HYUN SONG SHIN

A RECONSTRUCTION OF JEFFREY'S NOTION OF RATIFIABILITY IN TERMS OF COUNTERFACTUAL BELIEFS

ABSTRACT. We formalize Jeffrey's (1983) notion of ratifiability and show that the resulting formal structure can be obtained more directly by means of a theory of counterfactual beliefs. One implication is that, under the appropriate formalizations, together with certain restrictions on beliefs, Bayesian decision theory and causal decision theory coincide.

Keywords: Ratifiability, counterfactuals, correlated equilibrium.

1. INTRODUCTION

In the second edition of his monograph, Jeffrey (1983) proposes the notion of ratifiability as a criterion of rational choice, intended to encompass (and supersede) his earlier theory (Jeffrey, 1965). Ratifiability is a type of stability of decision. In his own words,

> The notion of ratifiability is applicable only where, during deliberation, the agent finds it conceivable that he will not manage to perform the act he finally decides to perform, but will find himself performing one of the other available acts instead. ... The option in question is ratifiable or not depending on whether or not the expected desirability of actually carrying it out (having chosen it) is at least as great as the expected desirability of actually carrying out each of the alternatives (in spite of having chosen to carry out a different option, as hypothesized). [Jeffrey (1983), pp. 18–20]

However, Jeffrey's discussion is somewhat brief, and it is more suggestive than systematic, relying as it does on examples. This leaves open some tantalizing questions. For example, what lies behind the apparently paradoxical claim that a theory of rationality rests on a degree of fallibility on the part of the decision maker? Also, how is ratifiability related to other theories of decision and to standard game-theoretic solution concepts?

This paper is devoted to a formal investigation of the notion of ratifiability in an attempt to address some of these issues. Broadly, we take on two tasks.

Theory and Decision **31**: 21–47, 1991.

(i) We formalize Jeffrey's notion of ratifiability and investigate the game-theoretic structure which emerges. This exercise has its own rationale, but its main role is to provide a backdrop for our second task. Namely,
(ii) we provide an alternative characterization of ratifiability in terms of counterfactual beliefs. We develop the theory of counterfactuals due to Stalnaker (1968) and Lewis (1973), and show that the formal structures obtained in (i) can be obtained more directly via the notion of counterfactuals.

Throughout this paper, the game-theoretic solution concept of correlated equilibrium (Aumann 1974, 1987) plays the key role. The formal structures which emerge from ratifiability and from our theory of counterfactuals both coincide with the notion of correlated equilibrium. These representations form the basis of our reconstruction of ratifiability in terms of counterfactual beliefs.

The outline of the paper is as follows. In Section 2, we present a particular formalization of the notion of ratifiability and show that the resulting structure coincides with the notion of correlated equilibrium (Theorem 1). In Section 3, we present a theory of conterfactuals in games. We motivate the discussion with an example, and then present the general theory. Our second result (Theorem 2) shows that the formal structure we obtain coincides with that in Section 2. This representation theorem constitutes our reconstruction of ratifiability by means of counterfactual beliefs. Section 4 sums up our discussion.

2. FORMALIZING RATIFIABILITY

We shall conduct the discussion in terms of a two player normal form game. This allows us to formalize the notion of ratifiability in a perspicuous way and focus attention on substantive issues of interpretation. A one-person decision problem is construed as a game between the decision maker and Nature.

2.1. The Game G

Let $G := (S^1, S^2, h^1, h^2)$ be a two-player normal form game, where S^i is player i's strategy set and $h^i : S^1 \times S^2 \to \mathbb{R}$ is player i's payoff function.

We assume that both S^1 and S^2 are finite, with K^1 and K^2 elements respectively. We denote by s^i_j the jth strategy of player i. For the rest of this paper, we shall follow the notational convention of denoting individuals by superscripts, and strategies by subscripts. The superscript "$-i$" refers to player i's opponent.

2.2. *The State Space* Θ

We define a set of *propositions* Ψ consisting of the following propositions.

$$D^i_k := \text{"player } i \text{ decides to play } s^i_k\text{"}$$
$$P^i_k := \text{"player } i \text{ performs } s^i_k\text{"}$$

where i ranges over $\{1,2\}$ and k ranges over $\{1,\ldots,K^i\}$. Ψ has $2(K^1+K^2)$ elements.

Consider a function $\theta: \Psi \to \{0,1\}$. We say that θ is a *state* if, for all i, k and $l \neq k$,

$$\theta(D^i_k)=1 \Leftrightarrow \theta(D^i_l)=0 \quad \text{and} \quad \theta(P^i_k)=1 \Leftrightarrow \theta(P^i_l)=0 \tag{2.1}$$

A proposition q is *true* at θ if $\theta(q)=1$. q is *false* at θ otherwise. Denote by Θ the set of all states. Θ has $(K^1K^2)^2$ elements. Define the following subsets of Θ.

$$\begin{aligned} \delta^i_k &:= \{\theta \mid \theta(D^i_k)=1\} \\ \pi^i_k &:= \{\theta \mid \theta(P^i_k)=1\} \end{aligned} \tag{2.2}$$

δ^i_k is the event that i decides to play s^i_k, and π^i_k is the event that i performs s^i_k. By construction, these events do not coincide, and we leave open as a logical possibility the divergence between decisions and performances. The mnemonics of "δ" for "decision" and "π" for "performance" should help the reader keep track of the notation.

Let $\Delta^i := \{\delta^i_k\}$ and $\Pi^i := \{\pi^i_k\}$, where $k \in \{1,\ldots,K^i\}$. By (2.1), both Δ^i and Π^i partition Θ. Denote by Δ the meet of the partitions Δ^1 and Δ^2 and by Π the meet of the partitions Π^1 and Π^2. Equivalently,

$$\Delta := \{\delta^1 \cap \delta^2 \mid \delta^1 \in \Delta^1 \text{ and } \delta^2 \in \Delta^2\}$$
$$\Pi := \{\pi^1 \cap \pi^2 \mid \pi^1 \in \Pi^1 \text{ and } \pi^2 \in \Pi^2\} \tag{2.3}$$

2.3. ε-Ratifiability

Suppose player i has a probability distribution p over Θ. We define the real-valued function H^i as follows.

$$H^i(k \mid j) := \sum_{l=1}^{K^{-i}} p(\pi_l^{-i} \mid \delta_j^i \cap \pi_k^i) h^i(s_k^i, s_l^{-i}) \tag{2.4}$$

$H^i(k \mid j)$ is the payoff expected by player i when he decides on s_j^i but plays s_k^i. $H^i(k \mid j)$ is defined whenever $\delta_j^i \cap \pi_k^i$ is non-null under p.

Let some $\varepsilon > 0$ be given. We take ε to be small. In particular, $\varepsilon < \min_i\{1/K^i\}$. Consider the following conditions on p.

(A1) $p(\pi_j^i \mid \delta_k^i) = \varepsilon \quad \forall j \neq k$, whenever defined

(A2) $p(\pi^i \mid \delta^i \cap \delta^{-i}) = p(\pi^i \mid \delta^i) \quad \forall \pi^i, \delta^i, \delta^{-i}$, whenever defined

(A3) $H^i(j \mid j) \geq H^i(k \mid j) \quad \forall j, k$, whenever defined

(A1) formalizes the existence of trembles of size ε so that, given a decision to play a certain strategy, each of the other strategies may be performed with probability ε. (A2) states that such trembles are independent of the opponent's decisions. (A3) formalizes the deliberational stability condition. It states that given i's decision to play s_j^i, he cannot do better by performing some other strategy.

DEFINITION 1. p is ε-ratifiable for i if p satisfies (A1), (A2) and (A3).

Note that we have applied the term "ratifiable" to the probability distribution p itself rather than to particular actions. The rationale for this is that p specifies, among other things, the decisions of the individual.

The notion of ε-ratifiability is very close in spirit to Jeffrey's own informal discussion of ratifiability. However, the prescriptions delivered by the notion of ε-ratifiability will depend on the particular ε chosen. For any particular decision problem, the notion of ε-ratifiability will specify a whole family of prescriptions, depending on what value of ε is chosen. Arguably, this is a shortcoming. The motive for introducing trembles at all was merely to avoid attaching probability zero to any act. The precise *magnitude* of such trembles should play no part in the analysis. Rather, what matters is that such trembles exist, and that they be "small".

An elegant way of overcoming this sort of problem is a method used by Selten (1975) in which we work with those tremble-free distributions, i.e., those distributions with $\varepsilon = 0$, which can be obtained as the limit of a sequence of distributions *with* trembles.

2.4. *Modest Ratifiability*

Consider the following condition on p.

(A4) $\quad p(\pi_k^{-i}) = p(\delta_k^{-i}), \forall k$

We say that p is *modest for i* if it satisfies (A4). When i has a modest distribution, he believes that his opponent is not susceptible to trembles. The notion of ratifiability we employ is the following.

DEFINITION 2. p is *modestly ratifiable for i* if there are sequences $\langle p_t \rangle$ and $\langle \varepsilon_t \rangle$ such that, for all t, p_t is modest for i and ε_t-ratifiable for i, and $p_t \to p$ as $\varepsilon_t \to 0$.

We say that p is *modestly ratifiable* if p is modestly ratifiable for both players.

2.5. *Correlated Equilibrium*

Our first result is that the class of modestly ratifiable distributions coincides with the class of correlated equilibrium distributions. We begin by reviewing the notion of correlated equilibrium. Let p be a

tremble-free distribution on Θ, i.e. one for which $p(\delta_k^i) = p(\pi_k^i)$ for all i and k. In such a case, p defines an unambiguous distribution $\tilde{p}$ over the strategy set $S^1 \times S^2$. Define:

$$\phi^i(l \mid j) := p(\delta_l^{-i} \mid \delta_j^i) . \tag{2.5}$$

The distribution $\tilde{p}$ over $S^1 \times S^2$ is said to be a *correlated equilibrium distribution* if, for all i, j and k whenever $\phi^i(l \mid j)$ is defined,

$$\sum_{l=1}^{K^{-i}} \phi^i(l \mid j)[h^i(s_j^i, s_l^{-i}) - h^i(s_k^i, s_l^{-i})] \geqslant 0 . \tag{2.6}$$

The notion of correlated equilibrium is due to Aumann (1974, 1987), and has the following interpretation. Suppose there is an impartial arbitrator who administers an experiment in which a random outcome is observed. The space of outcomes is the strategy set $S^1 \times S^2$, and both players know the distribution of probabilities over this set. However, only the arbitrator can observe the actual outcome of the experiment. When the arbitrator observes the outcome $s = (s^1, s^2)$, he recommends to player 1 that s^1 should be played, and recommends to player 2 that s^2 should be played. Crucially, one player does not know the recommendation made to his opponent. A probability distribution over $S^1 \times S^2$ is a correlated equilibrium distribution if no player can achieve a higher expected payoff by departing from the arbitrator's recommendations, given that his opponent heeds the recommendations.

We illustrate this concept with an example. Consider the following game – the familiar "chicken game".

	L	R
T	6, 6	2, 7
B	7, 2	0, 0

There are three Nash equilibria in this game. Namely, (T, R), (B, L), and a mixed strategy equilibrium in which each player receives $4\frac{2}{3}$. However, both players can receive an expected payoff of 5 if they agree to the following coordination scheme. Before the game is

played, they appoint an impartial arbitrator who will administer an experiment in which a fair die is cast. The arbitrator then issues suggestions to each player on which action should be performed. A player only hears his *own* message and can only make probability judgements about the likely messages received by his opponent. The rule followed by the arbitrator is as follows:

Outcome	*Suggestion to 1*	*Suggestion to 2*
1 or 2	play T	play L
3 or 4	play T	play R
5 or 6	play B	play L

It can readily be verified that neither player obtains a higher payoff by departing from the suggestions (given that his opponent follows the suggestions). For example, if player 1 is recommended to play T, he infers that the outcome of the die is in $\{1, 2, 3, 4\}$. Conditional on this information, he infers that player 2 has been suggested L and R with equal probability. Thus, on the assumption that 2 follows the arbitrator's suggestions, if 1 plays T his expected payoff is $\frac{1}{2}(6) + \frac{1}{2}(2) = 4$, whereas if he departs from the suggestion and plays B, his payoff is $\frac{1}{2}(7) + \frac{1}{2}(0) = 3\frac{1}{2}$. By following the mechanism above, both players can expect a payoff of $\frac{1}{3}(7) + \frac{1}{3}(2) + \frac{1}{3}(6) = 5$, which exceeds the maximum symmetric payoff obtainable as a Nash equilibrium.

On a point of terminology, when p is a correlated equilibrium distribution, we shall simply refer to p as a correlated equilibrium.

THEOREM 1. *p is modestly ratifiable if and only if p is a correlated equilibrium.*

The proof of this and all other theorems appear in the appendix. For now, we remark on the interpretation of modest ratifiability supplied by the notion of correlated equilibrium. The distinction between *decisions* and *performances* is exactly analogous to the distinction between the *recommendations* issued by the arbitrator and the *actions* taken by the players. Notice the crucial role played by the assumption that p is shared by both players. Without it, we could not obtain an *equilibrium*.

In addition to the interpretation supplied by the notion of correlated equilibrium, we can draw on the discussions of the properties of correlated equilibrium (as in Aumann, 1974), and translate them directly into the idiom of ratifiability. Thus, when p is modestly ratifiable we have the following corollaries.

(i) No player will place positive probability on a strictly dominated strategy. (So that, in Newcomb's problem, both boxes are taken, and in the prisoners' dilemma, the players confess.)

(ii) In a two-person zero-sum game, the payoffs achievable by modest ratifiability cannot exceed the "value" of the game.

(iii) However, in non-zero sum games, the payoffs achievable by modest ratifiability can exceed the payoffs achievable as a Nash equilibrium (as in our example above).

3. COUNTERFACTUALS AND GAMES

In this section, we shall develop the Stalnaker–Lewis approach to counterfactuals by constructing a framework which supplies a determinate criterion for similarity of possible worlds. Specifically, by constructing a possible worlds space for a given normal form game, we shall capture the notion of similarity by means of a *metric* on this space. Thus, two possible worlds are "similar" if the distance between them is "small". This approach to the analysis of counterfactuals seems particularly suited to game theory, since the subject matter of game theory (strategies, payoffs, and probabilities) supplies some very natural metrics on the space of possible worlds. We exploit this to the full.

A flavour of our approach can be conveyed by the diagram below. Φ is the set of possible worlds, $|p|$ is the subset Φ at which the proposition p is true, and $|q|$ is the subset of Φ at which the proposition q is true. Suppose φ is the true world, and we are concerned with the truth or falsity of the counterfactual; "if p were the case, q would be the case". We denote this counterfactual by $p \mathbin{\Box\!\!\rightarrow} q$.

We implement the Stalnaker–Lewis criterion by introducing a metric m on Φ. We identify the closest possible world to φ in this metric in which p is true, and see whether q is also true there. If so, then

$p \Box\!\!\rightarrow q$ is true at φ. Otherwise, $p \Box\!\!\rightarrow q$ is false at φ. In Figure 1, the closest possible world to φ in which p is true is $\tilde{\varphi}$. But q is also true at $\tilde{\varphi}$. Thus, we conclude that $p \Box\!\!\rightarrow q$ is true at φ.

This account is only intended to be suggestive, and cannot be a rigorous definition, since there may be more than one "closest" possible world. Our formal definition will be to say that $p \Box\!\!\rightarrow q$ is true at φ if and only if, there is a closed sphere C around φ in the metric m such that $C \cap |p|$ is non-empty and $C \cap |p| \subseteq |q|$.

The counterfactuals which will be of particular interest to us are those of the form; "if player i were to play strategy x, his payoff would be higher". When the truth value of these counterfactuals are known, a natural rationality criterion suggests itself. Namely, a player should never find himself at a possible world at which, according to his metric m, his payoff would be higher if he were to deviate. This is the principle which motivates our rationality criterion.

We motivate our general theory with the chicken game introduced in the last section. We define the *state space* Ω for this game to be the set $\{\omega_{tl}, \omega_{tr}, \omega_{bl}, \omega_{br}\}$, where ω_{tl} is the state in which 1 plays T and 2 plays L, ω_{tr} is the state in which 1 plays T and 2 plays R, and so on. We shall suppose that each player has a partition of Ω so that, at any state, a player knows his own action, but cannot exclude any action on the part of his opponent. For this, the players' partitions $\mathcal{P}^1$ and $\mathcal{P}^2$ must be as in Figure 2.

We suppose that each player has a prior probability distribution p over Ω, and forms beliefs by conditioning on the element of his partition which contains the true state. In particular, a player attaches probabilities to propositions of the form; "player i plays strategy s".

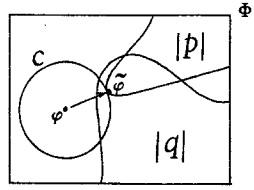

Fig. 1.

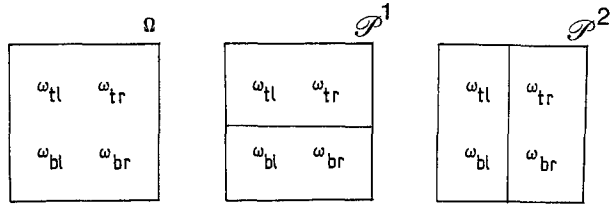

Fig. 2.

For example, if player 1 has the prior distribution p given by:

$$p(\omega_{tl}) = p(\omega_{tr}) = p(\omega_{bl}) = 1/3\,,$$

then at ω_{tl}, he has the following probability beliefs.

(i) T is played with probability 1.

(ii) L is played with probability 1/2.

(iii) R is played with probability 1/2.

More succinctly, we can combine (ii) and (iii) to give the proposition:

(ii)′ Player 2 randomizes (1/2, 1/2) over his strategy set.

In general, at any state $\omega \in \Omega$, each player attaches probability 1 to one of his own actions, and believes that his opponent randomizes with some pair of probabilities. For player i, we represent these beliefs by the pair $\langle x, y \rangle$, where x is a pure strategy of i, and y is a probability distribution over the strategies of i's opponent. The interpretation of the pair $\langle x, y \rangle$ is that, player i attaches probability 1 to his own strategy x, and believes that his opponent randomizes with distribution y. For example, player 1's beliefs at the state ω_{tl} is represented by the pair $\langle T, x \rangle$, where x is the randomization $\langle 1/2, 1/2 \rangle$. The set of all such pairs for player 1 is given by the product $\{T, B\} \times \mathbb{S}$, where $\mathbb{S}$ is the one-dimensional unit simplex representing the set of all probability distributions over $\{L, R\}$.

We shall define $\{T, B\} \times \mathbb{S}$ to be player 1's *possible worlds space*, and denote it by Φ^1. We denote by φ a typical element of this set, and

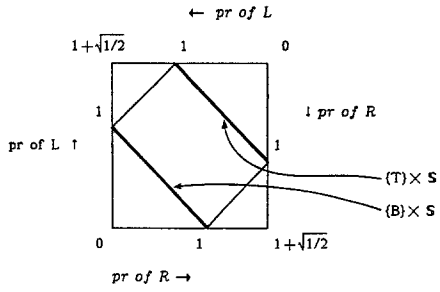

Fig. 3.

call it a *possible world* for player 1. Geometrically, we can represent Φ^1 as in Figure 3. Φ^1 is represented by the two parallel bold lines. The upper line is the set $\{T\} \times \mathbb{S}$, while the lower line is $\{B\} \times \mathbb{S}$. As we move toward the top left hand corner, the probability of L increases, and as we move toward the bottom right hand corner, the probability of R increases. The reason for this particular representation will become clear below when we introduce a particular metric on this space.

Given a prior probability distribution p for player 1, we can define a function $\beta^1 : \Omega \rightarrow \{T, B\} \times \mathbb{S}$ such that $\beta^1(\omega)$ represents the beliefs held by player 1 at ω. When $\beta^1(\omega) = \varphi$, we shall use the metaphor; "at ω, player 1 believes he is at the possible world φ". In Figure 4, we show the image of the function β^1.

Each possible world φ in Φ^1 determines a unique action for player 1

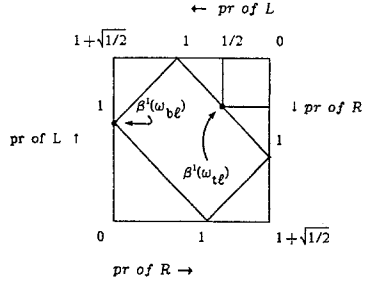

Fig. 4.

and a unique probability distribution over 2's actions. Hence, each φ determines an expected payoff for player 1. This is shown in Figure 5, where $H^1(\varphi)$ denotes the expected payoff of player 1 at φ.

Having thus defined the possible worlds space for player 1, the next step is to introduce a metric on this space which has the interpretation of player 1's "theory of the world". Each point in Φ^1 is a pair $\langle x, y\rangle$, where x is either T or B and $y = \langle y_1, y_2\rangle \in \mathbb{R}^2$, where $y_1 + y_2 = 1$. For any two points $\langle x, y\rangle$ and $\langle \tilde{x}, \tilde{y}\rangle$ in Φ^1, define the distance between them as:

$$\left[\sum_{i=1}^{2} |y_i - \tilde{y}_i|^2\right]^{1/2} \quad \text{if } x = \tilde{x},$$

$$\left[\sum_{i=1}^{2} |y_i - \tilde{y}_i|^2\right]^{1/2} + 1 \quad \text{if } x \neq \tilde{x}.$$

Geometrically, this metric could be dubbed the "library stack metric". Refer to Figure 6.

Imagine the two bold lines representing Φ^1 to be two parallel corridors in a library. There are stacks of shelves at right angles to the corridors. In order to get from one point in the library to another, one must follow the corridors and the spaces between the stacks. Thus, for example, the distance between φ and $\tilde{\varphi}$ in Figure 6 is the sum of two distances – between φ and $\hat{\varphi}$ (which is 1) and between $\hat{\varphi}$ and $\tilde{\varphi}$ (which is the Euclidean distance between $\hat{\varphi}$ and $\tilde{\varphi}$). We denote this metric by λ. As we shall see below, the general version of this metric plays a prominent role in our discussion.

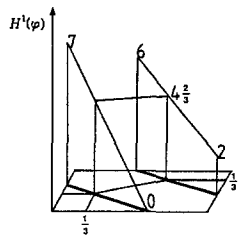

Fig. 5.

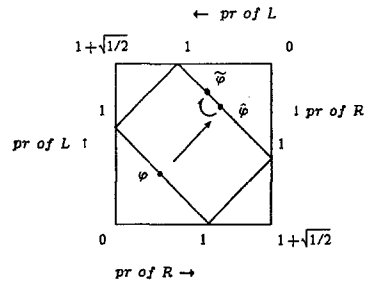

Fig. 6.

Given a player's possible worlds space and his metric, we have the apparatus to analyse the counterfactuals entertained by that player. Refer to Figure 7. Suppose player 1 believes that he is at the possible world φ. At φ, 1 plays T, and his payoff is 4. We are interested in the counterfactual; "If player 1 were to play B, his payoff would be higher". To evaluate this counterfactual at φ, we find the closest possible world(s) to φ at which player 1 plays B, and see whether 1's payoff is higher here than at φ. As we see in Figure 7, there is a unique closest possible world in which 1 plays B – namely, $\tilde{\varphi}$. But at $\tilde{\varphi}$, 1's payoff is 3.5. Thus, according to our criterion, the above counterfactual is false at φ.

Next, refer to Figure 8. Suppose player 1 believes he is at the possible world ψ. His action at ψ is B, and his payoff at ψ is 7. The closest possible world to ψ in which 1 plays T is $\tilde{\psi}$, and his payoff there is 6. Thus, the counterfactual "If player 1 were to play T, his payoff would be higher" is false at ψ.

This suggests a very natural rationality criterion for player 1. Namely, that he should never find himself at a possible world at which, according to his metric λ, he would be strictly better off if he were to deviate. We give the formal definition of this rationality criterion below. For now, notice that this criterion is satisfied by player 1 in our example, since

$$\beta^1(\omega_{tl}) = \beta^1(\omega_{tr}) = \varphi \,,$$
$$\beta^1(\omega_{bl}) = \beta^1(\omega_{br}) = \psi \,,$$

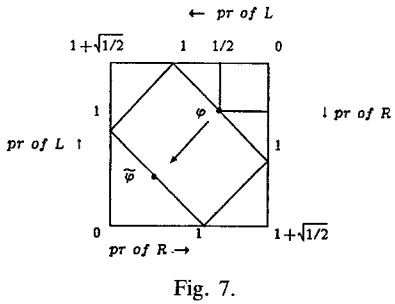

Fig. 7.

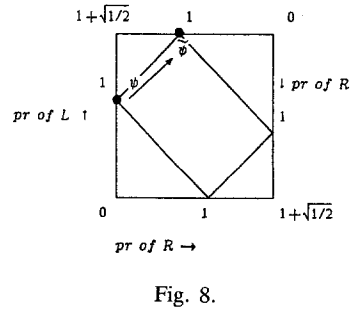

Fig. 8.

and we showed that both φ and ψ satisfy the rationality requirement sketched above.

Let us now proceed to make precise the concepts introduced above. The order of the discussion follows the discussion above, except that we consider the n-person case.

3.1. *The Game G*

Let G be a game in normal form between n players. $G = \langle S, h \rangle$, where $S = \times_{i=1}^{n} S^i$ and $h = \langle h^1, \ldots, h^n \rangle$. S^i is player i's strategy set, and h^i is his payoff function $h^i : S \to \mathbb{R}$. We assume that each S^i is finite and has K^i elements. Denote by s_j^i the jth strategy of player i. Let $K := \prod_{i=1}^{n} K^i$, so that S has K elements. Denote by S^{-i} the product $S^1 \times S^2 \times \cdots \times S^{i-1} \times S^{i+1} \times \cdots \times S^n$. S^{-1} has K/K^i elements. We order this set in some well-defined manner by the index set $\{1, 2, \ldots, K^{-i}\}$, where $K^{-i} = K/K^i$. Thus, $S^{-1} = \{s_1^{-i}, s_2^{-i}, \ldots, s_{K^{-i}}^{-i}\}$.

3.2. *The State Space* Ω

With each strategy n-tuple $s \in S$, we associate a unique *state* ω. The set of all such states is the *state space* Ω. Then we can define a function $\mathbf{s} : \Omega \to S$ such that $\mathbf{s}(\omega)$ is the strategy n-tuple associated with ω. The function $\mathbf{s}$ is therefore the n-tuple $\langle \mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^n \rangle$, where $\mathbf{s}^i$ is the function $\mathbf{s}^i : \Omega \to S^i$ such that $\mathbf{s}^i(\omega)$ is the strategy of player i associated with the state ω. We denote by $\mathbf{s}^{-i}$ the $(n-1)$-tuple $\langle \mathbf{s}^1, \ldots, \mathbf{s}^{i-1}, \mathbf{s}^{i+1}, \ldots, \mathbf{s}^n \rangle$. Let $\mathcal{P}^i$ be the partition of Ω generated by the equivalence relation $\equiv^i$ defined as; $\omega \equiv^i \omega' \Leftrightarrow \mathbf{s}^i(\omega) = \mathbf{s}^i(\omega')$. We denote by $P^i(\omega)$ the element of $\mathcal{P}^i$ containing the state ω. Finally, each player i has a probability distribution p^i over Ω.

3.3. *The Possible Worlds Space* Φ^i

For each player i, we define his *possible worlds space* Φ^i as the product

$$\Phi^i := S^i \times \Delta(S^{-i}), \tag{3.1}$$

where $\Delta(S^{-i})$ is the unit simplex of dimension $K^{-i} - 1$, representing

the set of all probability distributions over the set S^{-i}. An element of Φ^i will be called a *possible world*, and be denoted by φ. In turn, $\varphi = \langle \varphi^i, \varphi^{-i} \rangle$, where φ^i is the projection of φ into S^i (so that $\varphi^i \in S^i$) and φ^{-i} is the projection of φ into $\Delta(S^{-i})$ (so that $\varphi^{-i} \in \mathbb{R}^{K^{-i}}$). In particular, φ^{-i} will be denoted by the K^{-i}-dimensional vector:

$$(3.2) \quad \langle \varphi^{-i}[s_1^{-i}], \varphi^{-i}[s_2^{-i}], \ldots, \varphi^{-i}[s_{K^{-i}}^{-i}] \rangle ,$$

where $\varphi^{-i}[s_k^{-i}]$ has the interpretation of the probability weight given to s_k^{-i} in φ^{-i}. When no confusion is likely, we shall abbreviate (3.2) as:

$$(3.3) \quad \langle \varphi_1^{-i}, \varphi_2^{-i}, \ldots, \varphi_{K^{-i}}^{-i} \rangle .$$

Finally, with each possible worlds space Φ^i, we associate a metric m on Φ^i. This metric has the interpretation of the theory with which player i forms counterfactual beliefs. Call the pair $\langle \Phi^i, m \rangle$ player i's *theory space*.

3.4. The Belief Function β^i

Consider the posterior probability attached to s_k^{-i} by player i at ω, obtained by conditioning on his partition $\mathcal{P}^i$. We denote this probability by $\varphi_k^{-i}(i, \omega)$. More precisely,

$$(3.4) \quad \varphi_k^{-i}(i, \omega) := p^i(\{\omega \mid \mathbf{s}^{-i}(\omega) = s_k^{-i}\} \mid P^i(\omega)) .$$

Denote by Ω_+^i the set $\{\omega \mid p^i(\omega) > 0\}$, and define the function $\mathbf{t}^i : \Omega_+^i \to \Delta(S^{-i})$ as follows:

$$(3.5) \quad \mathbf{t}^i(\omega) := \langle \varphi_1^{-i}(i, \omega), \varphi_2^{-i}(i, \omega), \ldots, \varphi_{K^{-i}}^{-i}(i, \omega) \rangle .$$

From this, we define player i's *belief function* $\beta^i : \Omega_+^i \to \Phi^i$ as;

$$(3.6) \quad \beta^i := \langle \mathbf{s}^i, \mathbf{t}^i \rangle .$$

We can give the following interpretation to this function. Player i's partition $\mathcal{P}^i$ serves as his information partition. At the state ω, i

computes posterior probabilities by conditioning on $P^i(\omega)$. Thus, at the state ω, i attaches probability 1 to one of his own actions and attaches various probabilities to his opponents' strategy combinations s^{-i}. Each point in the possible worlds space Φ^i constitutes a possible state of belief for player i obtained in this manner. The belief function β^i is constructed so that, at ω, player i has the set of probability beliefs given by the possible world $\beta^i(\omega)$. More figuratively, we say that, at the state ω, player i believes that he is "at" the possible world $\beta^i(\omega)$.

3.5. *The Payoff Function* H^i

Each possible world $\varphi \in \Phi^i$ determines a probability distribution over S. By taking the weighed sum of $h^i(s)$ over $s \in S$ with the weights given by φ, we arrive at the expected payoff of player i at the possible world φ. We shall denote player i's expected payoff at the possible world φ as $H^i(\varphi)$. To define this formally, let I^i_j be the indicator function $I^i_j : S^i \to \{0, 1\}$ such that,

$$(3.7) \qquad \begin{aligned} I^i_j(s^i) &= 1 \quad \text{if} \quad s^i = s^i_j \\ &= 0 \quad \text{otherwise} . \end{aligned}$$

The function $H^i : \Phi^i \to \mathbb{R}$ is defined as follows.

$$(3.8) \qquad H^i(\varphi) = \sum_{j=1}^{K^i} \sum_{k=1}^{K^{-i}} I^i_j(\varphi^i) \varphi^{-i}_k h^i(s^i_j, s^{-i}_k) .$$

To verify that H^i is in accordance with the intuition outlined above, note that from (3.4) and the definition of β^i, when $\beta^i(\omega) = \varphi$, $I^i_j(\varphi^i)\varphi^{-i}_k$ is the probability of the event $\{\omega \mid \mathbf{s}(\omega) = \langle s^i_j, s^{-i}_k \rangle\}$ according to p^i, conditional on $P^i(\omega)$.

3.6. *The Proposition Set* Ψ^i

We shall associate with each player i a set of *propositions* Ψ^i. This set is defined by the following three clauses.

(i) The following are elements of Ψ^i.
$\sigma^i_j :=$ "player i plays s^i_j", for all $j \in \{1, \ldots, K^i\}$,

σ_μ^{-i} := "the set of players except i play the (possibly correlated) strategy μ", for all $\mu \in \Delta(S^{-i})$,

π_r^i := "player i's payoff does not exceed r", for all $r \in \mathbb{R}$.

(ii) Suppose $\psi, \chi \in \Psi^i$. Then $\psi \,\square\!\!\rightarrow\, \chi$ is an element of Ψ^i.

(iii) Ψ^i is the smallest set satisfying (i) and (ii).

3.7. *Truth Conditions and Events*

Given player i's theory space $\langle \Phi^i, m \rangle$, consider a function $e: \Phi^i \times \Psi^i \to \{0, 1\}$. For any such function, denote by $|\psi|_e$ the set $\{\varphi \in \Phi^i \mid e(\varphi, \psi) = 1\}$. We say that the function e is the *truth function* relative to the metric m if, for all $\varphi \in \Phi^i$ and $\psi \in \Psi^i$,

(i) $e(\varphi, \sigma_j^i) = 1 \quad \Leftrightarrow \quad \varphi^i = s_j^i, \ \forall j \in \{1, \ldots, K^i\}$

(ii) $e(\varphi, \sigma_\mu^{-i}) = 1 \quad \Leftrightarrow \quad \varphi^{-i} = \mu, \ \forall \mu \in \Delta(S^{-i})$

(iii) $e(\varphi, \pi_r^i) = 1 \quad \Leftrightarrow \quad H^i(\varphi) \leqslant r, \ \forall r \in \mathbb{R}$

(iv) $e(\varphi, \psi \,\square\!\!\rightarrow\, \chi) = 1 \quad \Leftrightarrow$ there is a closed sphere C around φ in the metric m such that $C \cap |\psi|_e$ is non-empty and $C \cap |\psi|_e \subseteq |\chi|_e$.

When e is a truth function, we say that ψ is *true at* φ if $e(\varphi, \psi) = 1$. ψ *is false at* φ otherwise. Worthy of note is clause (iv) formalizing the truth condition for counterfactual propositions. It states that $\psi \,\square\!\!\rightarrow\, \chi$ is true at the possible world φ if, and only if, in the closest possible world(s) to φ in which ψ is true, χ is also true.

When e is a truth function, we shall drop the subscript e from $|\psi|_e$. In this case, we call $|\psi|$ the *event* corresponding to ψ.

3.8. *Rationality*

Suppose $\beta^i(\omega) = \varphi$, and player i holds the metric m. We say that i is *m-rational at* ω if,

(3.9) For all $j \in \{1, \ldots, K^i\}$, there is $r \leqslant H^i(\varphi)$ such that $\sigma_j^i \,\square\!\!\rightarrow\, \pi_r^i$ is true at φ.

In other words, i is m-rational at ω if i believes that he is at a possible world at which, according to his metric m, he would not gain if he were to deviate. In general, we say that i is *m-rational* if i is m-rational at all states on which β^i is defined. The notion of m-rationality is a stipulation on a player's probability distribution. Namely, a player's probability distribution should be such that, given his metric m, he will not find himself in a situation in which he believes that he *would* do better if he *were* to deviate. Thus, m-rationality describes a type of consistency – the consistency of a player's probability distribution with his metric.

3.9. The Metric λ

We explore the equilibrium structure arising from the "library stack metric" in the n-player case. This metric will be the sum of two metrics on the two component sets of Φ^i. Denote by λ^i the discrete metric on S^i and by λ^{-i} the Euclidean norm on $\Delta(S^{-i})$. That is,

$$\lambda^i(s^i_j, s^i_q) = \begin{cases} 0 & \text{if} \quad j = q \\ 1 & \text{otherwise} \end{cases} \tag{3.10}$$

$$\lambda^{-i}(\varphi^{-i}, \hat{\varphi}^{-i}) = \|\varphi^{-i} - \hat{\varphi}^{-i}\| = \left(\sum_{k=1}^{K^{-i}} |\varphi^{-i}_k - \hat{\varphi}^{-i}_k|^2\right)^{1/2} \tag{3.11}$$

Thus, suppose $\varphi = \langle \varphi^i, \varphi^{-i} \rangle$ and $\hat{\varphi} = \langle \hat{\varphi}^i, \hat{\varphi}^{-i} \rangle$. We define λ as the sum of λ^i and λ^{-i}.

$$\lambda(\varphi, \hat{\varphi}) := \lambda^i(\varphi^i, \hat{\varphi}^i) + \lambda^{-i}(\varphi^{-i}, \hat{\varphi}^{-i}) . \tag{3.12}$$

3.10. Aumann Rationality

We now come to our second theorem. We shall demonstrate that a player who is rational with respect to the metric λ will act in accordance with the criterion of rationality as set out by Aumann (1987), and conversely. Denote by c^i_q the constant function $c^i_q : \Omega \to S^i$ such that $c^i_q(\omega) = s^i_q$, for all ω. We say that player i is *Aumann-rational* at the state ω if:

$$(3.13)\quad E(h^i(\mathbf{s}) \mid P^i(\omega)) \geq E(h^i(c_q^i, \mathbf{s}^{-i}) \mid P^i(\omega)),$$
$$\text{for all } q \in \{1, \ldots, K^i\}.$$

where E denotes the expectation with respect to the distribution p^i.

3.11. The Theorem

THEOREM 2. *Player i is λ-rational at ω if and only if i is Aumann-rational at ω.*

The following corollaries are straightforward consequences of this result. (See also Aumann's 1987 main theorem.)

COROLLARY 3.1. *Suppose $p^i = p$, $\forall i$. Then all players are λ-rational if and only if p is a correlated equilibrium.*

COROLLARY 3.2. *Suppose $p^i = p$, $\forall i$, and p is independent. Then all players are λ-rational if and only if the mixed strategies given by p is a Nash equilibrium.*

We are now in a position to tie together the results in Sections 2 and 3. Theorem 1 has identified the class of ratifiable distributions, while Theorem 2 and its corollaries have identified the class of counterfactually rational distributions. In both cases, they coincide with the class of correlated equilibria. Thus, we conclude that, under our formalizations, ratifiability and counterfactual rationality coincide. We have thus accomplished the main task of this paper.

4. AN OVERVIEW

We conclude with some general comments surveying the terrain we have covered. It was claimed in Section 2 that, by working with tremble-free distributions which could be obtained as the limit of a sequence of distributions *with* trembles, we could, as it were, have our

cake and eat it too. That is, we could help ourselves to the *content* of ε-ratifiability without actually having trembles at all.

The objection to this claim is clear enough – that, although we are working with tremble-free distributions, as long as we make reference to trembles of any form in the definition of ratifiability, we cannot remain untinged of associations therewith. This is a criticism which must be taken seriously, and it is sufficiently forceful to persuade us to search for alternative formulations of rational choice which goes beyond ratifiability. Having worked through the constructions in Sections 2 and 3, the reader will have gained some idea of the role played by ratifiability. The device of trembles is brought in, not because of the intrinsic interest of players "trembling", but because we want to evaluate propositions of the form:

(4.1) "I play x, but if I were to play y, the consequence would be z".

In other words, ratifiability is a device for formalizing counterfactual beliefs about those acts which are given zero probability in equilibrium. The postulates (A1) to (A4) are instrumental in providing a particular theory of counterfactuals within which to evaluate these statements.

It is this which motivates the discussion of counterfactuals in Section 3. By tackling counterfactuals head on, we can by-pass the enterprise of defining rationality in terms of concepts such as trembles. In short, rather than packing away all the assumptions into the postulates (A1) to (A3), we are able to lay bare the workings of the relevant counterfactuals.

ACKNOWLEDGEMENT

I am grateful to Michael Bacharach, Ken Binmore, Cristina Bicchieri, Bill Harper, Peter Gärdenfors, Brian Skyrms and the referee of this journal for their comments. A part of the material in this paper was presented at the workshop on game theory, Castiglioncello, Italy, in June 1989.

APPENDIX 1

Proof of Theorem 1.

We begin with some preliminary remarks. Let some $\varepsilon > 0$ be given. Since Π^i partitions Θ, $p(\pi^i_k \mid \delta^i_j) = 1 - \Sigma_{m \neq k}\, p(\pi^i_m \mid \delta^i_j)$. Thus, when p satisfies (A1) for player i,

$$(5.1) \qquad p(\pi^i_k \mid \delta^i_j) = \begin{cases} \varepsilon & \text{if} \quad j \neq k \\ 1 - (K^i - 1)\varepsilon & \text{if} \quad j = k \end{cases}$$

Then, since $p(\delta^i_j \cap \pi^i_k) = p(\pi^i_k \mid \delta^i_j)p(\delta^i_j)$, we have:

$$(5.2) \qquad p(\delta^i_j \cap \pi^i_k) = 0 \Leftrightarrow p(\delta^i_j) = 0$$

Suppose p satisfies (A1) and (A2) for i and is modest for i. Then, consider the probability $p(\pi^{-i}_l \mid \delta^i_j \cap \pi^{-i}_k)$ when $p(\delta^i_j) > 0$. We have two cases.

Case (i) $p(\delta^i_j \cap \delta^{-i}_l) > 0$

$$p(\pi^{-i}_l \mid \delta^i_j \cap \pi^i_k) = \frac{p(\pi^{-i}_l \cap \delta^i_j \cap \pi^i_k)}{p(\delta^i_j \cap \pi^i_k)}$$

$$= \frac{p(\delta^{-i}_l \cap \delta^i_j \cap \pi^i_k)}{p(\delta^i_j \cap \pi^i_k)}\,, \text{ by modesty}$$

$$= \frac{p(\pi^i_k \mid \delta^{-i}_l \cap \delta^i_j)p(\delta^{-i}_l \cap \delta^i_j)}{p(\pi^i_k \mid \delta^i_j)p(\delta^i_j)}$$

$$= \frac{p(\delta^{-i}_l \cap \delta^i_j)}{p(\delta^i_j)}\,, \text{ by (A2)}.$$

Case (ii) $p(\delta^i_j \cap \delta^{-i}_l) = 0$

$$p(\pi^{-i}_l \mid \delta^i_j \cap \pi^i_k) = \frac{p(\pi^{-i}_l \cap \delta^i_j \cap \pi^i_k)}{p(\delta^i_j \cap \pi^i_k)}$$

$$= \frac{p(\delta_l^{-i} \cap \delta_j^i \cap \pi_k^i)}{p(\delta_j^i \cap \pi_k^i)}, \text{ by modesty}$$

$$= 0$$

$$= \frac{p(\delta_l^{-i} \cap \delta_j^i)}{p(\delta_j^i)} .$$

In other words, when p satisfies (A1), (A2) and modesty for i, whenever $p(\delta_j^i) > 0$,

$$(5.3) \qquad p(\pi_l^{-i} \mid \delta_j^i \cap \pi_k^i) = p(\delta_l^{-i} \mid \delta_j^i), \quad \forall l, j, k .$$

We can then prove Theorem 1. First, we show that when p is modestly ratifiable, p is a correlated equilibrium. We fix a player i and let $\langle p_t \rangle$ and $\langle \varepsilon_t \rangle$ be sequences such that, p_t is ε_t-ratifiable for i for all t, p_t is modest for i for all t, and $p_t \to p$ as $\varepsilon_t \to 0$. Then, by (A3) and (5.2), whenever $p_t(\delta_j^i) > 0$,

$$(5.4) \qquad \sum_l p_t(\pi_l^{-i} \mid \delta_j^i \cap \pi_j^i) h^i(s_j^i, s_l^{-i})$$
$$\geqslant \sum_l p_t(\pi_l^{-i} \mid \delta_j^i \cap \pi_k^i) h^i(s_k^i, s_l^{-i}), \quad \forall j, k .$$

where l ranges over $\{1, \ldots, K^{-i}\}$. Denote the left hand side of (5.4) by L_t and the right hand side by R_t. Then,

$$\lim_{t \to \infty} L_t = \sum_{l=1}^{K^{-i}} \lim_{t \to \infty} p_t(\pi_l^{-i} \mid \delta_j^i \cap \pi_j^i) h^i(s_j^i, s_l^{-i})$$

$$= \sum_{l=1}^{K^{-i}} \lim_{t \to \infty} p_t(\delta_l^{-i} \mid \delta_j^i) h^i(s_j^i, s_l^{-i}), \text{ by (5.3)}$$

$$= \sum_{l=1}^{K^{-i}} p(\delta_l^{-i} \mid \delta_j^i) h^i(s_j^i, s_l^{-i}), \text{ since } p_t \to p .$$

Similarly,

$$\lim_{t\to\infty} R_t = \sum_{l=1}^{K^{-i}} p(\delta_l^{-i} \mid \delta_j^i) h^i(s_k^i, s_l^{-i}) .$$

Since $L_t \geq R_t$ for all t, $\lim_{t\to\infty} L_t \geq \lim_{t\to\infty} R_t$. This argument can be repeated for both players. Thus, whenever δ_j^i is non-null,

$$(5.5) \quad \sum_{l=1}^{K^{-i}} \phi^i(l \mid j)[h^i(s_j^i, s_l^{-i}) - h^i(s_k^i, s_l^{-i})] \geq 0 , \quad \forall i, j, k .$$

which is the condition for p being a correlated equilibrium.

We now prove the converse. Namely, when (5.5) holds for non-null δ_j^i, p is modestly ratifiable. Take some player i. Let $\langle \varepsilon_t \rangle$ be some sequence which converges to zero. We construct a sequence $\langle p_t \rangle$ as follows.

(i) $p_t(\delta) = p(\delta)$, $\forall \delta, t$,

(ii) p_t is modest for i, $\forall t$,

(iii) p_t satisfies (A1) and (A2) for i, where $\varepsilon = \varepsilon_t$, $\forall t$.

Then $p_t \to p$ as $\varepsilon_t \to 0$. Moreover, p_t is modest for i, for all t. Thus, to prove that p is modestly ratifiable for i, it remains to check that p_t is ε_t-ratifiable for i, for all t. Of all three conditions for ε_t-ratifiability, (A1) and (A2) hold by construction. To see that (A3) holds as well, note that from (i) and (5.5), whenever δ_j^i is non-null,

$$(5.6) \quad \sum_{l=1}^{K^{-i}} p_t(\delta_l^{-i} \mid \delta_j^i)[h^i(s_j^i, s_l^{-i}) - h^i(s_k^i, s_l^{-i})] \geq 0, \quad \forall j, k, t .$$

Then, from (5.3) and (5.2),

$$(5.7) \quad \sum_l p_t(\pi_l^{-i} \mid \delta_j^i \cap \pi_j^i) h^i(s_j^i, s_l^{-i}) \geq \sum_l p_t(\pi_l^{-i} \mid \delta_j^i \cap \pi_k^i) h^i(s_k^i, s_l^{-i}), \quad \forall j, k, t ,$$

whenever such expressions are defined. This is the condition (A3). Thus, p is modestly ratifiable for i. We can construct such a sequence for both players, so that p is modestly ratifiable for both players. This proves our theorem.

APPENDIX 2

Proof of Theorem 2.

Denote by $\varphi\backslash s_q^i$ the vector $\langle s_q^i, \varphi^{-i}\rangle$. In other words, $\varphi\backslash s_q^i$ is the possible world obtained from φ by replacing φ^i with s_q^i. We also use the following abbreviations.

$$[s_l^i] = \{\omega \mid \mathbf{s}^i(\omega) = s_l^i\},$$

$$[s_k^{-i}] = \{\omega \mid \mathbf{s}^{-i}(\omega) = s_k^{-i}\},$$

$$[s] = \{\omega \mid \mathbf{s}(\omega) = s\}.$$

A final piece of notation: we define $S_{il} = \{s \in S \mid s^i = s_l^i\}$. That is, S_{il} is the set of all strategy combinations in which player i plays s_l^i.

PROPOSITION 6.1. *Suppose* $\beta^i(\omega) = \varphi$. *Then,*

(i) $H^i(\varphi) = E(h^i(\mathbf{s}) \mid P^i(\omega))$

(ii) $H^i(\varphi\backslash s_q^i) = E(h^i(c_q^i, \mathbf{s}^{-i}) \mid P^i(\omega))$.

Proof. (i) $\mathbf{s}^i(\omega) = s_l^i$ for some $l \in \{1, \ldots, K^i\}$. Then,

$$H^i(\varphi) = \sum_{j=1}^{K^i} \sum_{k=1}^{K^{-i}} I_j^i(\varphi^i)\varphi_k^{-i} h^i(s_j^i, s_k^{-i}).$$

$$= \sum_k \varphi_k^{-i} h^i(s_l^i, s_k^{-i}), \text{ since } \varphi^i = s_l^i.$$

$$= \sum_k p^i([s_k^{-i}] \mid P^i(\omega)) h^i(s_l^i, s_k^{-i}), \text{ from (3.4)}$$

$$= \sum_k p^i([s_k^{-i}] \cap [s_l^i] \mid P^i(\omega)) h^i(s_l^i, s_k^{-i}),$$

since $P^i(\omega) = [s_l^i]$.

$$= \sum_k p^i(\{\omega \mid \mathbf{s}(\omega) = \langle s_l^i, s_k^{-i}\rangle\} \mid P^i(\omega)) h^i(s_l^i, s_k^{-i}),$$

$$= \sum_{s \in S_{il}} p^i([s] \mid P^i(\omega)) h^i(s) ,$$

$$= \sum_{s \in S} p^i([s] \mid P^i(\omega)) h^i(s), \text{ since } [s] \cap P^i(\omega) \text{ is null}$$

for all $s \notin S_{il}$.

$$= E(h^i(\mathbf{s}) \mid P^i(\omega)) .$$

(ii) $$H^i(\varphi \backslash s^i_q) = \sum_{j=1}^{K^i} \sum_{k=1}^{K^{-i}} I^i_j(s^i_q) \varphi^{-i}_k h^i(s^i_j, s^{-i}_k)$$

$$= \sum_k \varphi^{-i}_k h^i(s^i_q, s^{-i}_k) ,$$

$$= \sum_k p^i([s^{-i}_k] \mid P^i(\omega)) h^i(s^i_q, s^{-i}_k), \text{ from (3.4)}$$

$$= \sum_{s^{-i} \in S^{-i}} p^i([s^{-i}] \mid P^i(\omega)) h^i(s^i_q, s^{-i}) ,$$

$$= E(h^i(c^i_q, \mathbf{s}^{-i}) \mid P^i(\omega)) . \qquad \square$$

PROPOSITION 6.2. *Suppose* $\beta^i(\omega) = \varphi$. *Then* i *is* λ*-rational at* ω *if and only if*

(6.1) $$H^i(\varphi) \geqslant H^i(\varphi \backslash s^i_q), \quad \forall q .$$

Proof. (if) Suppose $H^i(\varphi) \geqslant H^i(\varphi \backslash s^i_q)$. Define C to be the closed sphere around φ in the metric λ with the radius $\lambda(\varphi, \varphi \backslash s^i_q)$. We claim that $c \cap |\sigma^i_q|$ is the singleton $\{\varphi \backslash s^i_q\}$. For, suppose not. Then there is $\hat{\varphi} \in C \cap |\sigma^i_q|$ such that $\hat{\varphi} \neq \varphi \backslash s^i_q$. However, then,

$$\lambda(\varphi, \hat{\varphi}) = \lambda^i(\varphi^i, \hat{\varphi}^i) + \lambda^{-i}(\varphi^{-i}, \hat{\varphi}^{-i})$$

$$= \lambda^i(\varphi^i, s^i_q) + \lambda^{-i}(\varphi^{-i}, \hat{\varphi}^{-i}), \text{ since } \hat{\varphi} \in |\sigma^i_q|$$

$$> \lambda^i(\varphi^i, \sigma^i_q), \text{ since } \hat{\varphi}^{-i} \neq \varphi^{-i} .$$

But this contradicts the supposition that $\hat{\varphi} \in C \cap |\sigma^i_q|$, since the radius of C is given by $\lambda(\varphi, \varphi \backslash s^i_q) = \lambda^i(\varphi^i, s^i_q)$, thereby establishing the claim.

Denote by r the number $H^i(\varphi\backslash s_q^i)$. Then $|\sigma_q^i| \cap C \subseteq |\pi_r^i|$, so that $\sigma_q^i \boxdot\!\!\rightarrow \pi_r^i$ is true at φ. However, $r \leqslant H^i(\varphi)$ by supposition. In other words, for all q, there is $r \leqslant H^i(\varphi)$ such that $\sigma_q^i \boxdot\!\!\rightarrow \pi_r^i$ is true at φ. This is the definition of λ-rationality at ω.

(only if) By λ-rationality at ω, for all q, there is $r \leqslant H^i(\varphi)$ such that $\sigma_q^i \boxdot\!\!\rightarrow \pi_r^i$ is true at φ. By the truth condition for $\boxdot\!\!\rightarrow$, there is a closed sphere C around φ such that $C \cap |\sigma_q^i|$ is non-empty and $C \cap |\sigma_q^i| \subseteq |\pi_r^i|$. Let $\tilde{\varphi} \in C \cap |\sigma_q^i|$. Then,

$$\begin{aligned}\lambda(\varphi, \tilde{\varphi}) &= \lambda^i(\varphi^i, \tilde{\varphi}^i) + \lambda^{-i}(\varphi^{-i}, \tilde{\varphi}^{-i}) \\ &\geqslant \lambda^i(\varphi^i, \tilde{\varphi}^i) \\ &= \lambda^i(\varphi^i, s_q^i)\end{aligned}$$

$$\begin{aligned}\lambda(\varphi, \varphi\backslash s_q^i) &= \lambda^i(\varphi^i, \sigma_q^i) + \lambda^{-i}(\varphi^{-i}, \hat{\varphi}^{-i}) \\ &= \lambda^i(\varphi^i, s_q^i)\,.\end{aligned}$$

Since the radius of C is no less than $\lambda(\varphi, \tilde{\varphi})$ and $\lambda(\varphi, \tilde{\varphi}) \geqslant \lambda(\varphi, \varphi\backslash s_q^i)$, we have $\varphi\backslash s_q^i \in C$. Also, $\varphi\backslash s_q^i \in |\sigma_q^i|$. By supposition, $C \cap |\sigma_q^i| \subseteq |\pi_r^i|$, so that $\varphi\backslash s_q^i \in |\pi_r^i|$. Hence, $H^i(\varphi\backslash s_q^i) \leqslant r$. But we know that $r \leqslant H^i(\varphi)$. Thus, $H^i(\varphi) \geqslant H^i(\varphi\backslash s_q^i)$. Moreover, this is the case for all q. □

Proof of Theorem 2.

i is λ-rational at $\omega \Leftrightarrow H^i(\varphi) \geqslant H^i(\varphi\backslash s_q^i)$, $\forall q$ (by Proposition 6.2) $\Leftrightarrow i$ is Aumann-rational at ω (by Proposition 6.1). □

REFERENCES

Aumann, R. J.: 1974, 'Subjectivity and Correlation in Randomized Strategies', *Journal of Mathematical Economics* **1**, 67–96.

Aumann, R. J.: 1987, 'Correlated Equilibrium as an Expression of Bayesian Rationality', *Econometrica* **55**, 1–18.

Jeffrey, R. C.: 1965, *The Logic of Decision*, McGraw-Hill, New York.

Jeffrey, R. C.: 1983, *The Logic of Decision*, second edition, Chicago University Press.

Lewis, D.: 1973, *Counterfactuals*, Blackwell, Oxford.

Savage, L.: 1954, *The Foundations of Statistics*, Wiley, New York.

Selten, R.: 1975, Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games', *International Journal of Game Theory* **4**, 25–55.
Stalnaker, R.: 1968, 'A Theory of Counterfactuals', in N. Rescher (ed.), *Studies in Logical Theory*, Blackwell, Oxford.

University College,
Oxford,
Oxford, OX1 4BH,
England.