for the constant function $\psi : S_M \to S_M$ defined by $\psi(i) = k$ for all $i \in S_M$, we have:

$$\begin{aligned}
\max_{\mu \in P[S_M]} u_M(\mu, \alpha_N) &= u_M(k, \alpha_N) \\
&= \sum_{j \in S_N} u_M(k, j)\alpha_N(j) \\
&= \sum_{j \in S_N} u_M(k, j) \sum_{i \in S_M} \alpha(i, j) \\
&= \sum_{(i,j) \in S_M \times S_N} u_M(\psi(i), j)\alpha(i, j)
\end{aligned}$$

If $\alpha \in \text{CE}$ then RHS $\leq u_M(\alpha)$, so the first condition defining MBR is satisfied. Since we could do as much taking the position of N (and therefore show the same property for player N), this is enough to prove that $CE \subseteq MBR$.

Moreover, $\max_\mu u_M(\mu, \alpha_N)$ actually equals $\sum_{i \in S_M, j \in S_N} u_M(\psi(i), j)\alpha(i, j)$ for some constant function ψ. But it is easy to produce examples where there exist non-constant ψ which outperform every constant choice. Consider the correlated measure, $\hat{\alpha}$, which assigns probability as before:

$$A = \begin{bmatrix} .20 & .30 & .00 \\ .00 & .20 & .05 \\ .15 & .00 & .10 \end{bmatrix}.$$

Further, suppose that the utility matrix for player M is the identity matrix. Then $u_M(\hat{\alpha}) = 0.5$ and

$$\hat{\alpha}_M = (0.50, 0.25, 0.25) \text{ and } \hat{\alpha}_N = (.35, .50, .15)$$

Thus,

$$\max_{\mu \in P[S_M]} u_M(\mu, \hat{\alpha}_N) = 0.5.$$

Thus $\hat{\alpha}$ is in MBR. However, if player M chooses $\psi(1) = 2, \psi(2) = 2$ and $\psi(3) = 1$, his expected payoff is:

$$\sum_{i \in S_M, j \in S_N} u_M(\psi(i), j)\alpha(\hat{i}, j) = .3 + .2 + .15 = .65$$

This is higher than $u_M(\alpha) = .5$ so $\hat{\alpha} \notin$ CE. In this example, the inequality is strict proving that $CE \subset MBR$.

The difference between a Nash equilibrium and a correlated one is that a Nash equilibrium assigns probabilities to each player's possible actions *on their own* while a correlated equilibrium assigns probabilities to all possible pairs of actions of players M and N.

To highlight the potential usefulness of correlated equilibria, consider, for example, the simple game called The Battle of the Buddies ([9]). To understand this game, consider two friends, Dave who wants to go to a hockey game and Marc who wants to go to a movie. Each, however, would rather go to his friend's choice than do his own thing alone. Thus Dave's first choice is for both of them to go to the hockey game but his second choice would be for both of them to go to a movie. (Of course, we are considering the situation in which they both choose independently.) Thus we can define the game by the following payoff matrices:

$$u_M = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}, \qquad u_N = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}.$$

As an aside, there is a legitimate question here about the plausibility or even the desirability of quantifying rewards. In order to be useful, game theory must necessarily reduce all rewards to a quantity. It cannot deal with qualitative rewards which may

nonetheless figure prominently in the decision process. This is an unfortunate limitation and one that needs very much to be kept in mind at all times. You cannot quantify a friendship. Nor can you quantify the value of a person's job, or the effect of an action on the environment. An excessive focus on the mathematics behind a problem may lead one to allow what is quantifiable to trump the qualitative aspects, to the detriment of all involved.

Returning, to the above example, let us first determine the existence of Nash equilibria. Recall that for a pair of measures, $(\hat{\mu}, \hat{\nu})$, to be a Nash equilibrium it must satisfy:

$$\begin{aligned} u_M(\hat{\mu}, \hat{\nu}) &\geq \max_{\mu \in P[S_M]} u_M(\mu, \hat{\nu}) \\ \Rightarrow 5\hat{\mu}(1)\hat{\nu}(1) + (1-\hat{\mu}(1))(1-\hat{\nu}(1)) &\geq \max_{\mu \in P[S_M]} [5\mu(1)\hat{\nu}(1) + (1-\mu(1))(1-\hat{\nu}(1))] \\ \Rightarrow \hat{\mu}(1)(6\hat{\nu}(1)-1) &\geq \max_{\mu \in P[S_M]} [\mu(1)(6\hat{\nu}(1)-1)]. \end{aligned}$$

There are therefore three possible cases allowing for the existence of a Nash Equilibrium. First, if $6\hat{\nu}(1) - 1 > 0$ then $\hat{\mu}(1)$ must be equal to 1. If, $6\hat{\nu}(1) - 1 < 0$ then $\hat{\mu}(1)$ must be equal to zero. The other case requires that $\hat{\nu}(1)$ be equal to $1/6$. Since we can do much the same using N, it is easy to verify that there exist only three Nash equilibria – with $[\mu, \nu] = [(0,1),(0,1)], [(1,0),(1,0)]$ or $[(5/6,1/6),(1/6,5/6)]$. The corresponding payoffs are $[u_M, u_N] = [1,5], [5,1]$ and $[5/6, 5/6]$. The following diagram illustrates the possible payoff region *provided we restrict ourselves to measures derived from the multiplication of marginals.*

However, the set of correlated equilibria is much larger. Recall that for a measure, $\hat{\alpha}$, to be a correlated equilibrium, it must satisfy:

$$\begin{aligned} u_M(\hat{\alpha}) &\geq \sum_{(i,j) \in S_M \times S_N} u_M(\psi(i), j)\alpha(i,j) \\ \Rightarrow 5\hat{\alpha}(1,1) + \hat{\alpha}(2,2) &\geq u_M(\psi(1),1)\alpha(1,1) + u_M(\psi(1),2)\alpha(1,2) + u_M(\psi(2),1)\alpha(2,1) \\ &\quad + u_M(\psi(2),2)\alpha(2,2) \end{aligned}$$

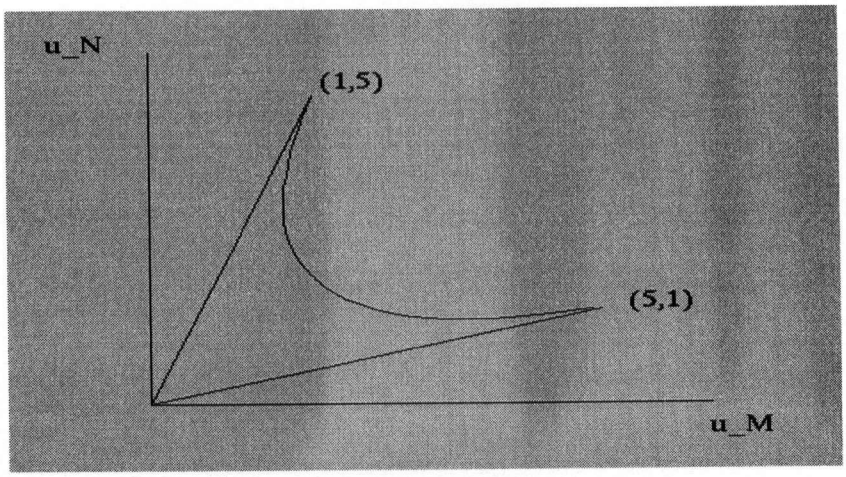


Figure 1.1: Payoff region allowing only for measures derived from the multiplication of marginals

as well as,

$$\begin{aligned} u_N(\hat{\alpha}) &\geq \sum_{(i,j)\in S_M\times S_N} u_N(i,\phi(j))\alpha(i,j) \\ \Rightarrow \hat{\alpha}(1,1)+5\hat{\alpha}(2,1) &\geq U_N(1,\phi(1))\alpha(1,1)+u_N(1,\phi(2))\alpha(1,2)+u_N(2,\phi(1))\alpha(2,1) \\ &\quad + u_N(2,\phi(2))\alpha(2,2) \end{aligned}$$

for arbitrary $\phi,\psi : \{1,2\} \to \{1,2\}$. This implies that if a measure $\hat{\alpha}$, is a correlated equilibrium then the following two conditions must be satisfied:

$$\begin{aligned} \hat{\alpha}(1,1) &\geq 5\hat{\alpha}(1,2) \\ \hat{\alpha}(2,2) &\geq 5\hat{\alpha}(2,1) \end{aligned}$$

These lead to a set of measures rather than three points. The following picture depicts the possible payoff region provided we allow for any correlated measure.

Note that these two diagrams highlight the potential usefulness of correlated measures that do not derive from the multiplication of two marginal measures. The payoff pair $(3,3)$ which seems to be the most obvious compromise is included only in the second diagram and not the first.

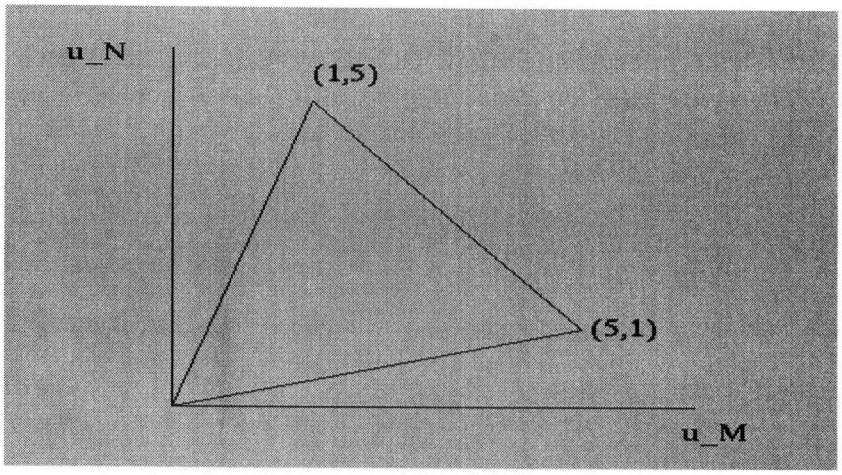


Figure 1.2: Payoff region allowing for all measures

Now, if we restrict ourselves to the set of correlated measures of product form, i.e., where $\alpha = \alpha_M^T \alpha_N$, then all three sets (CE, MBR and NE) are indistinguishable. That MBR and NE are equivalent under this restriction is obvious from the definitions. (MBR's definition is precisely that of NE in this case.) To show that CE and MBR are also equivalent, we need only show that some constant transformation, $\psi(i) = k$ for all $i \in S_M$, is an upper bound on $\sum_{i \in S_M, j \in S_N} u_M(\psi(i), j)\alpha(i,j)$ for all possible transformations $\psi : S_M \to S_M$. It would then follow that MBR=CE since the maximization in the definition of MBR turns out to be the largest constant transformation. Now, since we are in the case where α is a product measure,

$$\begin{aligned} \sum_{(i,j) \in S_M \times S_N} u_M(\psi(i), j)\alpha(i,j) &= \sum_{(i,j) \in S_M \times S_N} u_M(\psi(i), j)\alpha_M(i)\alpha_N(j) \\ &= \sum_{i \in S_M} \left(\sum_{j \in S_N} u_M(\psi(i), j)\alpha_N(j) \right) \alpha_M(i) \end{aligned}$$

Furthermore, since S_M is finite, it follows that there exists an $k \in S_M$ such that

$$\sum_{j \in S_N} u_M(k, j)\alpha_N(j) \geq \sum_{j \in S_N} u_M(i, j)\alpha_N(j) \quad \text{for all } i \in S_M$$

Thus, letting $\psi(i) = k$ for all $i \in S_M$ will give the desired upper bound and therefore CE

and MBR are equivalent.

1.5 Repeated Plays

Of course, game theory is hardly confined to one-shot games, the application of which is fairly limited. In the more involved scenario of repeated play, player M must impose a measure on his set of strategies, S_M, each time the game is played. (It is hopefully clear that actual actions are single-strategy choices, but that the choice of which action to play may very well be governed by a pre-chosen probability measure over the set of all possible actions.)

In the simplest scenario, a player may decide to use the same measure for all rounds of the game. Thus his empirical distribution of play (that is the distribution of the actual history of the game) derives from repeated draws from a fixed distribution. He may, on the other hand, decide to vary his mixed strategy at each round. (In fact, if he doesn't and the other player is at all rational then he is not likely to fare very well in the long run – unless he happens to have hit upon an equilibrium.) How, and for what reasons, a player might decide to vary his play is a question at the heart of this thesis.

The most logical idea would be for each player to attempt to learn from the history of the game. So, for instance, if t rounds have already been played, then each player has the experience of t rounds from which to draw in order to better gauge the behaviour of his opponent. For ease of notation, we can define a history matrix, h^t:

$$h^t = \begin{bmatrix} s_M^t \\ s_N^t \end{bmatrix} = \begin{bmatrix} s_M(1)...s_M(t) \\ s_N(1)...s_N(t) \end{bmatrix}$$

where $s_M(k)$ is the actual strategy played by M during round k (similarly for player N). I will denote by $S_M^t \times S_N^t$, the set of all histories of length t. Thus, M's measure or

behaviour rule for round $t+1$ may be a function of h^t. We will assume, however, that each player's payoff matrix is stationary with respect to time. The matching pennies example illustrates this notation perfectly. If M were to notice that, no matter what he chooses, nature seems always to choose heads (i.e. $s_N^t = (h, h, ..., h)$) then it would make sense for player M to do likewise.

Thus game theory becomes interested in the incorporation of learning into a game. This is, obviously, no trivial question. How do you incorporate into a mathematical model the idea that people learn from their past experience? We will look at some attempts to do just that later on in this paper. The bulk of this paper will be concerned with the long run performance of a number of different behaviour rules (policies used to determine action) that attempt to incorporate learning. More specifically, I am interested in each behaviour rule's differing convergence properties and whether or not they do actually converge to some stable equilibrium where each player eventually sticks to one measure (or set of measures) over his set of strategies.

Of course, one might still wonder what purpose there is in discussing equilibria. What reason do we have for believing that players will eventually come to decide upon measures that lead to an equilibrium? The study of game theory itself assumes a certain amount of rationality in the players. (If you take issue with that assumption then game theory will have little to offer.) The idea behind the study of equilibria is that, given time, the players will come to some compromise. It is this idea of a compromise that equilibria are meant to represent.

However, not all correlated equilibria result in great compromises. Recall, for example, the Battle of the Buddies where two equilibria resulted in one player caving in every time. So even if we can insure convergence to CE, this may not be all that desirable a result. Thus, even within CE, we may want to insure convergence to certain specific

distributions.

Questions of how to incorporate learning into a game and how to insure convergence have, of course, been around for a while. A number of methods of deciding on one's play (behaviour rules) have been developed which will be discussed in the remainder of this thesis (along with some original ones). I will be focusing specifically on their differing convergence results – what they converge to (if at all) and under what circumstances.

Chapter 2
Behaviour Rules

2.1 Overview

Though the concept of a behaviour rule has already arisen, we have until now left it somewhat informal. This needs now to be remedied.

Definition: A *behaviour rule for M* is a measure-valued mapping, $\mu : \cup_{t\geq 0}(S_M^t \times S_N^t) \rightarrow P[S_M]$ which returns a mixed strategy for M based on a given vector of observed plays. In other words, we imagine player M choosing a fixed mapping μ, and choosing as the mixed strategy to play at time $t+1$, the measure, $\mu^{t+1} = \mu(s_M(1), ..., s_M(t), s_N(1), ..., s_N(t))$.

Of course, it is quite possible to have a behaviour rule that only takes a portion of the history into account or which even ignores the history completely. For instance, in Matching Pennies, one might choose to play heads all the time no matter what happens (a constant mapping) or to play heads in the morning (first n rounds) and tails in the afternoon (next n rounds). We will, however, concentrate on history dependent rules.

Any given pair of behaviour rules (μ, ν) defines a probability distribution over all possible histories of any fixed length t. I will denote by $p(\mu, \nu)[h^t]$ – the probability that the history h^t will result given that M uses the fixed rule μ and N uses the fixed rule ν. Recall that we are assuming that each player acts independently. Similarly, if $H^t \subset S_M^t \times S_N^t$, we can define $p(\mu, \nu)[H^t]$ as the probability that a history $h^t \in H^t$ will

result given that M uses the fixed rule μ and N uses the fixed rule ν.

One means of evaluating a behaviour rule is to determine whether it guarantees a certain lower bound pay-off. To that end the following two quantities are useful. First, for any history vector, $h^t = (s_M(1), ..., s_M(t), s_N(1), ..., s_N(t))$, we can define the time-averaged realized pay-off,

$$U_M(h^t) = \frac{1}{t} \sum_{\tau=1}^{t} u_M(s_M(\tau), s_N(\tau)).$$

Second, a useful performance index associated with a particular history h^t is

$$\hat{U}_M(h^t) = \max_{\mu \in P[S_M]} u_M(\mu, \overline{\nu}(h^t))$$

where $\overline{\nu}(h^t)$ represents the empirical distribution of play of player N up to time t. In words, this represents the pay-off to M if N plays the empirical measure of N's past plays and M plays the corresponding best response.

2.2 Properties of Adaptive Behaviour Rules

To express desirable properties for M's behaviour rule μ, several criteria have been proposed. Each involves a different assumption concerning the behaviour rule of player N and attempts to insure that player M's time-averaged realized pay-off exceeds some lower bound. The first criteria, called *safety*, allows player N to use any available behaviour rule, ν, and seeks to guarantee that M receives at least his min-max pay-off. This property is obviously desirable, but perhaps not essential. It seems reasonable to think that in pursuing safety, one might be forced to ignore opportunities for much larger gains.

Safety:

A behaviour rule, $\hat{\mu}$, is said to be ϵ-*safe* for player M if there exists a $\bar{t}$ such that for any possible behaviour rule, ν, available to nature and for any $t \geq \bar{t}$ there is a subset of

histories of length t, $H^t \subset S_M^t \times S_N^t$, with $p(\hat{\mu}, \nu)[H^t] \geq 1 - \epsilon$, such that

$$U_M(h^t) + \epsilon \geq \min_{\nu} \left[\max_{\mu} u_M(\mu, \nu) \right] \quad \forall h^t \epsilon H^t.$$

A behaviour rule is said to be *safe* if it is ϵ-safe for every positive ϵ ([4]).

A second proposed criterion is *consistency.* This property is satisfied when player M does no worse than if he had played a best response against the empirical distribution. There are no less than three variations on the definition of consistency, each making different assumptions on the behaviour rule of player N. In the first definition given below, player N is assumed to employ a specific constant behaviour rule, $\hat{\nu}$. In the second definition, the behaviour rule is again assumed to be constant but now player N can use any fixed measure, $\nu \in P[S_N]$. Finally, in the definition of *universal consistency,* all restrictions on player N's behaviour rule are lifted. In each case however, the goal is the same.

Consistency:

Let $\hat{\nu}$ be a fixed measure in $P[S_N]$. A behaviour rule, $\hat{\mu}$, is said to be ϵ-*consistent against* $\hat{\nu}$ if there exists a $\bar{t}$ such that for any $t \geq \bar{t}$ there is a subset of histories of length t, $H^t \subset S_M^t \times S_N^t$, with $p(\hat{\mu}, \hat{\nu})[H^t] \geq 1 - \epsilon$, and

$$U_M(h^t) + \epsilon \geq \hat{U}_M(h^t) \quad \forall h^t \epsilon H^t$$

([4]).

If the above holds true for all fixed measures $\nu \in P[S_N]$ then $\hat{\mu}$ is said to be ϵ-*consistent.* In words, a behaviour rule is ϵ-consistent if, given that the mixed strategy played by nature is constant, M does about as well as playing a best response against the empirical distribution of play. A behaviour rule is said to be *consistent* if it is ϵ-consistent for all positive ϵ.

Finally, a behaviour rule, $\hat{\mu}$, is said to be ϵ-*universally consistent* if there exists a $\bar{t}$ such that for *any* behaviour rule ν available to nature and for any $t \geq \bar{t}$ there is a subset of histories of length t, $H^t \subset S_M^t \times S_N^t$, such that $p(\hat{\mu}, \nu)[H^t] \geq 1 - \epsilon$,

$$U_M(h^t) + \epsilon \geq \hat{U}_M(h^t), \quad \forall h^t \epsilon H^t.$$

The key change here is that if $\hat{\mu}$ is ϵ-universally consistent then ν is allowed to vary with time (i.e. there is no guarantee that $\nu^i = \nu^j$ for $i \neq j$). Again, a behaviour rule is said to be *universally consistent* if it is ϵ-universally consistent for all positive ϵ ([4]).

It is easy to show, directly from the definitions, that universal consistency implies safety. Universal-consistency of a measure, $\hat{\mu}$, implies that for all $\epsilon > 0$ there exists a $\bar{t}$ and a set of histories $H^t \subseteq S_M^t \times S_N^t$ such that, for any behaviour rule ν by player N and for any $t \geq \bar{t}$, $p(\hat{\mu}, \nu)[H^t] \geq 1 - \epsilon$ and

$$U_M(h^t) \geq \hat{U}_M(h^t) = \max_{\mu} u_M(\mu, \overline{\nu}(h^t)) \quad \forall h^t \in H^t$$

But $\max_\mu u_M(\mu, \overline{\nu}(h^t)) \geq \min_\nu [\max_\mu u_M(\mu, \nu)]$, for all $h^t \epsilon H^t$. So the above inequality implies the defining inequality for safety.

2.3 Forecasting and Response

It is useful to divide the larger set of all possible behaviour rules into two subsets. The first subset of behaviour rules attempts to predict N's play and then proceeds accordingly. The first step in this process – prediction – is the function of a forecast. A forecast is an attempt to determine how the *other* player will act (possibly expressed probabilistically) in the next round.

The most obvious forecast is the empirical average of the opponent's past play. Thus if the game has been played for t rounds and N has played strategy i, x number of

times then the forecast would simply predict that player N will play strategy i with a probability of x/t. Thus, we define

$$\overline{\nu}_j(h^t) = \frac{1}{t}\sum_{\tau=1}^{t} I_j(s_N(\tau)), \quad j = 1, 2, ..., |S_N|, \tag{2.1}$$

representing the empirical probability of playing strategy j observed in the history, h^t. Here, $I_j(p) = 1$ if $p = j$, and zero otherwise.

More generally, a forecast, γ, will generate, after round t of the game, a k-tuple, $\hat{\nu} = (\hat{\nu}_1(h^t), ..., \hat{\nu}_k(h^t))$ where $k = |S_N|$ and $\hat{\nu}_j(h^t)$ is M's forecasted probability that nature will play strategy j at time $t+1$. The forecast, of course, depends on whatever factors player M deems relevant. It would seem reasonable, however, to let the forecast depend only on the past history, $h^t \in S_M^t \times S_N^t$. Thus a forecast is a function, $\gamma : \bigcup_{t=1}^{\infty} S_M^t \times S_N^t \to P[S_N]$.

Calibration:

The idea behind a *calibrated forecast* is that eventually the empirical distribution of play of the opponent (nature) should converge to that which is predicted by the forecast (or alternatively, the forecast should adapt to fit the empirical distribution). If this occurs then the forecast is said to be well-calibrated.

Naturally, this idea can be represented mathematically. Fix $\hat{\nu} \in P[S_N]$ and a sequence of plays as recorded in history matrices $h^1, h^2, ..., h^t, ...$ Let $N(\hat{\nu}, t)$ denote the number of the first t rounds where M's forecast, $\gamma(h^t)$, generated the measure $\hat{\nu}$. This can be written as:

$$N(\hat{\nu}, t) = \sum_{\tau=1}^{t} I_{\hat{\nu}}(\gamma(h^t))$$

Let $\rho(\hat{\nu}, j, t)$ be the fraction of these rounds for which nature plays j.

$$\rho(\hat{\nu}, j, t) = \begin{cases} 0 & \text{if } N(\hat{\nu}, t) = 0 \\ \frac{1}{N(\hat{\nu},t)} \sum_{\tau=1}^{t} I_{\hat{\nu}}(\gamma(h^\tau)) I_j(s_N(\tau)) & \text{otherwise} \end{cases}$$

The forecast γ is *calibrated with respect to the history sequence* $h^1, h^2, \ldots$ if:

$$\lim_{t\to\infty} \sum_{\hat{\nu}} \mid \rho(\hat{\nu}, j, t) - \hat{\nu}_j \mid \frac{N(\hat{\nu}, t)}{t} = 0 \quad \forall j \in S_N$$

where the summation is taken over all possible $\hat{\nu}$ ([3]).

In other words, calibration merely insists that, in the long term, the forecast agrees with reality. However, it is clear that a good forecast need not guarantee a good behaviour rule or even that a good forecast is necessary for a good behaviour rule. The following are a couple of forecast-based behaviour rules that seek to make good use of forecasts in order to incorporate some form of learning into a player's behaviour.

Fictitious Play:

One of the most popular behaviour rules is that of Fictitious Play (FP) which attempts to make an educated guess (updated after each round) at the measure chosen by the opponent and then plays a best response to this guess. Player M makes some initial guess at nature's measure, ν_0, and then proceeds to modify it using the empirical probability distribution, $\overline{\nu}(h^t)$, which is accumulated as the game progresses. Mathematically, this "educated guess" or *forecast rule*, takes the following form:

$$\gamma(h^t) = \frac{n_0}{n_0 + t}\nu_0 + \frac{t}{n_0 + t}\overline{\nu}(h^t)$$

where n_0 is a fixed number ([4]). The FP behaviour rule is then given by

$$\mu^{t+1} = argmax_{\mu \in P[S_M]} u_M(\mu, \gamma(h^t))$$

(see definition of best response, chapter 1, pg.4). In the cases where there is more than one argmax, recall from chapter 1, pg. 4, the rule for deciding on a pure strategy.

As time goes on, FP places less and less weight on the initial guess, ν_0, and more and more weight on the accumulated empirical distribution, $\overline{\nu}(h^t)$. The long term effect

is that the forecast rule eventually mirrors the empirical distribution of past play by the opponent (and is thus calibrated). As time goes on, therefore, FP begins to behave more and more like a best-response to the forecast $\gamma(h^t) = \overline{\nu}(h^t)$, the empirical average of play. Thus we have a form of learning incorporated into the behaviour rule. This rule is deterministic in the sense that although it forecasts a probability for each strategy available to the opponent, its output is a pure strategy.

Though FP is consistent against certain measures it is not safe, as Fudenberg and Levine have shown ([4]). Therefore it is not universally-consistent either. Fudenberg and Levine did show that, given a situation where switching between actions becomes more and more infrequent, FP does turn out to be consistent (see the Shapley game discussed below). The lack of universal consistency has consequences for the convergence properties of FP which will be discussed later.

Past Response:

The past response behaviour rule, PR, is much like FP. The real difference occurs in the method of forecast. Suppose player M plays strategy j in round t. Whereas FP's forecast is based on the empirical distribution of the entire game up to the present, PR forecasts based only on the empirical distribution of those plays by N that followed a round in which M played j –that is, on observed "past responses" of player N to action j. Like FP, it would then place positive measure only on those actions which are a best response to this empirical distribution. Again an initial guess, ν_0, is needed. Assuming $s_M(t) = j$, PR's forecast rule for round t+1 has the following form:

$$\gamma(h^t) = \frac{n_0}{n_0 + t}\nu_0 + \frac{t}{n_0 + t}\overline{\nu}(h_j^t)$$

where n_0 is a fixed number and $\overline{\nu}(h_j^t)$ is the empirical distribution of play by N based

only on those rounds $\tau \leq t$, where $s_M(\tau - 1) = j$. Specifically, for all $k \in S_N$,

$$\overline{\nu}(h_j^t)[k] = \frac{1}{T}\sum_{\tau=2}^{t} I_j(s_M(\tau-1))I_k(s_N(\tau)) \quad \text{where } T = \sum_{\tau=1}^{t} I_k(s_M(\tau)).$$

PR therefore is an attempt to account for learning in the opponent while at the same time learning oneself. For the same reasons as FP, however, PR is neither safe nor universally-consistent. This being said, PR does seem to do better against the other behaviour rules (at least in the Shapley game – see Chapter 5) and may be an example of how insuring consistency requires that one forego certain opportunities for greater gain. PR, though doing better against the other behaviour rules, when played against itself has a much lower payoff than any of the others. Thus it has the possibility of much greater gain and the risk of much greater loss.

2.4 Forecast-free Behaviour Rules

Both FP and PR depend on h^t indirectly through an explicit forecast, $\gamma(h^t)$. However, an intermediate forecast is not essential as the next four behaviour rules demonstrate. Though they all rely on the history of the game, they never directly attempt to forecast the measure of the other player.

κ-exponential Fictitious Play:

κ-exponential FP, an invention of Fudenberg and Levine, assigns probability to each strategy, $i \in S_M$, in the following way:

$$\mu(h^t)[i] = \frac{w_i \exp\left[\kappa u(i, \overline{\nu}(h^t))\right]}{\sum_{j \in S_M} w_j \exp\left[\kappa u(j, \overline{\nu}(h^t))\right]}$$

where $u(i, \overline{\nu}(h^t)) = u_M(\delta_i, \overline{\nu}(h^t))$ (see Chapter 1, equation 1.1). Here, $w_1, ..., w_{|S_M|}$ is a collection of positive weights independent of time t and chosen in advance ([4]).

At first glance, this behaviour rule bears scant resemblance to FP. Looking for similarities, the first most obvious is that, like FP, it also depends on the empirical distribution of player N's play, namely $\overline{\nu}(h^t)$. In fact, closer examination shows that κ-exponential FP is a smoothing of the best-response strategy of FP that places equal weight on all maximizers. In other words, for κ sufficiently large and for fixed weights, $w_i > 0$, this behaviour rule assures that player M will play an ϵ-best response to $\overline{\nu}(h^t)$ with high probability and the remaining strategies with small probability. To see this, consider equal weights, w_i, and define $u(i, \overline{\nu}(h^t)) = u_i$ for ease of notation. The best response would simply be the argmax over i of $\{u_i : i = 1, .., |S_M|\}$. κ-exponential play however now has the form:

$$\begin{aligned}
\mu(h^t)[i] &= \frac{\exp[\kappa u_i]}{\sum_j \exp[\kappa u_j]} \\
&= \frac{1}{\sum_j \exp[\kappa(u_j - u_i)]} \\
&= \frac{1}{\sum_{j:u_j=u_i} 1 + \sum_{j:u_j<u_i} \exp[-\kappa(u_i - u_j)] + \sum_{j:u_j>u_i} \exp[\kappa(u_j - u_i)]}
\end{aligned}$$

Now, if we let κ go to infinity, we will have two options. If i is a best response then the third sum will be empty and the second sum will go to zero, leaving only the first sum. If i is not a best response then the third sum will be non-empty forcing $\mu(h^t)[i]$ to zero. Thus,

$$\lim_{\kappa \to \infty} \mu(h^t)[i] = \begin{cases} \frac{1}{\sum_{j:u_j=u_i} 1} & \text{if } i \text{ is a best response} \\ 0 & \text{otherwise} \end{cases}$$

So, as κ goes to infinity, κ-exponential FP converges to the best response measure that places equal weight on all maximizers. Unlike FP, κ-exponential FP does allow for other strategies to be played if only with small probability. This is called a *tremble* from the best-response. Exponential FP does, in some sense, still use the empirical distribution of play as a forecast even though it is not explicitly stated as such.

The major advantage of the κ-exponential FP over standard FP is that M's measure

now depends smoothly on the empirical distribution. Thus, since the empirical distribution adjusts as the inverse of the sample size, M's play cannot oscillate wildly. As Fudenberg and Levine have shown, it is this property of κ-exponential FP which causes it to be ϵ-universally consistent ([4]).

Regrets 1:

This behaviour rule was developed by Andreu Mas-Collel and Sergiu Hart. The idea is to define a function that in some sense measures the "regret" that M has for having played one strategy in the past rather than another. Suppose the game has already been played for t rounds. Then, given any two strategies $j, k \in S_M$, the regret function for player M is defined in the following way:

$$R_M^t(j,k) = \left[\frac{1}{t} \sum_{\tau \leq t : s_M(\tau) = j} \left(u_M(k, s_N(\tau)) - u_M(j, s_N(\tau)) \right) \right]^+$$

where $s_N(\tau)$ is the strategy played by N at time (or round) τ and $[z]^+ = \max z, 0$ ([8]).

This quantity measures the difference in payoff between what M received and what he could have received if he had consistently played the strategy k whenever he had, in the past, played strategy j. Note that $R_M^t(j,j) = 0$. (This function, R_M, obviously requires a certain amount of knowledge about M's payoff function which may or may not be reasonable depending on the application.) Computing this for every possible strategy $j \in S_M$ yields a $|S_M| \times |S_M|$ matrix, R_M^t, where the (j,k)th entry is $R_M^t(j,k)$ as defined above.

We can convert this into a stochastic matrix, P_M^t, by dividing by an appropriate constant, κ, and replacing the diagonal entries by 1 minus the sum of their corresponding

rows. In other words,

$$P_M^t = I + \frac{1}{\kappa}\left\{R_M^t - diag(R_M^t e)\right\} \tag{2.2}$$

where e is a column of ones and κ – which is independent of both time and history – is chosen to insure that the sum of the non-diagonal entries in each row is strictly less than one.

The regrets 1 behaviour rule then proceeds in the following manner. Given that $S_M(t) = j$, the measure or behaviour rule chosen by player M is simply the jth row of the matrix P_M^t. In detail,

$$\mu^{t+1}(k) = \frac{1}{\kappa} R_M^t(j,k) \quad \forall k \neq j \tag{2.3}$$

$$\mu^{t+1}(j) = 1 - \sum_{k \epsilon S_M : k \neq j} \mu^{t+1}(k) \tag{2.4}$$

Note that unlike FP, this rule assigns positive probability to all strategies having positive regret as well as to the previously played strategy. This differs from fictitious play which assigns positive probability only to those strategies that are a best response.

This regret 1 rule however is not universally consistent, so Mas-Collel and Hart derived the following method, also based on regrets, that is universally consistent.

Regrets 2:

The idea behind this method is to find an invariant vector, q, to the probability matrix P_M^t. That is, q^t must satisfy the following equation:

$$q^t P^t = q^t \tag{2.5}$$

where q^t is a row vector of length $|S_M|$ whose components sum to one. That such a vector q^t exists is a result of the following theorem (taken from Isaacson and Madsen, [5]):

Theorem: All finite stochastic matrices P have 1 as an eigenvalue and, moreover, there exist non-negative eigenvectors corresponding to $\lambda = 1$.

In fact, Isaacson and Madsen actually show that there is a left-eigenvector corresponding to $\lambda = 1$ that has non-negative components that sum up to one ([5]). (Note that the subscript M (or N) has been dropped on R, P and q for ease of notation.)

By multiplying through by κ, equation 2.5 can be written as:

$$\begin{aligned} \kappa q^t &= \kappa q^t I + q^t \left[R^t - diag(R^t e)\right] \\ \Leftrightarrow 0 &= q^t \left(R^t - diag(R^t e)\right) \end{aligned} \tag{2.6}$$

The regrets 2 behaviour rule assigns probabilities to playing each strategy $i \in S_M$, based on solutions to the above equation. That is,

$$\mu^{t+1}(i) = q^t(i) \quad \forall i \in S_M$$

for some q^t satisfying equation 2.6.

Both regrets behaviour rules have an interesting interpretation in terms of Continuous-time Markov Chain (CTMC) theory. If we let $Q^t = R^t - diag(R^t e)$ then Q^t has the properties of an infinitesimal generator of a CTMC. That is,

$$-Q^t(i,i) = \sum_{j \in S_M : j \neq i} Q^t(i,j) \quad \forall i \in S_M.$$

So we can interpret the regret, $R^t(j,k)$, as the rate of discarding strategy j in favour of strategy k. Therefore solving equation 2.6 is equivalent to solving for the stationary distribution of a CTMC. In essence, Regrets 2 runs a fictitious CTMC, between each round of the game, and uses its stationary distribution as the measure for the following round.

In terms of this matrix Q^t, the stochastic matrix, P^t, has a very simple form:

$$P^t = I + \frac{1}{\kappa} Q^t.$$

Interestingly, this is precisely the formulation for the stochastic matrix derived by the uniformization of the CTMC process (see [5] or [6]).

Regrets 2 has been shown by Mas-Collel and Hart to be universally consistent.

2.5 Modified Regrets

The idea for this behaviour rule stemmed from the results of the Battle of the Buddies game. When the game was played using any of the above behaviour rules (Regrets 2 not included due to programming difficulties), the result was that one of the two friends caved in to the wishes of the other every time. Which one ended up caving was entirely dependent on what happened in the first few rounds. This hardly seems satisfactory though not entirely unexpected. Regrets1 for instance only changes action if there is a positive regret *while holding the other player's action constant.* Now if we assume that player N chooses his preferred action then there will never be a positive regret for player M, as his choice of payoff is between one (if he caves in) and zero (if he doesn't). Thus, M will inevitably cave in even though his gain will remain one by not switching and might be 5 if he switched and convinced his friend to do likewise.

In other words, all these behaviour rules do not take into account the ability of one player to affect the action of the other. This, I think, is a major defect but not one that is easily overcome. For instance, how does one go about including in one's idea of a regret, the fact that if one had played differently one might have been able to change the action of one's opponent? In the standard Regrets-based behaviour rule, the regret for not having played action k when one actually played j is simply a matter of the difference between what one would have received if one had played k and what one actually did receive, *assuming the other player does not deviate from his play.* Somehow one would like to also incorporate into the idea of a regret, $R(j,k)$, some means of determining how

much one could possibly receive if, by switching from strategy j to k, one also persuaded one's opponent to switch as well.

It is this idea that motivated the following variant of Regrets 1 which I will call Modified Regrets (MR). MR also looks at the difference between what one could have received if one had played strategy k in the past when one actually played strategy j but now we allow the strategy of nature to vary over the whole set S_N. Thus, if j was played in round t then the modified regret is defined in the following way:

$$R_M^t(j,k) = \left[\sum_{\tau \leq t : s_M(\tau)=j} \{u_M(k, s_N(\tau)) - u_M(j, s_N(\tau))\} + \frac{M_1 + M_2}{|M_1| + |M_2|} \sum_{\tau \leq t : s_M(\tau)=j} \sum_{i \in S_N} \{u_M(k,i) - u_M(j,i)\} \right]^+ \quad (2.7)$$

where $M_1 = max(U_M)$ and $M_2 = min(U_M)$. Note that the first sum is simply the "normal" regret from the Regrets 1 behaviour rule. The second sum allows player M to account for losses arising because he has not tried to force the other player to switch strategy. The weighting on the second sum turns out to be useful since MR is of little benefit when the game is zero-sum. Thus, with the above weighting, MR reduces to normal regrets 1 when we have a zero-sum game.

Unfortunately, the above MR behaviour rule causes both players to be too stubborn in the Battle of the Buddies game (whereas, with the other rules, they are too compliant) so some idea of a compromise has to be introduced. The necessary twist is to have player M use MR unless his own payoff has diminished over the last two rounds due to a change in strategy by player N. This turn of events is likely to occur because player N sees some sort of benefit to playing his new strategy. Thus in such a scenario, player M, for that round only, responds with a best response to the last strategy played by his opponent. Results from this version of MR will be discussed later but turn out to be fairly encouraging.

Chapter 3
Blackwell's Approachability Theorem

3.1 Overview

Before analyzing the performance of the different behaviour rules, it is useful to make a short digression into a versatile and abstract result called Blackwell's Approachability Theorem (B.A.T.). Though only explicitly used in the proof of the convergence results of Regrets 2, B.A.T. also formed the spring board for Regrets 1 which in turn led to Modified Regrets.

In order to give an intuitive understanding of this theorem, we need to define a general "pay-off" function, $v_M(s^t)$ which is the payoff to player M when the players play $s^t = (s_M(t), s_N(t))$. This $v_M(s^t)$ need not be a scalar, so in this general setting, we define it as a vector (or matrix) in R^L. We also define a generalized time-averaged pay-off function, $V_M(h^t) = \frac{1}{t}\sum_{\tau \leq t} v_M(s^\tau)$, also a vector in R^L. Thus, $v_M(s^\tau)$ is the realized pay-off in round τ and $V_M(h^t)$ is the realized, time-averaged pay-off for a particular history of length t. An underlying assumption in the following discussion is that all pay-offs lie in the same bounded subset of R^L.

The question that B.A.T. attempts to answer concerns whether or not it is possible for M to insure that $V_M(h^t)$ approaches a pre-determined set C with probability one.

Blackwell's approach is to impose a condition. Suppose that for any $V_M(h^t) \notin C$ there exists a measure $\mu \in P[S_M]$ (dependent on $V_M(h^t)$) such that:

$$\left[v_M(\mu, j) - proj_C V_M(h^t)\right] \vec{n} \leq 0 \quad \forall j \in S_N$$

where $\vec{n} = V_M(h^t) - proj_C V_M(h^t)$ (that is, the outward normal of C at the closest point in C to $V_M(h^t)$) and $v_M(\mu, j) = v_M(\mu, \delta_j)$ (see Chapter 1, equation 1.1).

Geometrically, this implies that for all $s_N \in S_N$, the expected-payoff in the next round, $v_M(\mu, s_N)$, will be, relative to the hyperplane through $proj_C V_M(h^t)$, on the same side as C (as demonstrated in Figure 3.1 below). This is called the Blackwell Condition. The dependence of the choice for μ on $V_M(h^t)$ and the set C is clear but will not be specifically indicated in order to ease the notation.

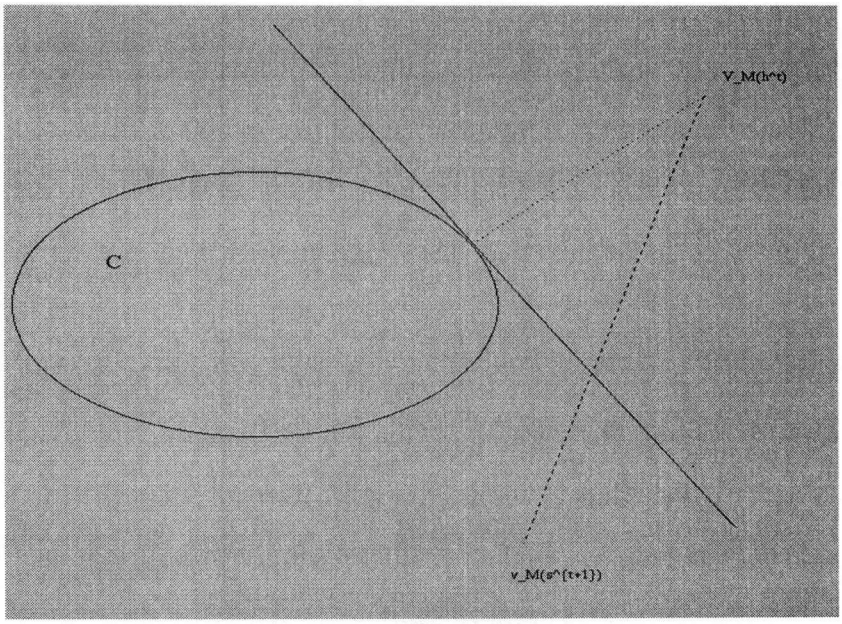


Figure 3.1: Geometrical representation of Blackwell's condition

Blackwell's theorem asserts that, given *any* predetermined set C, if the above condition can be satisfied then M can insure that $V_M(h^t)$ approaches C with probability one.

Before delving into the actual proof, I will give a somewhat geometrical interpretation of why B.A.T. makes sense. Given any $\mu \in P[S_M]$ and $j \in S_N$, the expected value of $V_M(h^{t+1})$, when N plays strategy $j \in S_N$ and M uses measure μ as a behaviour rule, can be written as follows:

$$V_M(\mu, j)[h^{t+1}] = \frac{t}{t+1} V_M(h^t) + \frac{1}{t+1} v_M(\mu, j).$$

Furthermore, if μ is chosen to satisfy the Blackwell condition and t is sufficiently large, then for all $j \in S_N$, $V_M(\mu, j)[h^{t+1}]$ is contained in the circle with center $proj_C V_M(h^t)$ and radius $||V_M(h^t) - proj_C V_M(h^t)||$. This follows from the fact that it is a convex combination of $V_M(h^t)$ and $v_M(\mu, j)$ (which lies on the same side of the hyperplane as C) and t is large.

Using this fact, it is easy to show that the euclidean distance between C and $V_M(h^t)$, denoted by $d(V_M(h^t), C)$, cannot increase very fast as t goes to infinity (since larger t favours smaller "correction" from $V_M(h^t)$). Indeed,

$$\begin{aligned} d(V_M(h^{t+1}), C) - d(V_M(h^t), C) &\leq ||V_M(\mu, s_N)[h^{t+1}] - proj_C V_M(h^t)|| - ||V_M(h^t) - proj_C V_M(h^t)|| \\ &= ||\frac{t}{t+1} V_M(h^t) + \frac{1}{t+1} v_M(\mu, s_N) - proj_C V_M(h^t)|| \\ &\quad - ||V_M(h^t) - proj_C V_M(h^t)|| \end{aligned}$$

This last expression obviously goes to zero as t goes to infinity (since all pay-offs are bounded). A precise computation shows that $d(V_M(h^t), C)$ doesn't merely converge to some limit greater than zero and remain bounded away from C but actually goes to zero itself. Thus by the Law of Large numbers, since the distance to C *in expectation* goes to zero so must the realized distance between the time-averaged pay-off and the set C. Blackwell's more detailed proof is outlined in the next two sections which may be skipped if the reader so desires.

3.2 Auxiliary Results

Blackwell's proof of his Approachability Theorem depends heavily on a version of the Strong Law of Large Numbers (SLLN), due to Blackwell himself, which he uses in order to prove a key lemma. These (Blackwell's version of the SLLN and the lemma) I will state first before diving in to the actual proof of B.A.T.

Blackwell's version of the SLLN is as follows:

Theorem: If $B_1, B_2, ...$ is a sequence of random variables such that $|B_k| \leq 1$ and there exists a $u > 0$ such that

$$E[B_k|B_1, ..., B_{k-1}] \leq -u \max(|B_k||B_1, ..., B_{k-1})$$

then for all $\sigma \in R$

$$Prob\{B_1 + ... + B_k \geq \sigma \text{ for some } k\} \leq \left(\frac{1-u}{1+u}\right)^{\sigma}.$$

I will not attempt to prove the above theorem here but refer the reader to the bibliography ([2]). For the following lemma, however, I will provide the proof, essentially as Blackwell presented it ([1]).

Lemma 3.2: Let $\delta^1, \delta^2, ...$ be a sequence of random variables for which there exist constants a, b and c such that, with probability one,

$$0 \leq \delta^t \leq a \quad \forall t \in N \tag{3.1}$$

$$|\delta^{t+1} - \delta^t| \leq \tfrac{b}{t+1} \quad \forall t \in N \tag{3.2}$$

$$E[\delta^{t+1}|\delta^1, \delta^2, ..., \delta^t] \leq (1 - \tfrac{2}{t+1})\delta^t + \tfrac{c}{(t+1)^2} \quad \forall t \in N \tag{3.3}$$

Then $\delta^t \to 0$ almost surely (a.s.). Indeed, for every $\epsilon > 0$, there exists a T_0 depending

only on ϵ, a, b and c such that for any δ^t satisfying equations 3.1, 3.2 and 3.3, we have

$$Prob\{\delta^t \geq \epsilon \text{ for some } t \geq T_0\} \leq \epsilon. \tag{3.4}$$

To see that this conclusion is truly equivalent to almost sure convergence, first assume almost sure convergence and fix $\eta > 0$. Then

$$\begin{aligned} \{\omega : \delta^t \to 0\} &= \{\omega : \forall \epsilon \in (0, \eta) \quad \exists T \text{ s.t. } \forall t \geq T, \delta^t < \epsilon, \} \\ &= \cap_{\epsilon \in (0,\eta)} \left[\cup_{T \in N} \{\omega : \delta^t < \epsilon, \quad \forall t \geq T\} \right] \end{aligned}$$

As ϵ decreases, the above sets that form the intersection shrink in a nested fashion. Therefore,

$$P[\delta^t \to 0] = \inf_{\epsilon \in (0,\eta)} P\left[\cup_T \{\delta^t < \epsilon, \quad \forall t \geq T\}\right] \tag{3.5}$$

And as T increases, the above sets that form the union also increase in a nested fashion so,

$$P\{\delta^t \to 0\} = \inf_{\epsilon \in (0,\eta)} \left[\sup_T P\{\delta^t < \epsilon, \quad \forall t \geq T\} \right]$$

Assuming $\delta^t \to 0$ a.s. implies that for all $\epsilon > 0$,

$$\sup_T P\{\delta^t < \epsilon, \quad \forall t \geq T\} = 1$$

Hence there exists a T_0 such that $P\{\delta^t < \epsilon, \quad \forall t \geq T_0\} \geq 1 - \epsilon$ which is equivalent to equation 3.4.

Conversely, if equation 3.4 holds, then for each $\epsilon \in (0, \eta)$ we can find $T = T_\epsilon$ such that

$$P\{\delta^t < \epsilon, \quad \forall t \geq T_\epsilon\} > 1 - \epsilon.$$

Consequently, from equation 3.5,

$$P\{\delta^t \to 0\} \geq \inf_{\epsilon \in (0,\eta)} [1 - \epsilon] = 1 - \eta.$$

Since this is true for any $\eta > 0$, equation 3.4 implies almost sure convergence.

Proof of Lemma:

Fix any $\epsilon > 0$ and any $t_0 \geq 2$. We first prove that there exists a $t_1 \geq t_0$ depending only on t_0, a, b and c such that

$$Prob\{\delta^t \geq \epsilon/2 \text{ for } t_0 \leq t \leq t_1\} < \epsilon/2. \tag{3.6}$$

To see this, define for $t \geq t_0$,

$$\alpha^t = \begin{cases} \delta^t & \text{if } \delta^i > 0 \text{ for } t_0 \leq i \leq t \\ 0 & \text{otherwise} \end{cases}$$

Then, $\alpha^t < \epsilon/2$ implies that $\delta^i < \epsilon/2$ for some $i \in [t_0, t]$.

By equation 3.3, for $t \geq t_0$,

$$E[\alpha^t | \alpha^{t_0}, ..., \alpha^{t-1}] \leq \left(1 - \frac{2}{t}\right) \alpha^{t-1} + \frac{c}{t^2}$$

It is clear, therefore, that this implies that the sequence of constants $E(\alpha^t)$, eventually decreases. To see that it does in fact converge to zero, let us define $e^t = E[\alpha^t]$. Then,

$$\begin{aligned} e^t &\leq \left(1 - \frac{2}{t}\right) e^{t-1} + \frac{c}{t^2} \\ \Rightarrow t(t-1)e^t &\leq (t-1)(t-2)e^{t-1} + \frac{c(t-1)}{t}. \end{aligned}$$

If we let $\beta^t = t(t-1)e^t$ then

$$\beta^t \leq \beta^{t-1} + c.$$

Fixing t_0, it follows that $\beta^t - \beta^{t_0} \leq (t - t_0)c$. Thus, substituting for β^t, we get

$$e^t \leq \frac{t_0(t_0-1)e^{t_0}}{t(t-1)} + \frac{(t-t_0)c}{t(t-1)} \tag{3.7}$$

$$\Rightarrow E[\alpha^t] \leq \frac{c_1}{t(t-1)} + \frac{c}{t} \tag{3.8}$$

which, of course, implies that $E[\alpha^t]$ converges to zero as $t \to \infty$ at a rate depeding only on t_0, a, c.

In other words, for all $\epsilon > 0$ there exists a $T > 0$ such that for all $t \geq T$,

$$\left(\frac{\epsilon}{2}\right)^2 > E(\alpha^t) \geq \int_{\{\alpha^t \geq \epsilon/2\}} \alpha^t dP \geq \frac{\epsilon}{2} P\{\alpha^t > \epsilon/2\}.$$

Therefore, there exists a $t_1 \geq t_0$ depending only on t_0, a, c such that

$$Prob\{\alpha^{t_1} < \epsilon/2\} > 1 - \epsilon/2.$$

This completes the first step in proving the lemma – demonstrating that equation 3.6 holds under the conditions stated in the lemma.

In order to continue, we define the following double sequence. (A new random sequence with index k associated with each fixed t.) For every t, k with $t \leq k$ we define the variable z_{tk} as follows:

$$z_{tk} = \begin{cases} 0 & \text{unless } \delta^{t-1} < \epsilon/2 \text{ and } \delta^t \geq \epsilon/2 \\ \delta^k & \text{if } \delta^{t-1} < \epsilon/2 \text{ and } \delta^i \geq \epsilon/2 \text{ for all } i \text{ such that } t \leq i \leq k \\ \delta^{k_0} & \text{for } k \geq k_0, \text{ if } \delta^{t-1} \leq \epsilon/2 \text{ and } \delta^i \geq \epsilon/2 \text{ for all } i, \text{ such that } t \leq i < k_0 \\ & \text{and } \delta^{k_0} < \epsilon/2 \end{cases}$$

Perhaps it would be helpful to unpack this definition a little bit. We can first break it down into two possibilities – either there is an upcrossing of the level $\epsilon/2$ by the function δ^i between round $t-1$ and round t or there isn't. If there isn't then z_{tk} is equal to zero for all k. If there is then again we have two possibilities – either δ^i remains above $\epsilon/2$ for all i such that $t < i \leq k$ or else it dips below $\epsilon/2$ at some round k_0, $t < k_0 < k$. These two possibilities account for the second two parts of the definition of z_{tk} (see diagram below).

Thus, z_{tk} monitors the upcrossings of $\epsilon/2$ made by δ^k. If an upcrossing is made at

time t, then z_{tk} keeps track of the value of δ^k up until it once again dips below $\epsilon/2$. z_{tk} then holds the constant value associated with the first value below $\epsilon/2$.

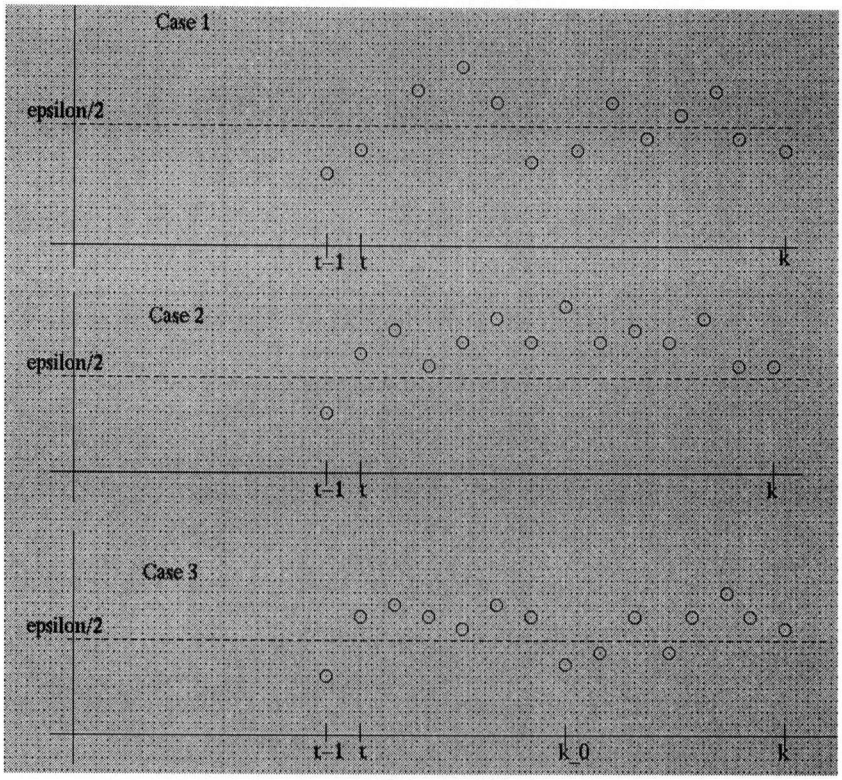


Figure 3.2: Three cases for z_{tk}

It follows from the above definition that if $\delta^k \geq \epsilon$ for some $k \geq t_1$ then either (i) $\delta^t \geq \epsilon/2$ for all t such that $t_0 \leq t \leq t_1$, or (ii) $z_{tk} \geq \epsilon$ for some $t \in [t_0, k]$. Indeed, if (i) fails then we must have $\delta^{\hat{t}} < \epsilon/2$ for some $\hat{t}$ with $t_0 \leq \hat{t} \leq t_1$ and there must exist an upcrossing somewhere between $\hat{t}$ and k. Therefore, there exists a $\bar{t} \in [t_0, k]$ such that $z_{\bar{t}k} = \delta^k \geq \epsilon$, i.e., (ii) must hold.

Now, we have already shown that outcome (i) is not very likely, i.e.,

$$Prob\{\delta^t \geq \epsilon/2 \ \forall t \in [t_0, t_1]\} < \epsilon/2.$$

Therefore, if we can show that the same is true for case (ii), i.e.,

$$Prob\{z_{tk} \geq \epsilon \text{ for some } t \geq t_0 \text{ and some } k \geq t\} \leq \epsilon/2 \tag{3.9}$$

then both options would be nicely bounded which would imply that $Prob\{\delta^k \geq \epsilon \text{ for some } k \geq t_1\} \leq \epsilon$, completing the proof.

In order to prove equation 3.9, fix $t \geq t_0$. If t is not associated with an upcrossing, then $z_{tk} = 0$ for all $k \geq t$. Hence such a t makes no contribution to the probability on the left in equation 3.9 and so we may assume that t is associated with an upcrossing. In this case, define $\beta_k = z_{tk} - z_{t,k-1}, k \geq t$ (with $\beta_t = 0$). Two cases arise. If $z_{t,k-1} \geq \epsilon/2$ we have both $z_{t,k-1} = \delta^{k-1}$ and $z_{tk} = \delta^k$, so $\beta_k = \delta^{k-1} - \delta^k$ and

$$\begin{aligned}
E[\beta_k | z_{tt}, \beta_t, ..., \beta_{k-1}] &= E[\delta^k - \delta^{k-1} | \delta^t, ..., \delta^{k-1}] \\
&= E[\delta^k | \delta^t, ..., \delta^{k-1}] - \delta^{k-1} \\
&\leq \left(1 - \frac{2}{k}\right)\delta^{k-1} + \frac{c}{k^2} - \delta^{k-1} \\
&= \frac{-2}{k}\delta^{k-1} + \frac{c}{k^2} \\
&\leq \frac{-\epsilon}{2k} \quad \text{for large enough } k \\
&\leq \frac{-\epsilon}{2b} \max(|\beta_k| | \beta_t, ..., \beta_{k-1}) \quad \text{since } |\beta_k| < b/k. \qquad (3.10)
\end{aligned}$$

If $z_{t,k-1} < \epsilon/2$ then $\beta_k = 0$ since we can't be in the second case in the definition of z_{tk}. So, in either case, equation 3.10 is satisfied.

We now use the aforementioned variation on the SLLN in the following manner. If we let $B_k = (t/b)\beta_{k+1}$, and we define $u = \epsilon/2b$ then, by equation 3.2:

$$\beta_k < \frac{b}{k} \quad \forall k \geq t \quad \Rightarrow \quad B_k < \frac{t}{k+1} \leq 1 \quad \forall k \geq t$$

and, by equation 3.10,

$$\begin{aligned}
E[B_{k-1} | B_t, ..., B_{k-2}] &= E[\frac{t}{b}\beta_k | \beta_{t+1}, ...\beta_{k-1}] \\
&\leq -\frac{\epsilon}{2b} \max(\frac{t}{b}|\beta_k| | \beta_{t+1}, .., \beta_{k-1}) \\
&= -\frac{\epsilon}{2b} \max(|B_{k-1}| | B_t, ..., B_{k-2})
\end{aligned}$$

Thus, this definition of B_k satisfies the hypotheses of the above theorem allowing us to assert that, for $u = \epsilon/2b$,

$$Prob\{B_t + ... + B_{k-1} \geq \sigma \text{ for some } k\} \leq \left(\frac{1-u}{1+u}\right)^{\sigma}.$$

But

$$\begin{aligned} B_t + ... + B_{k-1} &= \frac{t}{b}\beta_{t+1} + ... + \frac{t}{b}\beta_k \\ &= \frac{t}{b}[(z_{t,t+1} - z_{t,t}) + ... + (z_{t,k} - z_{t,k-1})] \\ &= \frac{t}{b}(z_{t,k} - z_{t,t}) \end{aligned}$$

Thus, taking $\sigma = st/b$ above, we have

$$Prob\{z_{tk} - z_{tt} > s \text{ for some } k \geq t\} < r^{ts} \quad \text{where } r = \left(\frac{1-u}{1+u}\right)^{1/b}.$$

Recall that $z_{tt} = \delta^t$. Therefore, since $\delta^{t-1} < \epsilon/2$ (by assumption) and $t \geq t_0$, we can insure that $z_{tt} < 3\epsilon/4$ (by taking t_0 large enough and using equation 3.2), so that $z_{tk} \geq \epsilon$ for some k implies that $z_{tk} - z_{tt} > \epsilon/4$. Thus,

$$Prob\{z_{tk} \geq \epsilon \text{ for some } k \geq t\} \leq (r^{\epsilon/4})^t.$$

This, finally implies that,

$$Prob\{z_{tk} \geq \epsilon \text{ for some } k \geq t, t \geq t_0\} \leq \sum_{t=t_0}^{\infty}(r^{\epsilon/4})^t = \frac{r^{\epsilon(t_0/4)}}{1 - r^{\epsilon/4}}$$

Now $r < 1$, so for large enough t_0, we have equation 3.9. Thus both options are nicely bounded completing the proof of the lemma.

One final theorem is required before we can complete the proof of B.A.T.

Theorem 1: For every closed subset $C \in R^N$, there exists a function $\pi_C : R^N \to R^N$, with the property that

$$< z - \pi_C(z), y - \pi_C(z) > \quad \leq \quad \frac{1}{2}|y - \pi_C(z)|^2 \quad \forall y \in C.$$

Indeed, it suffices to take for π_C any selection of the nonempty-valued (since C is closed) multifunction $z \to argmin\{|z-y| : y \in C\}$.

Proof of Theorem 1: The proof is a straightforward application of the properties of an inner product. If $z \in R^n$ and $\pi = \pi_C(z)$ is a nearest point in C to z, then

$$|z-\pi|^2 \leq |z-y|^2 \quad \forall y \in C$$

Thus,

$$\begin{aligned} |z-\pi|^2 &\leq |(z-\pi)+(\pi-y)|^2 \\ &= |z-\pi|^2 + 2 < z-\pi, \pi-y > + |\pi-y|^2 \\ \Rightarrow \quad < z-\pi, y-\pi > &\leq \frac{1}{2}|y-\pi|^2 \quad \forall y \in C \end{aligned}$$

completing the proof.

3.3 Concise Statement and Detailed Proof

Before we can even make sense of the statement of BAT, we will need the following definition.

Definition: Let C be a set in L-space. We shall say that C is *approachable* if there exists a behaviour rule (for player M) such that for every sequence of mixed strategies available to player N, the sequence $\{\delta^t : t \in L\}$ converges to zero almost surely. Here δ^t is the squared distance to C of the empirical $V_M(h^t)$, i.e.,

$$\delta^t = \min\{|V_M(h^t) - \pi|^2 : \pi \in C\}. \tag{3.11}$$

That is, for every $\epsilon > 0$ (and every possible behaviour rule by player N) there is a T_0 such that

$$Prob\{\delta^t \geq \epsilon \text{ for some } t \geq T_0\} \leq \epsilon.$$

Note that, as shown earlier, this is equivalent to almost sure convergence of δ^t to zero. This being said, we are finally in a position to actually give a concise statement of B.A.T.

Blackwell's Approachability Theorem: Let C be any closed set in R^L. Suppose there exists a map $g : (R^L \backslash C) \rightarrow P[S_M]$ such that for every $z \in R^L \backslash C$, the measure $\mu = g(z)$ obeys

$$< v_M(\mu, s_N) - \pi_C(z), z - \pi_C(z) > \quad \leq \quad 0 \quad \forall s_N \in S_N. \tag{3.12}$$

Then C is approachable. Indeed, the behaviour rule $\gamma(h^t) := g(V_M(h^t))$ will serve.

In other words, if we find ourselves at round t, then the Blackwell condition requires that there exists a measure $\mu^{t+1} \in P[S_M]$ such that the expected value of the next payoff, $v_M(s^{t+1})$, is on the C-side of the hyperplane that separates this new payoff from the time-averaged payoff up to time t. The theorem then promises convergence of $V_M(h^t)$ to the set C.

Proof of B.A.T.:

The method of this proof is to show that, given Blackwell's condition, the random sequence δ^t defined by equation 3.11 satisfies the three inequalities of Lemma 3.2. It then follows that δ^t converges to zero almost surely, which is equivalent to C being approachable as shown earlier.

If we denote $\pi_C(V_M(h^t))$ by y^t, then Blackwell's condition requires that, given the history up to time t, we have:

$$< v_M(\mu^{t+1}, s_N) - y^t, V_M(h^t) - y^t > \quad \leq \quad 0 \qquad \forall t \in R \text{ and } s_N \in S_N. \tag{3.13}$$

Therefore, since $y^t \in C$ and using theorem 1,

$$\begin{aligned}
\delta^{t+1} &\leq |V_M(h^{t+1}) - y^t|^2 \\
&= |V_M(h^{t+1}) - V_M(h^t)|^2 + 2 < V_M(h^t) - y^t, V_M(h^{t+1}) - V_M(h^t) > \\
&\quad + |V_M(h^t) - y^t|^2 \qquad (3.14)
\end{aligned}$$

Now,

$$\begin{aligned}
V_M(h^{t+1}) - V_M(h^t) &= \frac{1}{t+1}\sum_{\tau=1}^{t+1} v_M(s^\tau) - \frac{1}{t}\sum_{\tau=1}^{t} v_M(s^\tau) \\
&= \frac{v_M(s^{t+1})}{t+1} + \frac{1}{t+1}\sum_{\tau=1}^{t} v_M(s^\tau) - \frac{1}{t}\sum_{\tau=1}^{t} v_M(s^\tau) \\
&= \frac{v_M(s^{t+1})}{t+1} + \frac{1}{t+1}\left[tV_M(h^t) - (t+1)V_M(h^t)\right] \\
&= \frac{v_M(s^{t+1}) - V_M(h^t)}{t+1}.
\end{aligned}$$

Using this fact, we can now write both

$$\begin{aligned}
< V_M(h^t) - y^t, V_M(h^{t+1}) - V_M(h^t) > &= \tfrac{1}{t+1}\left[< V_M(h^t) - y^t, v_M(s^{t+1}) - V_M(h^t) >\right] \\
= \tfrac{1}{t+1}\left[< V_M(h^t) - y^t, v_M(s^{t+1}) - y^t > + < V_M(h^t) - y^t, y^t - V_M(h^t) >\right] \\
= \tfrac{1}{t+1} < V_M(h^t) - y^t, v_M(s^{t+1}) - y^t > - \tfrac{1}{t+1}|y^t - V_M(h^t)|^2
\end{aligned}$$

and

$$|V_M(h^{t+1}) - V_M(h^t)|^2 = \left|\frac{v_M(s^{t+1}) - V_M(h^t)}{t+1}\right|^2 \leq \frac{c}{(t+1)^2}$$

since all payoffs are assumed to be bounded. Thus, given the history up to time $t-1$ and choosing $\mu^t \in P[S_M]$ such that Blackwell's condition (equation 3.13) is satisfied, equation 3.14 implies:

$$\begin{aligned}
E[\delta^t|\delta^1, ..., \delta^{t-1}] &= |V_M(h^{t-1}) - y^{t-1}|^2 + 2\frac{E[< V_M(h^{t-1}) - y^{t-1}, v_M(s^t) - y^{t-1} > |\delta^1, ..., \delta^{t-1}]}{t} \\
&\quad - 2\frac{|y^{t-1} - V_M(h^{t-1})|^2}{t} + E[|V_M(h^t) - V_M(h^{t-1})|^2|\delta^1, ..., \delta^{t-1}] \\
&\leq \delta^{t-1} - \frac{2}{t}\delta^{t-1} + \frac{c}{t^2} \\
&= \left(1 - \frac{2}{t}\right)\delta^{t-1} + \frac{c}{t^2} \quad \forall t \in R \text{ and } \forall s_N \in S_N \qquad (3.15)
\end{aligned}$$

Since all payoffs are bounded, we know that

$$0 \leq \delta^t \leq a \qquad \text{for some scalar } a \text{ and } \forall t \in R \text{ and } \forall s_N \in S_N. \tag{3.16}$$

Finally,

$$\begin{aligned}
\delta^t - \delta^{t-1} &\leq ||V_M(h^t) - \pi||^2 - ||V_M(h^{t-1} - \pi||^2 \quad \text{for all } h^t \\
&\leq ||V_M(h^t) - V_M(h^{t-1}||^2 \\
&= ||\frac{t-1}{t} V_M(h^t) + \frac{1}{t} v_M(s^t) - V_M(h^{t-1}||^2 \\
&\leq \frac{b}{t^2} \qquad \text{for some scalar } b \\
&\leq \frac{b}{t} \qquad \forall t \in R \text{ and } \forall s_N \in S_N
\end{aligned} \tag{3.17}$$

These last three numbered equations are equivalent to those required for the lemma. Thus B.A.T. follows.

Chapter 4
Convergence Results

What then can be said about the convergence properties of these various behaviour rules? Do they converge at all? If so, can we characterize the set to which they converge *independent of the game being played*? In this chapter, I will focus on the convergence of the empirical distribution of play, namely the sequence in $P[S_M \times S_N]$ defined by:

$$z^t(i,j) = \frac{1}{t}\sum_{\tau \leq t} I_i(S_M(\tau))I_j(S_N(\tau)).$$

4.1 Fictitious Play, Past Response and κ-Exponential Fictitious Play

Fictitious Play

Foster and Vohra have shown that, given a certain condition, the set of all distributions which are limit points of z^t is equal to CE provided that the behaviour rules involved are both a best-response to a calibrated forecast ([3]). Let us denote the set of limit points of this subset of behaviour rules by BR. Thus, Foster and Vohra proved that, given a certain condition, CE=BR. The required condition is that the set of measures over S_N for which i is a best-response, $M_b(i)$, must have a non-empty interior for all $i \in S_M$. More specifically, this implies that $M_b(i) \neq M_b(j)$ for all $i \neq j$ and $i, j \in S_M$. In those cases where this does not occur then there may exist a correlated equilibrium which is not contained in BR but any point in BR will still be a correlated equilibrium (i.e. BR

is contained in but not necessarily equal to CE). Fortunately, the above condition can be shown to be true for almost every game.

Since FP is a best-response to a calibrated forecast, it follows that if FP converges then it must converge to a correlated equilibrium. However, Foster and Vohra make no claim that any *specific* best-response behaviour rule based upon a calibrated forecast (such as FP) must converge. Their result merely implies that given any correlated equilibrium and given a game where the above condition is satisfied, there exists *some* best-response behaviour rule based on a calibrated forecast that will converge to it. The Shapley game described below is an example of a game in which FP does not converge to CE simply because there are no limit points. Thus the known convergence results for FP are entirely game specific.

Past Response

Past Response is a behaviour rule that I developed myself essentially as a potential opponent for the other rules. The idea behind it makes intuitive sense but I have made no attempt to determine any convergence results for it. Numerical results from three different games are given in the following chapter. These numerical results suggest that PR does converge to a correlated equilibrium when played against itself but not necessarily to a very desirable one (a defect that is common to most of the other behaviour rules as well).

κ-exponential Fictitious Play:

Fudenberg and Levine [4] prove that if all players use a universally consistent behaviour rule then the empirical distribution of play will eventually remain within MBR. Moreover, if all players use κ-exponential FP then the conclusion can be strengthened to insure that the empirical distribution of play will eventually remain within the *exact*

MBR.

To be a little more rigorous, assume that all players are using universally consistent behaviour rules. We can then derive a probability distribution, $p^T(\mu, \nu)$, over the set of possible empirical distributions arising from the use of the given behaviour rules up to any time T. Now, there is no a priori reason to assume that this probability distribution will converge at all as time progresses. However, since the space of measures on a compact set is compact (in the topology of weak convergence), we can at least be assured of the existence of accumulation points. What Fudenberg and Levine prove is that, with probability one, these accumulation points will be found entirely within MBR.

4.2 Regrets 2

Mas-Collel and Hart make stronger claims for both Regrets 1 and 2. Their Regrets 2 convergence result depends heavily on the aforementioned B.A.T., which allows Mas-Collel and Hart to establish the following theorem.

Theorem 2.3.1: Suppose that at every period t, M (or N) chooses strategies as outlined by Regrets 2. Then M's (or N's) regrets $R^t(j,k)$, converge to zero almost surely for every $j, k \in S_M$.

Proof:

Recalling the setup outlined in section 3.1, let $L = \{(j,k) \in S_M \times S_M\}$, so that R^L can be viewed as the space of all $|S_M| \times |S_N|$ matrices, and let C be the non-positive orthant of R^L. We define M's vector payoff, $v_M(s^t)$, as the following square matrix:

$$[v_M(s^t)](j,k) = \begin{cases} u_M(k, s_N(t)) - u_M(j, s_N(t)) & \text{if } s_M(t) = j, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

(Thus, $v_M(s^t)$ has only one non-zero row, namely row $j = s_M(t)$.)

So then, the regret function defined earlier is now equal to the positive part of the time-averaged pay-off. That is,

$$R_M^t(j,k) = \left[[V_M(h^t)](j,k)\right]^+.$$

Moreover, $V_M(h^t) \in C$ if and only if all of player M's regrets are zero.

Now, suppose $\mu \in P[S_M]$ is an invariant vector of the stochastic matrix P_M^t associated with R_M^t. This is the same vector used in Regrets 2, to determine player M's measure. As in Chapter 2, μ satisfies,

$$\mu^T R_M^t = \mu^T diag(R_M^t e). \tag{4.2}$$

Thus, to prove the theorem it would suffice to show that this μ satisfies Blackwell's condition for the set C since then, by B.A.T., $V_M(h^t)$ will approach C almost surely. (Hence $R_M^t \to 0$ almost surely.)

Note that for any $V_M(h^t) \in R^L \backslash C$,

$$V_M(h^t) - proj_C V_M(h^t) = V_M(h^t) - (V_M(h^t) - V_M(h^t)^+) = R_M^t.$$

Moreover,

$$proj_C V_M(h^t) \cdot (V_M(h^t) - proj_C V_M(h^t)) = V_M(h^t)^- \cdot V_M(h^t)^+ = 0.$$

Thus Blackwell's condition can be re-written as,

$$\sum_{j,k \in S_M} [v_M(\mu, s_N)](j,k) R_M^t(j,k) \leq 0 \quad \forall s_N \in S_N, \tag{4.3}$$

or, in matrix notation, where "$\cdot$" represents the standard matrix inner product ($A \cdot B = \text{tr}(A^T B)$),

$$v_M(\mu, s_N) \cdot R_M^t \leq 0.$$

Now since $[v_M(\mu, s_N)](j,k) = \mu(j)[u_M(k, s_N) - u_M(j, s_N)]$, equation 4.3 is equivalent to

$$\sum_{j,k \in S_M} \mu(j)[u_M(k, s_N) - u_M(j, s_N)] R_M^t(j,k) \leq 0$$

$$\Leftrightarrow \sum_{j \in S_M} \left[\sum_{k \in S_M} \mu(k) R_M^t(k,j) - \mu(j) \sum_{k \in S_M} R_M^t(j,k) \right] u_M(j, s_N) \leq 0$$

$$\Leftrightarrow \sum_{j \in S_M} \left[\mu^T R_M^t e_j - \mu^T (e_j^T R_M^t)^T \right] u_M(j, s_N) \leq 0$$

Let $\beta(j) = [\mu^T R_M^t e_j - \mu^T (e_j^T R_M^t)^T]$. So then $\beta = \mu^T R_M^t - \mu^T diag(R_M^t e)$. But this we know to be equal to zero by Equation 4.2 . Thus the inequality demanded here actually holds as an equality in this case. So Equation 4.3, and thus Blackwell's condition, holds. It follows, therefore, that $V_M(h^t)$ converges to the set C almost surely and therefore so does $R_M^t(j,k) = \max\{[V_M(h^t)](j,k), 0\}$, for all $j, k \in S_M$. That is, all regrets go to zero.

Mas-Collel and Hart go on to show that having the regrets go to zero *for all players* is a necessary and sufficient condition for convergence of the empirical distribution to CE. (Note that CE need not be a single point. Thus, the guarantee is not that Regrets 1 and 2 will ensure convergence of the empirical distribution to a single element of CE but that, given either of these behaviour rules, the empirical distribution will eventually remain within a small neighbourhood of the set of correlated equilibria.)

Proposition: Let $(s^t)_{t=1,2...}$ be a sequence of plays such that $limsup_{t \to \infty} R_i^t \leq 0$ for $i = M$ and N. Then the sequence of empirical distributions, z_t, converges to the set CE.

Proof: Let $\alpha \in P[S_M \times S_N]$ be an accumulation point of the sequence, z^t, of empirical distributions of play and consider arbitrary strategies $j, k \in S_M$. Then

$$[V_M(h^t)](j,k) = \sum_{s \in S_M \times S_N : s_M = j} z^t(j, s_N)[u_M(k, s_N) - u_M(j, s_N)].$$

In the limit, this is bounded above by zero (since all regrets go to zero). Therefore,

$$\limsup_{t\to\infty} \sum_{s\in S_M\times S_N : s_M=j} z^t(s)[u_M(k,s_N) - u_M(s)] \leq 0.$$

Thus taking the limit along the subsequence where z^t converges to α gives,

$$\begin{aligned} & \sum_{s\in S_M\times S_N : s_M=j} \alpha(s)[u_M(k,s_N) - u_M(s)] \;\leq\; 0 \\ \Leftrightarrow \sum_{s\in S_M\times S_N : s_M=j} & \alpha(j,s_N)[u_M(\psi(j),s_N) - u_M(j,s_N)] \;\leq\; 0 \qquad \text{where } \psi(j)=k \end{aligned}$$

Now, since this is true for any arbitrary pair $j,k \in S_M$, we get

$$\sum_{s\in S_M\times S_N} u_M(\psi(s_M),s_N)\alpha(s) - u_M(\alpha) \leq 0 \tag{4.4}$$

for all $\psi : S_M \to S_M$. Since the same could be shown using N instead of M, this proves that α is a correlated equilibrium.

4.3 Regrets 1

The proof that, if all players use Regrets 1, the empirical distribution of play converges to CE is a little more complicated. (Note that this hypothesis is also more restrictive than Regrets 2, which only required that all players use an adaptive procedure that insures that their regrets go to zero.) It does share similarities with the Regrets 2 approach in that the idea is to show that for each individual player, the regret goes to zero as the game progresses. Thus, by the above proposition, we have convergence to CE.

Theorem: If every player plays according to Regrets 1, then the empirical distribution of play z^t will converge almost surely as $t \to \infty$ to the set CE.

Proof: We will drop the subscript M whenever this will cause no confusion. Thus once again we define the set C to be the non-positive orthant of R^L, where $L = \{(j,k) \in$

$S_M \times S_M\}$. Recall that $V(h^t) = \frac{1}{t}\sum_{\tau \leq t} v(s^t)$. Regrets going to zero is then equivalent to $V(h^t)$ converging to the set C as t goes to infinity. A logical quantity to study therefore is the distance between $V(h^t)$ and the set C. We define:

$$\rho_t := [dist(V(h^t), C)]^2.$$

This proof will depend on taking a subsequence, $\{t_n\}_{n=0,1,2...}$, of the rounds of the game and showing that along this subsequence all regrets go to zero. That is, $\rho_{t_n} \to 0$, as $n \to \infty$. Then, by showing that all rounds in between t_n and t_{n+1} are bounded by the inverse of n, we will prove that all regrets go to zero. Thus, it is no use simply investigating the difference $\rho_{t+1} - \rho_t$. Instead, we must look at $\rho_{t+s} - \rho_t$ where $s = t_{n+1} - t_n$ and $t = t_n$. Specifically we need to look at the expected value of ρ_{t+s} where only the history up to time t is known. The proof will depend on the fact that we can keep s small relative to t.

This proof makes use of the standard "O" notation. For two real-valued functions, $f(\cdot)$ and $g(\cdot)$, defined on a domain X, "$f(x) = O(g(x))$" means that there exists a constant $K < \infty$ such that $|f(x)| \leq Kg(x)$ for all $x \in X$.

To simplify this process it is helpful to notice the following recursive relationship. Recalling the definition of $v(s^t)$ as given in Equation 4.1, it follows that:

$$V(h^{t+s}) = \frac{t}{t+s}V(h^t) + \frac{1}{t+s}\sum_{w=1}^{s} v(s^{t+w}) \tag{4.5}$$

Keeping all this in mind, we may now proceed with the proof of the convergence of the empirical distribution arising from the regrets 1 behaviour rule to the set of correlated equilibria (essentially following the proof developed by Mas-Collel and Hart, [8]).

Step 1:

With this background (equation 4.5) and since $[V(h^t)]^- \in C$, we can now derive a bound on ρ_{t+s}:

$$\begin{aligned}
\rho_{t+s} &\leq \left|\left|\left[\frac{t}{t+s}V(h^t) + \frac{1}{t+s}\sum_{w=1}^{s} v(s^{t+w})\right] - [V(h^t)]^-\right|\right|^2 \\
&= \left|\left|\frac{t}{t+s}\left(V(h^t) - [V(h^t)]^-\right) + \frac{1}{t+s}\sum_{w=1}^{s}\left(v(s^{t+w}) - [V(h^t)]^-\right)\right|\right|^2 \\
\Rightarrow (t+s)^2\rho_{t+s} &\leq t^2\rho_t + 2t\sum_{w=1}^{s}\left(v(s^{t+w}) - [V(h^t)]^-\right)\cdot\left(V(h^t) - [V(h^t)]^-\right) \\
&\quad + \left|\left|\sum_{w=1}^{s}\left(v(s^{t+w}) - [V(h^t)]^-\right)\right|\right|^2 \\
&\leq t^2\rho_t + 2t\sum_{w=1}^{s} v(s^{t+w})\cdot R^t + O(s^2) \quad \text{since } R^t = [V(h^t)]^+ \\
&\leq t^2\rho_t + O(ts + s^2)
\end{aligned}$$

This last inequality follows from the fact that all payoffs are bounded and all strategy sets are finite.

The above derivation leads to two useful equations:

$$E[(t+s)^2\rho_{t+s}|h^t] \leq t^2\rho_t + 2t\sum_{w=1}^{s} E[v(s^{t+w})|h^t]\cdot R^t + O(s^2), \tag{4.6}$$

$$(t+s)^2\rho_{t+s} - t^2\rho_t \leq O(ts + s^2). \tag{4.7}$$

Step 2:

Thus, we need to bound the expectation on the right hand side of equation 4.6, namely

$$E[v(s^{t+w})|h^t] = \sum_{s_N \in S_N} diag(\phi_{s_N})U_{s_N}$$

where $\phi_{s_N}(j) = Pr[s^{t+w} = (j, s_N)|h^t]$ and $U_{s_N}(j,k) = [u_M(k, s_N) - u_M(j, s_N)]^+$.

The difficulty is that the stochastic process $(s^{t+w})_{w=0,1,2,..}$ is not stationary so the probability that M plays strategy j in round $t+w$ is dependent on the play of N in rounds t through $t+w-1$. Thus, given only the history up to time t, the above expectation is very difficult to analyse. We proceed, therefore, by introducing a second stochastic process $(\hat{s}^{t+w})_{w=0,1,2...}$. This new stochastic process uses the stationary transition probability matrices P_M^t, P_N^t for all rounds greater than or equal to t. That is, $\hat{s}^\tau = s^\tau$ for all $\tau \leq t$ and

$$Pr[\hat{s}^{t+w} = (j, s_N)|\hat{s}^t, ..., \hat{s}^{t+w-1}] = P_M^t(\hat{s}_M(t+w-1), j)P_N^t(\hat{s}_N(t+w-1), s_N)$$

where P_M^t is the transition matrix defined by equation 2.2 on page 27. We can then define $\hat{\phi}$ for the process $(\hat{s}^{t+w})_{w=0,1,2...}$ in an analogous way to ϕ.

We will proceed by first showing that the difference between these two processes can be bounded for fixed w and then prove that if we use the stationary process instead of the original one then we would be able to get a useful bound for Equation 4.6. Thus we will have proven that the original process gives a similar bound.

Step 3:

In round $t+w$, the original process will use the matrix P^{t+w} while the new stationary process will use the matrix P^t. Thus, we would like to get a bound on the difference between these two transition matrices. Towards this end consider,

$$\begin{aligned} V(h^{t+s}) - V(h^t) &= \frac{t}{t+s}V(h^t) + \frac{1}{t+s}\sum_{w=1}^{s} v(s^{t+w}) - V(h^t) \\ &= \frac{1}{t+s}\sum_{w=1}^{s} v(s^{t+w}) - \frac{s}{t+s}V(h^t) \\ &\leq O(s/t). \end{aligned} \tag{4.8}$$

Therefore, it follows that the difference between the matrices P^t and P^{t+s} can be no more than $O(s/t)$ either.

Step 4:

However, we are really interested in comparing $\hat{\phi}$ and ϕ. Based on step 3, we get that

$$
\begin{aligned}
&Pr[\hat{s}^{t+w} = (j, s_N)|s^t, ..., s^{t+w-1}] - Pr[s^{t+w} = (j, s_N)|s^t, ..., s^{t+w-1}] \\
&= P_M^t(s_M(t+w-1), j)P_N^t(s_N(t+w-1), s_N) \\
&\qquad - P_M^{t+s}(s_M(t+w-1), j)P_N^{t+s}(s_N(t+w-1), s_N) \\
&= P_M^t(s_M(t+w-1), j)P_N^t(s_N(t+w-1), s_N) \\
&\qquad - (P_M^t(s_M(t+w-1), j) + O(w/t))(P_N^t(s_N(t+w-1), s_N) + O(w/t)) \\
&= O(w/t) \qquad \text{since } w < t
\end{aligned}
$$

However, this only bounds the one-step difference where the history up to time $t + w - 1$ is known and is equivalent in the two processes, whereas to compare $\hat{\phi}$ and ϕ we need to bound the s-step difference where only the history up to time t is known and equivalent. To do this we proceed by induction. Consider the following proposition involving n:

$$
Pr[\hat{s}^{t+w} = (j, s_N)|s^t, ..., s^{t+n}] - Pr[s^{t+w} = (j, s_N)|s^t, ..., s^{t+n}] \leq \sum_{r=n+1}^{w} O(r/t) \qquad (4.9)
$$

We know this is true by the above, for $n = w - 1$. We need to show that this is true for $n = 0$. But this is proven by the following

$$
\begin{aligned}
&Pr[\hat{s}^{t+w} = (j, s_N)|s^t, ..., s^{t+n-1}] \\
&\quad = \textstyle\sum_{s^{t+n}} Pr[\hat{s}^{t+w} = (j, s_N)|s^t, ..., s^{t+n}]Pr[\hat{s}^{t+n} = s^{t+n}|s^t, ..., s^{t+n-1}] \\
&\quad \leq \textstyle\sum_{s^{t+n}}(Pr[s^{t+w} = (j, s_N)|s^t, ..., s^{t+n}] + \sum_{r=n+1}^{w} O(r/t))Pr[\hat{s}^{t+n} = s^{t+n}|s^t, ..., s^{t+n-1}] \\
&\quad \leq \textstyle\sum_{s^{t+n}} Pr[s^{t+w} = (j, s_N)|s^t, ..., s^{t+n}](Pr[s^{t+n} = s^{t+n}|s^t, ..., s^{t+n-1}] + O(n/t)) \\
&\qquad + \textstyle\sum_{r=n+1}^{w} O(r/t) \\
&\quad \leq \textstyle\sum_{s^{t+n}} Pr[s^{t+w} = (j, s_N)|s^t, ..., s^{t+n}]Pr[s^{t+n} = s^{t+n}|s^t, ..., s^{t+n-1}] + \sum_{r=n}^{w} O(r/t) \\
&\quad = Pr[s^{t+w} = (j, s_N)|s^t, ..., s^{t+n-1}] + \textstyle\sum_{r=n}^{w} O(r/t)
\end{aligned}
$$

Thus, by induction, this proves that equation 4.9 is true for $n = 0$. Therefore,

$$\begin{aligned}
\hat{\phi}_{s_N}(j) - \phi_{s_N}(j) &\leq \sum_{r=0}^{w} O(r/t) \\
\Rightarrow \hat{\phi}_{s_N}(j) - \phi_{s_N}(j) &\leq \frac{K}{t} \sum_{r=0}^{w} r \quad \text{for some large } K \\
&\leq \frac{K}{t}\left(\frac{w(w+1)}{2}\right) \\
&\leq \frac{K}{t}\left(\frac{w(w+w)}{2}\right) \\
&= K\left(\frac{w^2}{t}\right).
\end{aligned}$$

It is important to note that this step requires that player N also does not change strategy too quickly. That is, player N must use a behaviour rule which insures that the transition probabilities do not change too drastically between round t and round $t + w$. The statement of the theorem is that all players use the same Regrets 1 behaviour rule. This is somewhat more restrictive than necessary but it does, at least, insure the validity of this proof.

Step 5:

So then

$$\begin{aligned}
&E[v(\hat{s}^{t+w})|h^t] \cdot R^t - E[v(s^{t+w})|h^t] \cdot R^t \\
&\quad = \left[\textstyle\sum_{s_N \in S_N} diag(\hat{\phi}_{s_N}) U_{s_N} - \sum_{s_N \in S_N} diag(\phi_{s_N}) U_{s_N}\right] \cdot R^t \\
&\quad = \left[\textstyle\sum_{s_N \in S_N} diag(\hat{\phi}_{s_N} - \phi_{s_N}) U_{s_N}\right] \cdot R^t \\
&\quad \leq 2 \max(u_M) \left[\textstyle\sum_{s_N \in S_N} diag(\hat{\phi}_{s_N} - \phi_{s_N}) U_{s_N}\right] \\
&\quad \leq 4 \max(u_M)^2 \left[\textstyle\sum_{j \in S_M} \sum_{s_N} \hat{\phi}_{s_N}(j) - \phi_{s_N}(j)\right] \\
&\quad = O(w^2/t) \qquad \text{by Step 4}
\end{aligned}$$

It remains to be shown, therefore, that $E[v(\hat{s}^{t+w})|h^t] \cdot R^t$ can be bounded for all w.

Step 6:

$$\hat{\phi}_{s_N}(j) = Pr[\hat{s}_N(t+w) = s_N|h^t]Pr[\hat{s}_M(t+w) = j|h^t] \quad (4.10)$$

$$= Pr[\hat{s}_N(t+w) = s_N|h^t][P^t]^w(s_M(t), j) \quad (4.11)$$

Therefore, since $P^t = I + \frac{1}{\kappa}(R^t - diag(R^t e))$

$$E[v(\hat{s}^{t+w})|h^t] \cdot R^t = \kappa \left(\sum_{s_N \in S_N} diag(\hat{\phi}_{s_N}) U_{s_N} \right) \cdot \left(P^t - I + \frac{1}{\kappa} diag(R^t e) \right)$$

$$= \kappa \sum_{s_N \in S_N} \left\{ \sum_{j,k \in S_M} \hat{\phi}_{s_N}(j)[u_M(k, s_N) - u_M(j, s_N)]P^t(j,k) \right\} \quad \text{(by defn of } U_{s_N})$$

$$= \kappa \sum_{s_N \in S_N} \left[\sum_{j,k \in S_M} \hat{\phi}_{s_N}(j) u_M(k, s_N) P^t(j,k) - \sum_{j,k \in S_M} \hat{\phi}_{s_N}(j) u_M(j, s_N) P^t(j,k) \right]$$

(Note that the second equality follows since $I - \frac{1}{\kappa} diag(R^t e)$ is a diagonal matrix and U_{S_N} has only zero elements along the diagonal.)

Now, if we switch the indices in the first sum over $j, k \in S_M$ and recognize that the second sum reduces to $\sum_{j \in S_M} \hat{\phi}_{s_N}(j) u_M(j, s_N)$ (since $\sum_{k \in S_M} P^t(j,k) = 1$ for all $j \in S_M$), we get that

$$E[v(\hat{s}^{t+w})|h^t] \cdot R^t = \kappa \sum_{s_N \in S_N} \sum_{j \in S_M} u_M(j, s_N) \left[\sum_{k \in S_M} \hat{\phi}_{s_N}(k) P^t(k,j) - \hat{\phi}_{s_N}(j) \right] \quad (4.12)$$

Thus, setting $\beta^{t,w}(j, s_N)$ equal to what is inside the square brackets and recalling equation 4.11, we have

$$\beta^{t,w}(j, s_N) = Pr[\hat{s}_N(t+w) = s_N|h^t] \left\{ \sum_{k \in S_M} [P^t]^w(s_M(t), k) P^t(k,j) - [P^t]^w(s_M(t), j) \right\}$$

$$= Pr[\hat{s}_N(t+w) = s_N|h^t] \left\{ [P^t]^{w+1}(s_M(t), j) - [P^t]^w(s_M(t), j) \right\}$$

$$= Pr[\hat{s}_N(t+w) = s_N|h^t][(P^t)^{w+1} - (P^t)^w](s_M(t), j)$$

Therefore,if we can bound $\beta^{t,w}$ then we will have bounded $E[v(\hat{s}^{t+w})|h^t] \cdot R^t$ as desired.

Step 7:

Step 7 depends on the following lemma:

Let P be an $m \times m$ stochastic matrix with all of its diagonal entries positive. Then $[P^{w+1} - P^w](j,k) = O(w^{-1/2})$ for all $j,k = 1,..,m$.

Given this lemma, it is clear that $\beta^{t,w} = O(w^{-1/2})$ since P^t is designed so that the diagonal entries are positive.

Step 8:

Thus, we have now shown that

$$E[v(\hat{s}^{t+w})|h^t] \cdot R^t = O(w^{-1/2})$$

Therefore, by step 5,

$$E[v(s^{t+w})|h^t] \cdot R^t = O(\frac{w^2}{t} + w^{-1/2})$$

Finally, returning to equation 4.6 from step 1, it follows that

$$\begin{aligned} E[(t+s)^2\rho_{t+s}|h^t] &\leq t^2\rho_t + 2t\sum_{w=1}^{s} O(\frac{w^2}{t} + w^{-1/2}) + O(s^2) \\ &= t^2\rho_t + O(s^3 + ts^{1/2}) \end{aligned}$$

Step 9:

Here then is where we need to make an intelligent choice for a subsequence, $\{t_n\}$. Let t_n be equal to the largest integer not exceeding $n^{5/3}$. Therefore, by letting $t = t_n$ and $s = t_{n+1} - t_n$, step 8 results in

$$E[t_{n+1}^2\rho_{t_{n+1}}|h^{t_n}] = t_n^2\rho_{t_n} + O(n^2)$$

since $s = O(n^{2/3})$ which implies that $s^3 = O(n^2)$ and $ts^{1/2} = O(n^2)$.

Step 10:

Finally, we can use the following theorem by Loeve, to show that along this subsequence ρ_{t_n} goes to zero as $n \to \infty$.

Theorem: Let X_n be a sequence of random variables and b_n a sequence of real numbers increasing to infinity, such that the series $\sum_n Var(X_n)/b_n^2$ converges. Then

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{r=1}^{n} (X_r - E[X_r|X_1, ..., X_{r-1}]) = 0 \qquad \text{a.s.}$$

If we let $b_n = t_n^2$ and $X_n = b_n\rho_{t_n} - b_{n-1}\rho_{t_{n-1}}$ then by equation 4.7 from step 1, it follows that $|X_n| \leq O(t_n s_n + s_n^2) = O(n^{7/3})$. Therefore,

$$\sum_n Var(X_n)/b_n^2 = \sum_n O(n^{14/3})/n^{20/3} = \sum_n O(1/n^2) < \infty$$

Moreover, Step 9 implies that,

$$\frac{1}{b_n} \sum_{r\leq n} E[X_r|X_1, ..., X_{r-1}] = O(n^{-10/3}) \sum_{r\leq n} O(r^2) = O(n^{-10/3}n^3) = O(n^{-1/3})$$

Therefore, since this obviously goes to zero as $n \to \infty$, it follows that

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{r=1}^{n} X_r = \lim_{n\to\infty} \left[\rho_{t_n} - (\frac{t_0}{t_n})^2 \rho_{t_0}\right] = \lim_{n\to\infty} \rho_{t_n} = 0 \qquad \text{a.s.}$$

Step 11:

Since $\rho_{t_n} = \sum_{j,k}[R^{t_n}(j,k)]^2$, step 10 implies that the subsequence R^{t_n} converges to the set C a.s. as $n \to \infty$. When $t_n \leq t \leq t_{n+1}$ we have $R^t - R^{t_n} \leq O(n^{2/3})/O(n^{5/3}) = O(1/n)$ by equation 4.8 in step 3. Thus the full sequence R^t converges to the set C a.s., thereby completing the proof.

Alternative Proof for Regrets 1 Convergence

The concept behind this alternative proof is to use what we already know concerning the convergence properties of the Regrets 2 behaviour rule. If we can show that eventually the Regrets 1 behaviour rule behaves similarly to that of Regrets 2 then we will have proven the desired result. The basic idea is to show that a subsequence of the transition matrices arising from Regrets 1 is essentially equivalent to the transition matrices arising from Regrets 2.

Towards that end it is necessary to introduce a tremble. Recall that a tremble is introduced to insure that all transitions have positive probability. In this case, I will introduce a tremble of $1/t^{1/4}$. That is, if S_1^τ is the original sequence of transition matrices arising from the Regrets 1 behaviour rule, then the new sequence of transition matrices has the following form:

$$S_2^\tau = \frac{1}{\tau^{1/4}} A + \left(1 - \frac{|S_M||S_N|}{\tau^{1/4}}\right) S_1^\tau \quad \text{for all } \tau.$$

where A is a matrix with each entry equal to one and τ is large enough so that the coefficient in front of S_1^τ is positive. Note that the difference between the two sequences must always be of the order of $1/\tau^{1/4}$ so that, for large enough τ, dealing with one sequence is essentially the same as dealing with the other. Therefore,

$$E[S_1^{t+s} - S_2^{t+s}|h^t] = O(1/(t+s)^{1/4}).$$

Now, recall that $R_M^t = [V_M(h^t)]^+$ and that $V_M(h^t)$ obeys the following recursion formula:

$$V_M(h^{t+s}) = \frac{t}{t+s} V_M(h^t) + \frac{1}{t+s} \sum_{w=1}^{s} v_M(s^{t+w}).$$

This, in turn led us to the fact that

$$P_M^{t+s} - P_M^t = O(s/t). \tag{4.13}$$

(see step 3 from the original proof for regrets 1.)

Moreover, as t gets large, the above recursion formula makes it clear that the Regrets 1 behaviour rule essentially uses the same transition matrix over and over again. In other words,

$$E[P_M^{t+s}|h^t] \approx [P_M^t]^s.$$

In fact, the difference is of the order of s/t.

But, by theorem 2.15 in Markov Processes for Stochastic Modeling [7], the difference between $[P_M^t]^s$ and the stationary distribution associated with P_M^t can be bounded in the following way:

$$[P_M^t]^s - \pi^t \leq C \left[\left(1 - \frac{1}{t^{1/4}} \right)^{\frac{2}{|S_M|(|S_M|+1)}} \right]^s \tag{4.14}$$

where C is a constant and π^t is the stationary distribution associated with the transition matrix P_M^t. Note that this is where it is necessary to assume that the transition matrices have non-zero entries. The above discussion on the addition of a tremble however shows that the perturbation introduced by the tremble can be easily controlled.

Therefore, if we are given an $\epsilon > 0$ and a $\delta > 0$, and we start with t large enough, then we can choose s large enough so that the bound in equation 4.14 is less that δ without having the bound in equation 4.13 exceed ϵ. The only difficulty would occur if as t increased the ratio s/t required in order to reduce the bound from equation 4.14 to a fixed number increased. However, it is easy to show computationally that this is not the case and that in fact this ratio decreases as t gets larger.

Thus we can take a subsequence of the transition matrices arising from the Regrets 1 behaviour rule – $P_M^{t_1+s_1}, P_M^{t_2+s_2}, \ldots$ (where $t_{i+1} = t_i + s_i$) so that

$$P_M^{t_{i+1}+s_{i+1}} - P_M^{t_i+s_i} \leq \epsilon \tag{4.15}$$

and

$$\begin{aligned} E[P_M^{t_{i+1}+s_{i+1}}|h^t] - \pi^{t_i+s_i} &\leq [P_M^{t_i+s_i}]^{s_{i+1}} - \pi^{t_i+s_i} + \epsilon \\ &\leq \delta + \epsilon \end{aligned} \tag{4.16}$$

Equation 4.16 proves that this subsequence of Regrets 1 essentially behaves like that of Regrets 2. Now, we know that Regrets 2 forces the regret matrix to zero and therefore forces the transition matrix to the identity. Therefore along this subsequence, Regrets 1 also forces the regrets to zero. Moreover, the transition matrices (and therefore the regrets) between the members of the subsequence are restricted by equation 4.15 to be within ϵ of the closest member of the above subsequence. Therefore, Regrets 1 must also force the regrets to zero and so insure convergence to CE.

4.4 Modified Regrets

The convergence results for modified regrets (MR) are entirely numerical. In the zero-sum game discussed in the next chapter, MR does not converge to a correlated equilibrium. However, it does converge towards the "best" compromise position where both players get an equal payoff. In the other two games, neither of which are zero sum, MR converges at a very fast rate to a correlated equilibrium. More particularly, it converges to the correlated equilibrium that results in equal payoff to both players. None of the other rules show anything close to the same kind of convergence rate – at least not in all three games. (These results will be given in more detail in the next chapter.) Intuitively, this makes sense. MR is based on the idea of compromise. It is an attempt to incorporate into a behaviour rule the idea of accomodation. Thus, it would make sense that, in games where there is a possibility of compromise (i.e., in non-zero sum games), MR would work very well. Even in a zero-sum game, the idea of a compromise would simply mean that both players should receive approximately zero payoff.

This brings into question, in my mind, the whole idea of convergence to correlated equilibria in the first place. As will be demonstrated, in the Battle of the Buddies game all behaviour rules converge to a correlated equilibrium but it is very obvious that the result (except in the case of MR) is far from satisfactory for one of the two players (see results next chapter). Thus convergence to the set of correlated equilibria is not necessarily a desirable outcome; or at least not a sufficiently restrictive one. If instead, we could guarantee convergence to the "best" compromise then surely that would be a more useful result. How one defines the "best" compromise is not an easy question to answer. Certainly, in a zero-sum game the best compromise would result in a zero payoff for each player. For a nonzero-sum game though the "best" compromise may not be definable in the general setting but rather depend on the specifics of each individual game.

There is one type of game that would most obviously cause the MR behaviour rule some difficulty. Consider a game with the following pay-off matrices:

$$u_M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad u_N = \begin{bmatrix} 0 & 3 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

In such a game, MR will lead player N to have a positive regret for not playing strategy 2. But if N plays strategy 2 then player M will play strategy 2 as well. Player N may continue to play strategy 2 (despite receiving zero pay-off) in the hopes of forcing player M to switch but in this case there is no reason why M ever would. Thus MR will only work well in those games in which both players can act in such a way as to create an adverse effect on the other player which might then influence him to switch strategy (such as Battle of the Buddies).

Without knowing the pay-off matrix of one's opponent, it is not easy to see how this might be remedied. One possible option would be to introduce a weighting on the two

sums in equation 2.7 on page 30. Such a weighting might consider monitoring the average of the components of u_M and increasing the weight on the first sum (the normal regret) if the average of past payoffs dipped below this level. Such a weighting might then prevent player M from continuing to try to influence player N when previous play had shown such influence to be minimal. My own experience, however, with the introduction of a weighting system has made it very clear that this is a delicate task. A weighting that may work well for one game can prove disastrous in another. Whether there does exist a "optimal" weighting system that would allow MR to perform well in all types of games remains an open question.

Chapter 5
Experimental Results

This chapter describes numerical experiments with the convergence results of various behaviour rules. I chose three different games to play, picking three whose forms varied quite dramatically. The first one is a zero-sum game which is an expanded version of the familiar scissors-paper-rock game. The second is a game developed by Lloyd Shapley and the third is the aforementioned Battle of the Buddies. I will use the following abbreviations to refer to the behaviour rules defined above: Regrets 1 - R1, Regrets 2 - R2, Modified Regrets - MR, Fictitious Play - FP and κ exponential-fictitious play - EFP.

5.1 Scissors-Paper-Rock-Glass-Water

This game (abbreviated SPRGW) works exactly the same way as the old game of scissors-paper-rock, except that now, instead of having three choices, each player has five ([9]). Which choice beats which is demonstrated in the following diagram. If the arrow goes from vertex u to vertex v then u beats out v. Thus, for instance, glass beats out stone (see diagram below).

A loss means a penalty of -1 and a win means a reward of $+1$. Thus $u_N = -u_M$,

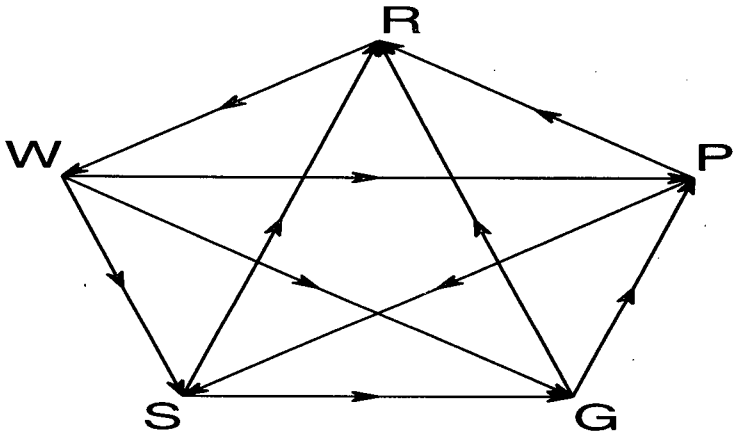


Figure 5.1: SPRGW – rules of the game

where u_M is of the following form:

$$u_M = \begin{bmatrix} 0 & -1 & 1 & 1 & -1 \\ 1 & 0 & 1 & -1 & -1 \\ -1 & -1 & 0 & -1 & 1 \\ -1 & 1 & 1 & 0 & -1 \\ 1 & 1 & -1 & 1 & 0 \end{bmatrix}$$

In this game, the convergence rate for FP and PR is extremely fast. FP and PR when played against themselves converge to the Nash equilibrium where the marginals are both $(1/9, 1/9, 1/3, 1/9, 1/3)$. The resulting payoff for both players is, of course, zero. EFP converges to the same Nash equilibrium, though at a slower rate. R1 seems to be doing the same though the convergence rate is extremely slow. Finally, MR played against itself leads to a fast convergence to the marginals (1/3,1/3,0,1/3,0) which results in zero payoff as well. However, this distribution is *not* an equilibrium. So, in each case, an "optimal" compromise is reached in which both players do equally well. When the different behaviour rules are played against each other, however, their convergence properties become much less obvious (see Appendix). Numerical results however did support Fudenberg and Levine's claim that EFP converges to MBR in all cases where

the opponent uses a consistent behaviour rule.

5.2 The Shapley Game

Julia Robinson proved, as early as 1951, that, in a two person zero-sum game, FP does insure the convergence of the empirical distribution to NE ([10]). The Shapley game was originally conceived by Lloyd Shapley in order to prove that FP does not necessarily converge to a NE outside of the narrow context assumed by Robinson ([11]). Shapley's game is also useful here because it turns out to be a good example for demonstrating the previously discussed convergence properties (or lack thereof) of the various behaviour rules.

The game is a nonzero-sum 3×3 game with the following payoff matrices:

$$u_M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad u_N = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Foster and Vohra [3] have proven that, not only does FP not converge to an equilibrium but that, in fact, it doesn't converge at all. A fairly simple program will show that, in this game, FP oscillates between 6 states: $(1,1), (1,2), (2,2), (2,3), (3,3)$ and $(3,1)$. Now the only correlated equilibrium with support on these states places equal weight, $1/6$, on each of them. If we order the coordinates of the matrix by labelling the first row across numbers 1 through 3 and then the second 4 through 6 and so on, FP produces the following empirical distribution of play:

After 1000 iterations:

$(.120, .281, 0, 0, .258, .058, .128, 0, .086)$

After 30000 iterations:

$(.064, .095, 0, 0, .139, .203, .201, 0, .298)$

These values and the graphical results in Appendix A provide no indication that FP is converging to the correlated equilibrium. In fact, FP cycles through these six states and with each state change the length of time to the next change becomes longer. Note, however, that according to Fudenberg and Levine, due to the increase in time between state changes, FP is consistent in this game and thus should remain within MBR (the table below provides numerical confirmation of this). Thus this shows that Fudenberg and Levine's result cannot be strengthened to insure convergence to CE.

The κ-exponential FP behaviour rule also cycles through the same six states and produces the following empirical distribution of play after a certain number of iterations, again showing that universal consistency is only enough to insure convergence to MBR:

After 1000 iterations:

$(.223, .329, 0, 0, .132, .066, .150, 0, .101)$

After 30000 iterations:

$(.243, .061, 0, 0, .089, .131, .283, 0, .193)$

EFP seems to do much the same as FP in that it cycles through the same six states with the rate of state change getting longer and longer.

The Regret 1 behaviour rule again cycles through the same six states. The convergence to CE is obviously extremely slow.

After 1000 iterations:

$(.218, .347, .000, .000, .237, .038, .093, .000, .068)$

After 30000 iterations:

$(.136, .149, .006, .015, .182, .197, .114, .004, .196)$

PR does something a little different. This behaviour rule cycles between all nine states and does seem to converge (though the evidence is entirely numerical) to the

correlated equilibrium which places the same weight on each of the six states mentioned above, and a little less weight on the remaining three states.

After 1000 iterations:

$(.131, .117, .103, .097, .107, .112, .123, .092, .119)$

After 30000 iterations:

$(.114, .115, .108, .107, .112, .110, .113, .109, .112)$

This, once again, highlights the fact that not all correlated equilibria are desirable results. Here, the payoff is close to $1/3$ for each player which is far from the optimal compromise of $1/2$ each.

Finally, MR not only converges to CE but does so at a rate much faster than that of Regrets 1. It also places all its weight on the same six states.

After 1000 iterations:

$(.1668, .1668, 0, 0, .1668, .1668, .1668, 0, .1665)$

After 30000 iterations:

$(.1665, .1666, 0, 0, .1665, .1665, .1671, 0, .1665)$

The following table summarizes the payoffs between behaviour rules after 30000 iterations. It also gives the defining quantities, $\max_\mu(\mu, \alpha_N)$ and $\max_\nu(\alpha_M, \nu)$, for MBR so that the results of Fudenberg and Levine can be easily confirmed.

Behaviour Rule (M vs N)	$\max_\mu u_M(\mu, \alpha_N)$	$u_M(\alpha)$	$\max_\nu u_M(\alpha_M, \nu)$	$u_N(\alpha)$
PR vs PR	.336	.337	.337	.338
FP vs PR	.403	.402	.481	.593
FP vs FP	.501	.501	.499	.499
EFP vs PR	.418	.418	.469	.578
EFP vs FP	.413	.413	.476	.583
EFP vs EFP	.526	.525	.475	.475
R1 vs PR	.359	.451	.356	.470
R1 vs FP	.453	.537	.463	.463
R1 vs EFP	.447	.485	.462	.462
R1 vs R1	.399	.514	.395	.461
MR vs PR	.333	.250	.333	.500
MR vs FP	.333	.800	.333	.200
MR vs EFP	.433	.525	.294	.392
MR vs R1	.335	.850	.345	.101
MR vs MR	.334	.500	.334	.500

Pathological as the Shapley game seems to be, it does, however ratify Fundenberg and Levine's result that any rule which is universally-consistent will eventually converge to the exact MBR (at least to a good approximation). Results show clearly that EFP converges to the exact MBR, no matter what behaviour rule is used by the opposing player (except MR). FP converges to the exact MBR except when played against MR or EFP. PR, on the other hand, converges to MBR only when played against itself. Recall that Fudenberg and Levine's result only insures convergence to MBR when all players are using a consistent behaviour rule. R1 has been touted by Mas-Collel and Hart to

converge to CE in any game. While there is indication that this is true (see Appendix A) and theoretically the proof is sound, the rate of convergence is clearly slower than one might hope. Even after 30000 iterations, R1 versus itself has yet to satisfy the criterion for MBR, let alone that of CE. Excluding MR, most of the above behaviour rule pairs do not converge to any distribution. The exceptions are PR against itself and R1 against itself. Thus, the actual figures given above for payoffs received are not all that useful (except as ratification for Fudenberg and Levine's result). Closer examination actually shows that the payoffs continue to vary a fair bit, favouring player M at one iteration and player N at another.

In contrast, MR causes rapid convergence when played against itself *as well as against FP, R1 and PR.* MR vastly outplays both FP and Regrets 1 (see results in the above table) while being itself outplayed by PR. This last result is due to the fact that PR will inevitably do well against a behaviour rule that uses a best response technique. PR when played against MR results in frequent switches causing MR to resort to the best-response option on a frequent basis. Against EFP, the empirical distribution oscillates in such a way that both players do about equally well. Against itself, MR converges rapidly to the correlated equilibrium that places equal weight (1/6) on the six states (1,1),(1,2),(2,2),(2,3),(3,1) and (3,3) which, of course, leads to both players receiving a payoff of 1/2.

5.3 The Battle of the Buddies

Recall that the Battle of the Buddies game is defined by the following matrices:

$$u_M = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}, \qquad u_N = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}.$$

For this game, the NE contains three strategy pairs – $[(0,1),(0,1)]$,$[(1,0),(1,0)]$ and $[(5/6,1/6),(1/6,5/6)]$ as shown earlier (chapter 1). Recall that, respectively, they result in payoffs of $(1,5),(5,1)$ and $(5/6,5/6)$. The first two equilibria result when one of the two friends caves in and agrees always to accede to his friend's choice. The last is more of a compromise but results in both of them getting even less of a reward than if either one had caved.

Excluding MR, the results from this game are rather uninteresting. For any combination of behaviour rules, the result is always that one of the two players caves in to the choice of the other. Which player caves in is entirely conditional on the first few rounds of play. In any case, the result, even though it is a correlated equilibrium, is anything but a compromise with one player always receiving a payoff of 5 and the other a payoff of 1. Only MR was able to do differently. Playing this behaviour rule against any of the others (except PR) leads to *the opponent* caving in every time – *regardless of the initial condition.* When played against PR, MR does not do well as both players tend to be too stubborn (see the table on the next page). When played against itself, MR comes close to the correlated equilibrium which places equal weight on strategy pairs $(1,1)$ and $(2,2)$ and none anywhere else. (After only 3000 iterations, the empirical distribution was $(.4389,.1256,0,.4355)$, which is itself a correlated equilibrium.) This leads to a much more satisfactory compromise where both players receive a payoff of approximately three.

After only 3000 iterations the following payoffs and best responses to the marginals were observed:

Behaviour Rule	$\max_\mu u_M(\mu, \alpha_N)$	$u_M(\alpha)$	$\max_\nu u_M(\alpha_M, \nu)$	$u_N(\alpha)$
MR vs FP	4.97	4.97	.999	.998
MR vs EFP	4.96	4.96	.998	.999
MR vs R1	4.94	4.94	.999	.994
MR vs PR	.985	.203	.866	.837
MR vs MR	2.19	2.63	2.18	2.62

Note that the initial conditions were set so that player N had the initial advantage. If they are set so that player M has the initial advantage then all of the first four converge directly to the Nash equilibrium that places all its weight on strategy pair $(1, 1)$. Thus, MR will not only outplay all but PR in any game of the same form as Battle of the Buddies but will, if played against itself, result in the "optimal" compromise.

Chapter 6
Conclusion

The problems related to the convergence of behaviour rules within game theory can be broken into three areas – whether or not a given behaviour rule can insure convergence of the empirical distribution, if it does, at what rate and, finally, to what point (or set of points) does it force convergence?

The three games described in the last chapter show clearly that it is no simple matter to insure convergence of the empirical distribution even to a set, let alone a point. One solution is to relax the goal. Thus, Fudenberg and Levine developed κ-exponential FP which concentrates on insuring convergence to MBR. As the Battle of the Buddies game shows, however, this may lead to a less than satisfactory result, depending on the nature of the game. One can very easily play a best-response to the marginal of the other player and still fail to maximize one's payoff or even reach an acceptable compromise. Fudenberg and Levine's result for κ-exponential fictitious play is especially nice, however, in that it guarantees convergence to MBR *no matter what behaviour rule is used by the opponent.*

Normal regrets, as developed by Mas-Collel and Hart, has the benefit of insuring convergence of the empirical distribution to a smaller set (CE) no matter what the game but again this does not address the fact that the set CE may contain points whose status as a compromise are less than ideal (e.g. Battle of the Buddies). Moreover, it is very clear

that though convergence is assured, the rate of convergence leaves much to be desired. Still there is something to be said for being able to insure convergence to CE as then, at least, one's payoff is assured. Unfortunately, Mas-Collel and Hart's result only insures convergence to CE if both players use normal regrets. Our numerical experiments showed that in the Shapley game, regrets 1 fails to converge to CE when played against anything but itself.

What would be the ideal situation? Clearly one would want to insure a fast convergence, but to what? I am not convinced that either MBR or CE are necessarily the best answers to that question; or at least not sufficiently restrictive answers. In many games there are clearly correlated equilibria that do not lead to a satisfactory result for both players. An alternative would be the set of "optimal" compromises. Obviously, in zero-sum games this set would only include those distributions which lead to both players receiving a payoff of zero. However, in a nonzero-sum game, though the concept of an "optimal" compromise is fairly intuitive, it is not so clear what the set of "optimal" compromises would entail. Clearly it is not simply those for which both players do equally well as this would allow both players to do equally badly. For instance, in the Shapley game there are numerous distributions that would lead to both players getting zero. The set would include some correlated equilibria, but not all, and would also include some distributions outside of CE. It would hopefully avoid such unsatisfactory results as occurred in the Battle of the Buddies.

Even more ideally (and perhaps unrealistically) one would like to be able to insure convergence to the above set *even against other behaviour rules.* In fact, one would expand the set so that the goal is for the empirical distribution to converge to the set that insures *at least* the "optimal" compromise. This may, however, be too ambitious a task.

MR was an attempt to provide a behaviour rule that forced convergence to a more satisfactory set. While the results when MR is played against itself are generally excellent (and definitely superior to the other behaviour rules studied), there are games (as described on page 62) where currently MR fails to converge to a distribution resulting in anything that might reasonably be called an "optimal" compromise. Moreover, MR failed to do well against PR in any of the games discussed in the previous chapter. Whether or not there exists a useful weighting of the two sums involved that would lead to more satisfactory convergence, *no matter what the game*, is an open question.

In view of our results above, it seems fairly clear that the three problems related to the convergence of behaviour rules in game theory (mentioned at the outset of this conclusion) have yet to be ideally answered. Certainly there is possible work to be done in determining a more satisfactory target set to replace MBR and CE. Even with CE as the target set, it would be beneficial to have a behaviour rule whose convergence rate was a little faster than that of Regrets 1; though as yet Mas-Collel and Hart, to their credit, are the only ones to guarantee convergence to CE at all. The above notwithstanding, there is much to be said for the work that has already been done. My hope is that this thesis has provided a clear understanding of the research so far and at least some indication of future potentialities.

Appendix A: Graphical and Tabular Results

This appendix presents graphical and tabular evidence of convergence of the different behaviour rule pairs in the games described in the chapter 5. I have not included the graphs of all possible pairs in all three games as that would be both tedious and unnecessary. I have included some examples from the Shapley game as an indication of the various convergence rates. Also included are tables of the payoffs and empirical distributions after 30000 iterations, along with an indication of each pair's convergence properties.

Legend:
* obvious convergence
+ evidence of convergence
– no evidence of convergence

Battle of the Buddies - 3000 iterations

(initial conditions set to favour player N)

Convergence	Behaviour Rule	Empirical Distribution
*	R1 vs FP	$(0, 0, 0, 1)$
*	R1 vs EFP	$(0, 0, 0, 1)$
*	R1 vs PR	$(0, 0, 0, 1)$
*	R1 vs R1	$(0, 0, 0, 1)$
*	MR vs R1	$(.9887, .0103, .0000, .0010)$
*	MR vs FP	$(.9943, .0050, .0000, .0007)$
*	MR vs EFP	$(.9917, .0070, .0000, .0013)$
*	MR vs PR	$(.0073, .8194, .0073, .1659)$
*	MR vs MR	$(.4389, .1256, .0000, .4355)$
*	FP vs FP	$(0, 0, 0, 1)$
*	FP vs EFP	$(0, 0, 0, 1)$
*	FP vs PR	$(0, 0, 0, 1)$
*	EFP vs EFP	$(0, 0, 0, 1)$
*	EFP vs PR	$(0, 0, 0, 1)$
*	PR vs PR	$(0, 0, 0, 1)$

Behaviour Rule	$\max_\mu u_M(\mu, \alpha_N)$	$u_M(\alpha)$	$\max_\nu u_M(\alpha_M, \nu)$	$u_N(\alpha)$
R1 vs FP	1	1	5	5
R1 vs EFP	1	1	5	5
R1 vs PR	1	1	5	5
R1 vs R1	1	1	5	5
MR vs R1	4.94	4.94	.999	.994
MR vs FP	4.97	4.97	.999	.998
MR vs EFP	4.96	4.96	.998	.999
MR vs PR	.985	.203	.866	.837
MR vs MR	2.19	2.63	2.18	2.62
FP vs FP	1	1	5	5
FP vs EFP	1	1	5	5
FP vs PR	1	1	5	5
EFP vs FP	1	1	5	5
EFP vs PR	1	1	5	5
PR vs PR	1	1	5	5

Behaviour rules	Marginal for player 1	Marginal for player 2
R1 vs FP	$(0, 1)$	$(0, 1)$
R1 vs EFP	$(0, 1)$	$(0, 1)$
R1 vs PR	$(0, 1)$	$(0, 1)$
R1 vs R1	$(0, 1)$	$(0, 1)$
MR vs R1	$(.9990, .0010)$	$(.9887, .0113)$
MR vs FP	$(.9993, .0007)$	$(.9943, .0057)$
MR vs EFP	$(.9987, .0013)$	$(.9917, .0083)$
MR vs PR	$(.8267, .1733)$	$(.0147, .9853)$
MR vs MR	$(.5645, .4355)$	$(.4389, .5611)$
FP vs FP	$(0, 1)$	$(0, 1)$
FP vs EFP	$(0, 1)$	$(0, 1)$
FP vs PR	$(0, 1)$	$(0, 1)$
EFP vs EFP	$(0, 1)$	$(0, 1)$
EFP vs PR	$(0, 1)$	$(0, 1)$
PR vs PR	$(0, 1)$	$(0, 1)$

SPRGW - 30000 iterations

Behaviour Rule	$\max_\mu u_M(\mu, \alpha_N)$	$u_M(\alpha)$	$\max_\nu u_M(\alpha_M, \nu)$	$u_N(\alpha)$
R1 vs FP	.0471	$-.0411$	.0424	$+.0411$
R1 vs EFP	.0739	$-.0333$	.0332	$+.0333$
R1 vs PR	.0821	$-.0695$	.0399	$+.0695$
R1 vs R1	.1565	$+.0239$	.0673	$-.0239$
MR vs R1	.6453	$+.5678$	.4192	$-.5678$
MR vs FP	.6666	$-.0673$	.3278	$+.0673$
MR vs EFP	.9027	$-.2456$	.2513	$+.2456$
MR vs PR	.7866	$-.5287$	.5405	$+.5287$
MR vs MR	.9993	$-.0003$	.9987	$+.0003$
FP vs FP	.0101	$+.0000$	.0101	$+.0000$
FP vs EFP	.0066	$-.0038$	.0123	$+.0038$
FP vs PR	.0018	$-.0120$	.0108	$+.0120$
EFP vs EFP	.0061	$-.0068$	.0048	$+.0068$
EFP vs PR	.0115	$-.0088$	.0121	$+.0088$
PR vs PR	.0101	$+.0000$	.0101	$+.0000$

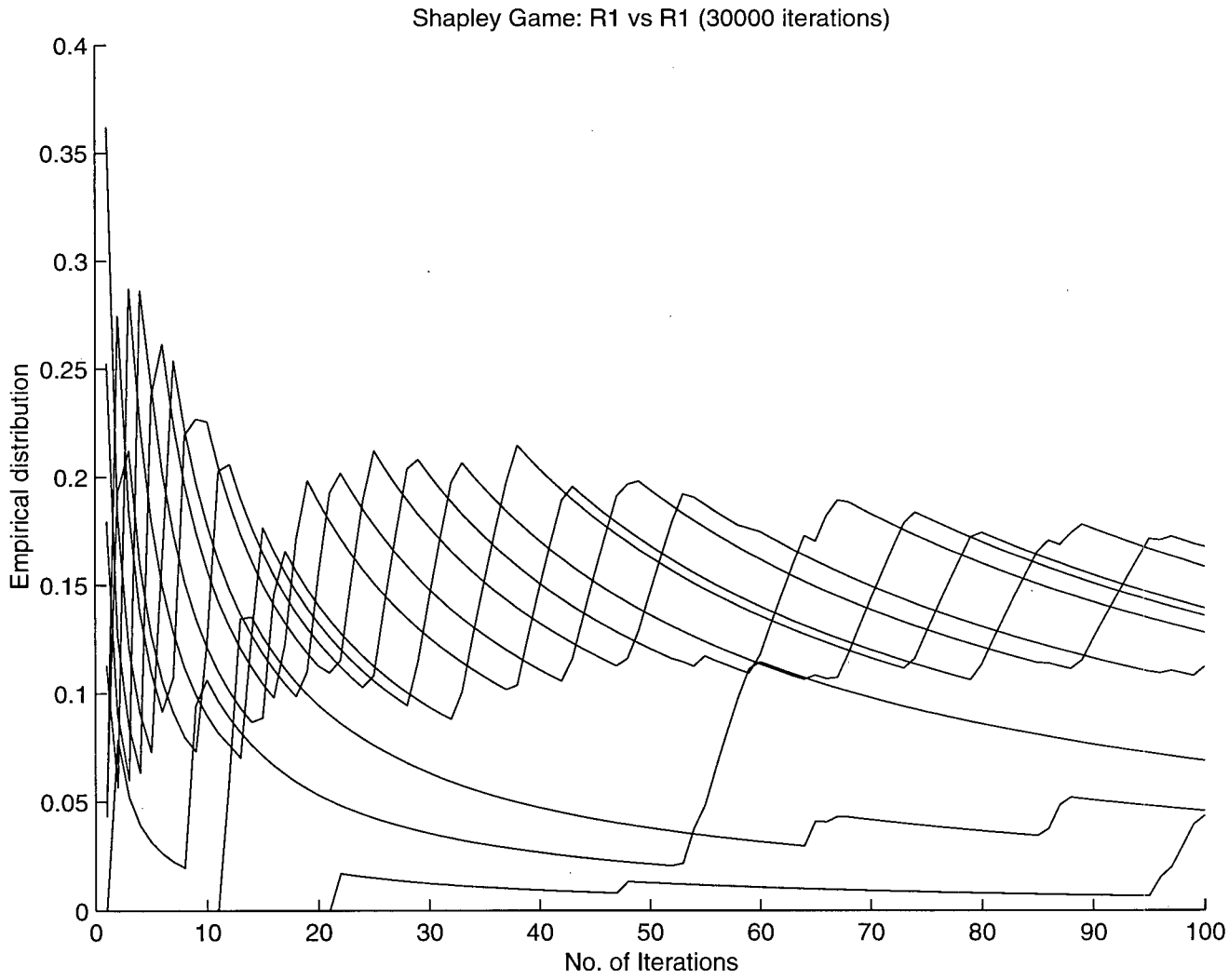
Behaviour rules	Marginal for player 1	Marginal for player 2
R1 vs FP	(.1011, .1358, .2991, .0924, .3717)	(.0756, .1569, .3027, .1173, .3475)
R1 vs EFP	(.1029, .1093, .3533, .1029, .3266)	(.1239, .0861, .3269, .0896, .3735)
R1 vs PR	(.0961, .1391, .3341, .0934, .3372)	(.1460, .0842, .3529, .1035, .3134)
R1 vs R1	(.1186, .0791, .3739, .1004, .3279)	(.1426, .1152, .2355, .1342, .3726)
MR vs R1	(.2636, .2229, .2814, .2141, .0130)	(.2301, .2332, .0617, .2437, .2313)
MR vs FP	(.2203, .2203, .3334, .2204, .0056)	(.2592, .21350, .1940, .3334)
MR vs EFP	(.1929, .1928, .3360, .1933, .0851)	(.2990, .2911, .0003, .3129, .0967)
MR vs PR	(.1917, .1175, .4187, .2558, .0164)	(.2887, .2129, .0002, .2851, .2131)
MR vs MR	(.3332, .3332, .0007, .3329, .0000)	(.3332, .3329, .0000, .3332, .0007)
FP vs FP	(.1102, .1117, .3300, .1153, .3329)	(.1102, .1117, .3300, .1153, .3329)
FP vs EFP	(.1036, .1115, .3259, .1230, .3360)	(.1096, .1152, .3286, .1076, .3390)
FP vs PR	(.1120, .1073, .3348, .1147, .3314)	(.1108, .1112, .3342, .1100, .3338)
EFP vs EFP	(.1055, .1155, .3288, .1126, .3376)	(.1157, .1092, .3275, .1083, .3393)
EFP vs PR	(.1159, .1087, .3298, .1173, .3283)	(.1178, .1241, .3260, .0918, .3404)
PR vs PR	(.1095, .1114, .3238, .1130, .3424)	(.1095, .1114, .3238, .1130, .3424)

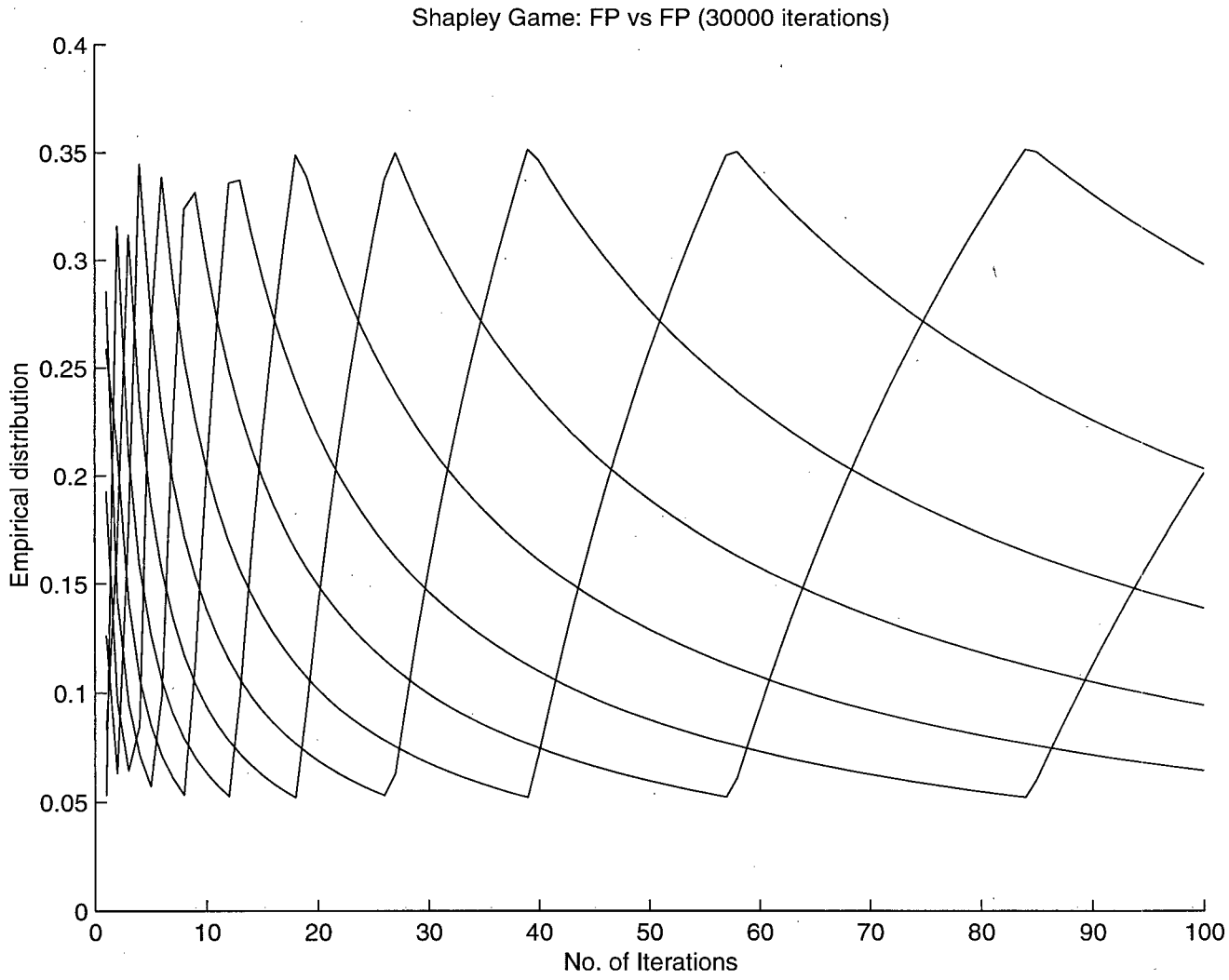
Convergence	Behaviour Rule	Empirical Distribution
−	R1 vs FP	(.0101, .0142, .0287, .0146, .0335, .0191, .0058, .0434, .0241, .0434, .0384, .0616, .0741, .0295, .0955, .0044, .0060, .0390, .0123, .0307, .0037, .0692, .1174, .0369, .1145)
−	R1 vs EFP	(.0022, .0015, .0257, .0226, .0509, .0221, .0007, .0386, .0255, .0225, .0575, .0483, .0982, .0247, .1246, .0041, .0001, .0486, .0000, .0551, .038, .0355, .1158, .0168, .1205)
−	R1 vs PR	(.0125, .0088, .0258, .0063, .0427, .0148, .0044, .0457, .0372, .0371, .0511, .0389, .1141, .0299, .1001, .0180, .0019, .0435, .0025, .0276, .0496, .0302, .1238, .0276, .1059)
+	R1 vs R1	(.0085, .0047, .0199, .0283, .0571, .0301, .0006, .0144, .0141, .0144, .0588, .0446, .01039, .0407, .1259, .0040, .0203, .0228, .0057, .0426, .0362, .0449, .0739, .0453, .1276)
+	MR vs R1	(.0128, .0177, .0439, .1715, .0177, .1687, .0139, .0016, .0238, .0149, .0329, .0249, .0118, .0304, .1815, .0112, .1690, .0044, .0150, .0145, .0047, .0077, .0000, .0029, .0027)
+	MR vs FP	(.0098, .1005, .0000, .0758, .0342, .0903, .0059, .0000, .0752, .0490, .0452, .0359, .0000, .0371, .2152, .1122, .0700, .0000, .0049, .0333, .0018, .0011, .0000, .0011, .0016)
+	MR vs EFP	(.0588, .0577, .0001, .0587, .0175, .0607, .0529, .0000, .0566, .0226, .0966, .0984, .0002, .1091, .0316, .0577, .0596, .0000, .0571, .0188, .0252, .0225, .0000, .0313, .0061)
*	MR vs PR	(.0384, .0381, .0001, .0384, .0767, .0385, .0012, .0000, .0760, .0017, .0844, .1709, 0, .0812, .0822, .1235, .0008, .0000, .0844, .0470, .0039, .0019, .0001, .0051, .0054)
*	MR vs MR	(.0000, .1666, .0000, .1666, .0000, .1666, .0000, .0000, .1666, .0000, .0003, .0000, .0000, .0000, .0003, .1663, .1663, .0000, .0000, .0003, .0000, .0000, .0000, .0000, .0000)
*	FP vs FP	(.1102, .0000, .0000, .0000, .0000, .0000, .1117, .0000, .0000, .0000, .0000, .0000, .3300, .0000, .0000, .0000, .0000, .0000, .1153, .0000, .0000, .0000, .0000, .0000, .3329)
+	FP vs EFP	(.0097, .0120, .0349, .0096, .0375, .0128, .0118, .0413, .0127, .0329, .0378, .0378, .1061, .0337, .1105, .0131, .0108, .0362, .0154, .0476, .0362, .0429, .1101, .0362, .1106)
*	FP vs PR	(.0087, .0119, .0381, .0135, .0397, .0135, .0121, .0351, .0110, .0355, .0381, .0396, .1089, .0380, .1102, .0125, .0131, .0373, .0126, .0392, .0380, .0345, .01148, .0350, .1092)
*	EFP vs EFP	(.0076, .0154, .0329, .0141, .0354, .016, .0076, .0371, .0149, .0399, .0392, .0352, .1116, .0353, .1075, .0145, .0152, .0370, .0078, .0382, .0385, .0358, .1088, .0362, .1183)
+	EFP vs PR	(.0132, .0129, .0388, .0112, .0398, .0131, .0165, .0341, .0081, .0370, .0413, .0418, .1038, .0300, .1129, .0136, .0130, .0378, .0111, .0418, .0366, .0399, .1114, .0315, .1089)
*	PR vs PR	(.1095, .0000, .0000, .0000, .0000, .0000, .1114, .0000, .0000, .0000, .0000, .0000, .3238, .0000, .0000, .0000, .0000, .0000, .1130, .0000, .0000, .0000, .0000, .0000, .3424)

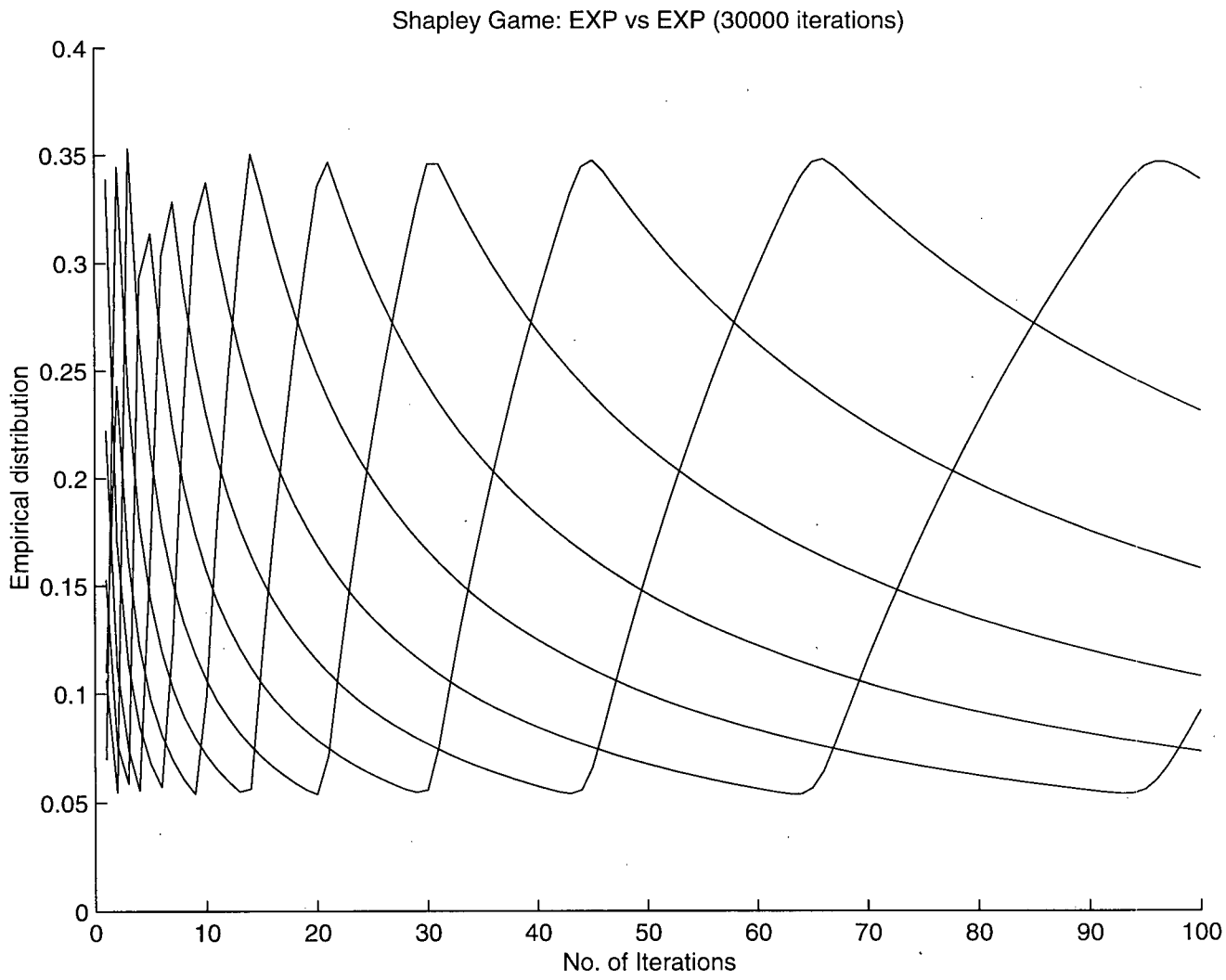
Shapley Game - 30000 iterations

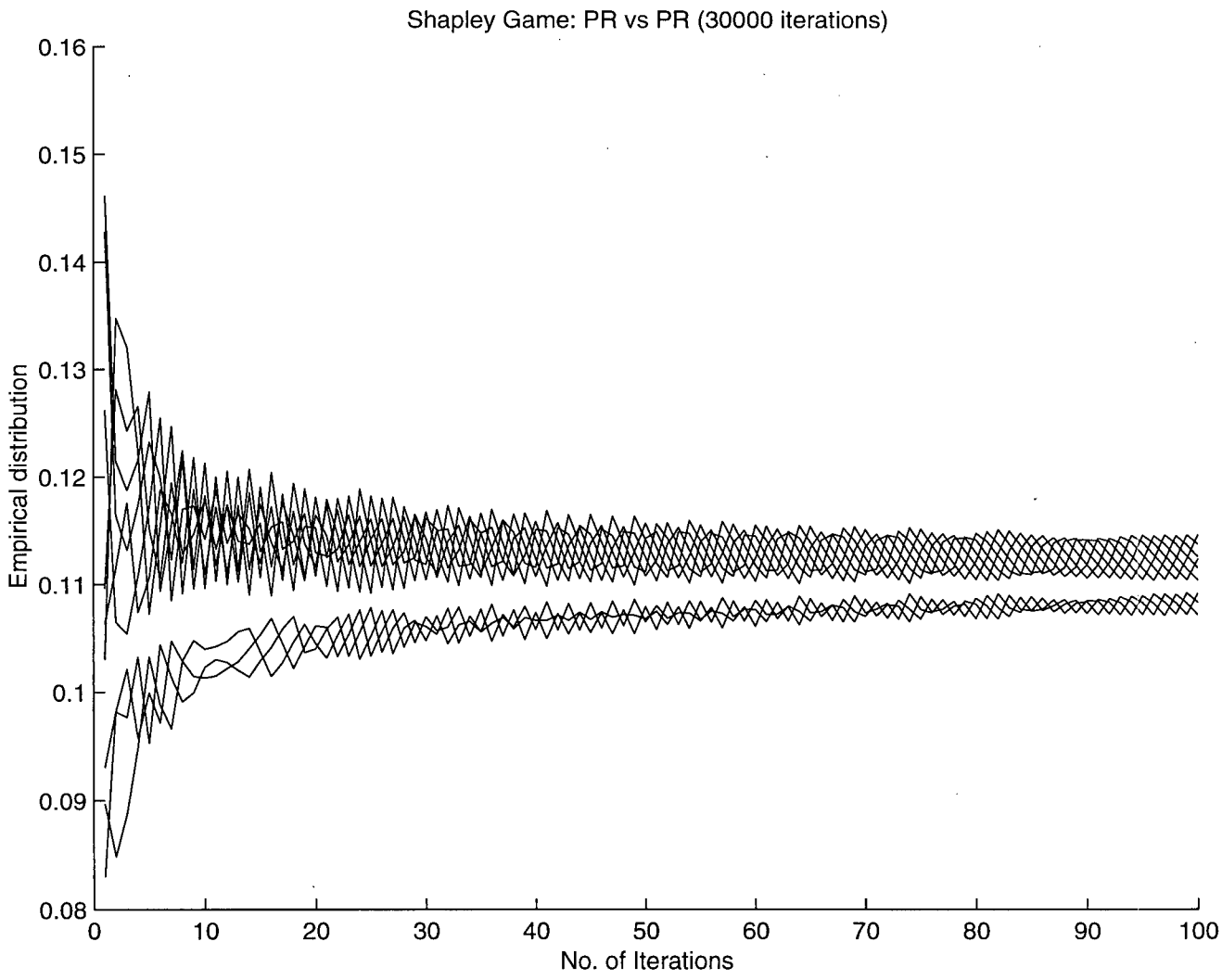
Convergence	Behaviour Rule	Empirical Distribution
−	R1 vs FP	(.1595, .0862, .0005, .0003, .1389, .1429, .2424, .0009, .2284)
−	R1 vs EFP	(.1892, .2093, .0000, .0376, .2678, .0677, .1216, .0018, .1050)
+	R1 vs PR	(.1389, .1608, .0302, .0393, .1610, .1637, .1386, .0365, .1310)
+	R1 vs R1	(.1363, .1492, .0058, .0153, .1818, .1975, .1141, .0038, .1961)
*	MR vs R1	(.2571, .0548, .0270, .0218, .2439, .0498, .0582, .0358, .2516)
*	MR vs FP	(.2667, .0667, .0000, .0000, .2667, .0667, .0667, .0000, .2666)
−	MR vs EFP	(.1489, .2094, .0142, .0774, .2427, .0559, .1158, .0624, .0734)
*	MR vs PR	(.0833, .1666, .0833, .0833, .0834, .1667, .1667, .0833, .0834)
*	MR vs MR	(.1667, .1667, .0000, .0000, .1667, .1667, .1667, .0000, .1667)
−	FP vs FP	(.0644, .0945, .0000, .0000, .1386, .2033, .2013, .0000, .2980)
−	FP vs EFP	(.0891, .1304, .0000, .0000, .1910, .2804, .0602, .0000, .2490)
−	FP vs PR	(.1137, .1833, .0013, .0015, .1835, .2962, .1135, .0011, .1059)
−	EFP vs EFP	(.2430, .0611, .0000, .0000, .0895, .1311, .2826, .0000, .1925)
−	EFP vs PR	(.1084, .1754, .0017, .0017, .1756, .2848, .1082, .0015, .1426)
*	PR vs PR	(.1138, .1147, 1083, .1072, .1124, .1104, .1126, .1092, .1115)

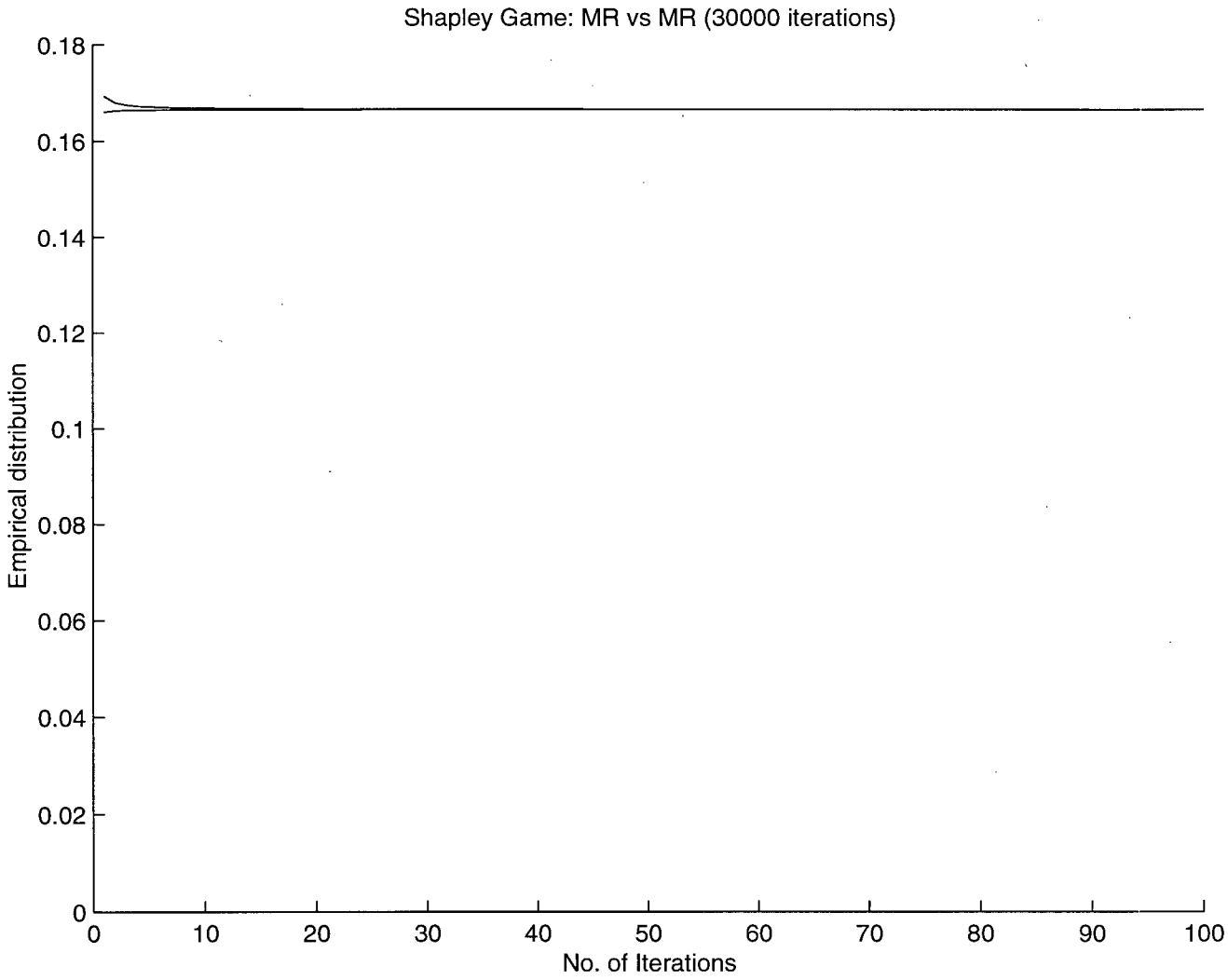
Behaviour Rule	$\max_\mu u_M(\mu, \alpha_N)$	$u_M(\alpha)$	$\max_\nu u_M(\alpha_M, \nu)$	$u_N(\alpha)$
R1 vs FP	.402	.527	.472	.472
R1 vs EFP	.479	.562	.399	.399
R1 vs PR	.358	.431	.364	.463
R1 vs R1	.399	.514	.395	.461
MR vs R1	.337	.752	.346	.163
MR vs FP	.333	.800	.333	.200
MR vs EFP	.514	.465	.376	.381
MR vs PR	.333	.250	.333	.500
MR vs MR	.333	.500	.333	.500
FP vs FP	.501	.501	.499	.499
FP vs EFP	.529	.529	.471	.471
FP vs PR	.403	.403	.481	.593
EFP vs EFP	.526	.525	.475	.475
EFP vs PR	.429	.427	.462	.568
PR vs PR	.336	.338	.337	.338


Shapley Game: R1 vs R1 (30000 iterations)
Empirical distribution
No. of Iterations
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
80
90
100


Shapley Game: FP vs FP (30000 iterations)
Empirical distribution
No. of Iterations
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
80
90
100


Shapley Game: EXP vs EXP (30000 iterations)
Empirical distribution
No. of Iterations
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
80
90
100


Shapley Game: PR vs PR (30000 iterations)
Empirical distribution
No. of Iterations
0.16
0.15
0.14
0.13
0.12
0.11
0.1
0.09
0.08
0
10
20
30
40
50
60
70
80
90
100


Shapley Game: MR vs MR (30000 iterations)
Empirical distribution
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60
70
80
90
100
No. of Iterations

Appendix B: Programs

All programs were written in Matlab. Though not presented here, they are available for public viewing at the following website: http://www.iam.ubc.ca/theses/ for any who are interested.

The main program is given first and is called game.m. Each behaviour rule has two separate programs – one for player M and one for player N. The line in program "game" which starts $j1 = ...$ determines what behaviour rule is used by player M while the line right below it determines what behaviour rule is used by player N. The program "game" will first ask the user to input the payoff matrices for both players. Thus, it will run any game the user wants.

Moreover by changing two lines, the user can run any pair of behaviour rules against each other. The output includes the empirical distribution (after however many iterations) of the game and, for each player, the payoff, the marginal distribution and the best-response payoff against the opponent's marginal distribution. (In other words, all the information displayed in the tables in Appendix A.)

Bibliography

[1] Blackwell, D. "An Analog of the Minimax Theorem for Vector Payoffs". (1954).

[2] Blackwell, D. "On Optimal Systems". Ann. Math. Stat. 25(1954), 394-397.

[3] Foster, D. and Vohra, R. "Calibrated Learning and Correlated Equilibrium". Games and Economic Behaviour 21(1997), 40-45.

[4] Fudenberg, D. and Levine D. "Consistency and Cautious Fictitious Play". Journal of Economic Dynamics & Control 19(1995), 1065-1089.

[5] Isaacson, D. Markov Chains: Theory and Applications. John Wiley & Sons. Toronto. 1976.

[6] Kao,E. An Introduction to Stochastic Processes. Duxbury Press. Toronto. 1997.

[7] Kijima, M. Markov Processes for Stochastic Modelling. Chapman & Hall. New York. 1997.

[8] Mas-Collel, A. and Hart, S. "A simple Adaptive Procedure Leading to Correlated Equilibrium", preprint (1998).

[9] Morris, Peter. Introduction to Game Theory. Springer-Verlag. New York: 1994.

[10] Robinson, Julia. "An Iterative Method of solving a Game". Ann. of Math. 54(1951), 296-301.

[11] Shapley, Lloyd. "Some Topics in Two-Person Games". Advances in Game Theory Princeton: Princeton University Press, 1964.