ON STALNAKER'S NOTION OF STRONG RATIONALIZABILITY AND NASH EQUILIBRIUM IN PERFECT INFORMATION GAMES

Giacomo Bonanno

and

Klaus Nehring

Department of Economics,
University of California,
Davis, CA 95616-8578, USA
E-mail: gfbonanno@ucdavis.edu
kdnehring@ucdavis.edu

September 1996
Forthcoming in *Theory and Decision*

Abstract

Counterexamples to two results by Stalnaker (*Theory and Decision*, 1994) are given and a corrected version of one of the two results is proved. Stalnaker's proposed results are: (1) if at the true state of an epistemic model of a perfect information game there is common belief in the rationality of every player and common belief that no player has false beliefs (he calls this joint condition "strong rationalizability"), then the true (or actual) strategy profile is path equivalent to a Nash equilibrium; (2) in a normal-form game a strategy profile is strongly rationalizable if and only if it belongs to C^{∞}, the set of profiles that survive the iterative deletion of inferior profiles.

Stalnaker (1994, Section 6) introduced the notion of Strong Rationalizability and stated a number of results. We provide counterexamples to the two main results (Theorem 3, p. 63, and Theorem 4, p.64) and give a proof of a corrected version of one of them. First we recall some basic notation and definitions.

Let $\Gamma = \langle N, \langle C_i, u_i \rangle_{i \in N} \rangle$ be a normal-form game, where $N = \{1, \ldots, n\}$ is the set of players and, for every $i \in N$, C_i is i's set of strategies and $u_i : C \to \mathbb{R}$ (where $C = C_1 \times \ldots \times C_n$) is i's payoff function. An *epistemic model for* Γ, which we denote by M_Γ, is a tuple $\langle W, \mathbf{a}, \langle R_i, P_i, S_i \rangle_{i \in N} \rangle$ where W is a finite set of states (or possible worlds), **a** is a designated member of W, representing the true state (or actual world), and, for every player $i \in N$,

- R_i is a serial, transitive and euclidean relation (the interpretation of xR_iy is that at state x player i considers state y possible). We denote the set $\{y \in W : xR_iy\}$ by $R_i(x)$.

- $P_i : W \to \Delta(W)$ (where $\Delta(W)$ denotes the set of probability distributions on W) is a function that associates with every $w \in W$, player i's subjective probabilistic beliefs $P_{i,w} \in \Delta(W)$ at state w, satisfying the condition that $P_{i,w}(y) > 0$ if and only if $y \in R_i(w)$.

- $S_i : W \to C_i$ is a function that associates with every state a strategy for player i, satisfying the property that if $y \in R_i(x)$ then $S_i(y) = S_i(x)$. Define $S : W \to C$ and $S_{-i} : W \to C_{-i}$ (where $C_{-i} = C_1 \times \ldots \times C_{i-1} \times C_{i+1} \times \ldots \times C_n$) as follows: $S(w) = (S_1(w), \ldots, S_n(w))$ and $S_{-i}(w) = (S_1(w), \ldots, S_{i-1}(w), S_{i+1}(w), \ldots, S_n(w))$.

Let R^* denote the transitive closure of the R_i relations. For every $w \in W$, $R^*(w)$ denotes the set $\{y \in W : wR^*y\}$. A proposition (or event) $E \subseteq W$ is *commonly believed* at

$w \in W$ if and only if $R^*(w) \subseteq E$. Player *i is rational at state w* if and only if , for all $c_i \in C_i$,

$$\sum_{y \in R_i(w)} P_{i,w}(y)\, u_i(S(y)) \geq \sum_{y \in R_i(w)} P_{i,w}(y)\, u_i(c_i, S_{-i}(y))$$

Let A_i be the set of states where player i is rational and let $A = \bigcap_{i \in N} A_i$. In model $\mathtt{M}_\Gamma$

(1) there is *common belief in rationality* whenever this obtains at the true state, that is, if and only if $R^*(\mathbf{a}) \subseteq A$, and

(2) there is *common belief that no player has false beliefs* if and only if the R_i relations restricted to the set $R^*(\mathbf{a})$ are reflexive, that is, for all $w \in R^*(\mathbf{a})$ and all $i \in N$, $w \in R_i(w)$.

Stalnaker (p. 61) calls a strategy profile *strongly rationalizable* if it is the strategy profile associated with the true state in some model $\mathtt{M}_\Gamma$ where there is common belief in rationality and common belief that no player has false beliefs. A strategy profile $c \in X$ is *inferior relative to X* if there exists a player i and a (possibly mixed) strategy m_i of player i such that (1) $u_i(c) < u_i(m_i, c_{-i})$ and (2) for all $s_{-i} \in C_{-i}$ such that $(c_i, s_{-i}) \in X$, $u_i(c_i, s_{-i}) \leq u_i(m_i, s_{-i})$. For every $k \geq 0$, define $C^k \subseteq C$ and $I^k \subseteq C$ as follows: $C^0 = C$, I^k is the set of profiles that are inferior relative to C^k and $C^{k+1} = C^k \setminus I^k$. Let $C^\infty = \bigcap_{k=1}^{\infty} C^k$.

Stalnaker states the following result (the sufficiency part of Theorem 3, p.63): If $c \in C$ is strongly rationalizable, then $c \in C^\infty$. We show that this result is not correct. Let Γ be the following two-player game.

		Player 2	
		d	a
Player	D	1 , 1	1 , 1
1	A	1 , 1	0 , 0

First note that (A, a) $\notin C^{\infty}$ since it is inferior relative to C (for Player 1 A is weakly dominated by D and $u_1((A, a)) = 0 < u_1((D, a)) = 1$). We now construct a model for Γ where there is common belief in rationality and common belief that no player has false beliefs and nevertheless $S(\mathbf{a}) = (A, a)$, so that (A,a) is strongly rationalizable. Let $W = \{\mathbf{a}, b, c\}$. For $x \in \{b, c\}$, $R_1(x) = R_2(x) = \{x\}$ and $R_1(\mathbf{a}) = \{c\}$ and $R_2(\mathbf{a}) = \{b\}$. Let $S(\mathbf{a}) = (A, a)$, $S(b) = (D, a)$ and $S(c) = (A, d)$. This model is represented in Figure 1 where there is an arrow for player i from state x to state y if and only if $y \in R_i(x)$.

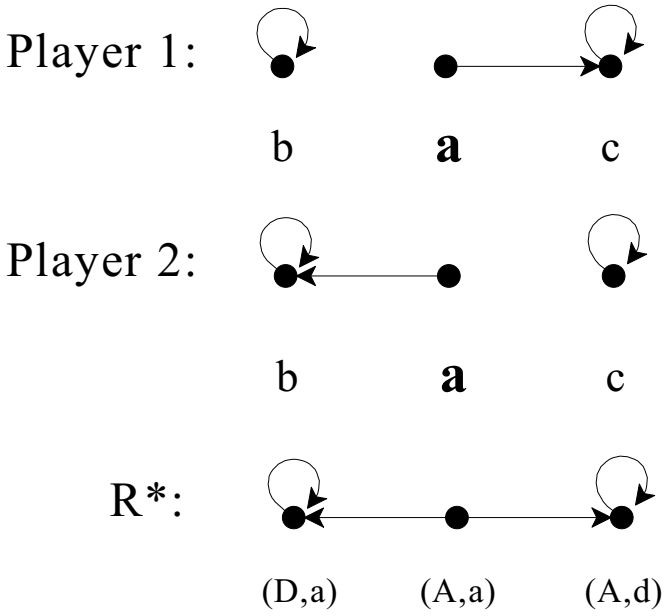

Figure 1

Note that $S_1(\mathbf{a}) = S_1(c)$, as required by the fact that $c \in R_1(\mathbf{a})$, and similarly for Player 2. It is easy to check that at every state every player is rational (indeed for $x \in \{b, c\}$ S(x) is a Nash equilibrium). Hence at **a** (indeed at every state) it is common belief that all the players are rational. Furthermore there is common belief (at **a**, indeed at every state) that no player has false beliefs.

The above is also a counterexample to Theorem 4 (p.64) which states the following: In the strategic form of any perfect information game, if a strategy profile is

strongly rationalizable then it is path equivalent to a Nash equilibrium strategy profile. To see this, consider the extensive game of Figure 2, whose normal form coincides with the game of the previous example.

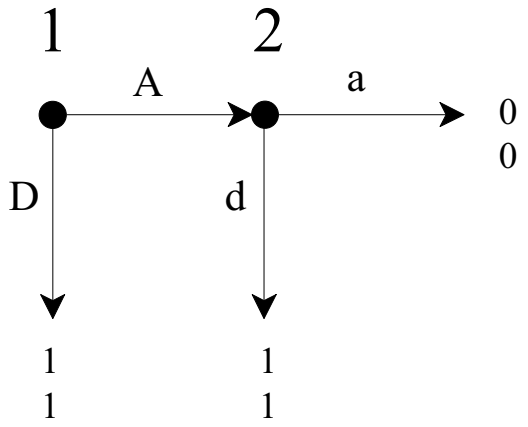

Figure 2

Since S(**a**) = (A, a) is not a Nash equilibrium and there is no other strategy profile yielding the same outcome as (A, a), we have a counterexample to the above claim.

We now state and prove a corrected version of the sufficiency part of Theorem 3.

THEOREM. Let Γ be a normal-form game and $\mathtt{M}_{\Gamma}$ a model for Γ where there is common belief (at the true state) that every player is rational and that no player has false beliefs. Then (at the true state) it is common belief that only strategy profiles in C^{∞} are chosen, that is, $R^{*}(\mathbf{a}) \subseteq \{w \in W : S(w) \in C^{\infty}\}$.

Proof. For every $w \in R^{*}(\mathbf{a})$ define j(w) as follows: $j(w) = \infty$ if $S(w) \in C^{\infty}$ and $j(w) = k \in \mathbb{N}$ (where $\mathbb{N}$ is the set of non-negative integers) if $S(w) \in C^{k}$ and $S(w) \notin C^{k+1}$. Clearly j(w) is well defined, since $S(w) \in C^{0}$ for all $w \in W$. Let $\overline{k}$ be the minimum of $\{j(w)\}_{w \in R^{*}(\mathbf{a})}$. Suppose that $R^{*}(\mathbf{a}) \not\subseteq \{w \in W : S(w) \in C^{\infty}\}$. Then $\overline{k} < \infty$. Let $\overline{w} \in R^{*}(\mathbf{a})$

be such that j($\overline{w}$) = $\overline{k}$. Then S($\overline{w}$)$\in I^{\overline{k}}$, that is, S($\overline{w}$) is inferior relative to $C^{\overline{k}}$. Thus there is a player i and a (possibly mixed) strategy m_i of player i such that:

$$u_i(m_i, s_{-i}) \geq u_i(S_i(\overline{w}), s_{-i}) \text{ for all } s_{-i} \in C_{-i} \text{ such that } (S_i(\overline{w}), s_{-i}) \in C^{\overline{k}}, \text{ and} \quad (1)$$

$$u_i(m_i, S_{-i}(\overline{w})) > u_i(S(\overline{w})). \quad (2)$$

Since $\overline{w} \in R^*$(**a**) and at **a** it is common belief that no player has false beliefs, $\overline{w} \in R_i(\overline{w})$, which implies that $P_{i,\overline{w}}(\overline{w}) > 0$. By definition of R^*, $R_i(\overline{w}) \subseteq R^*(\overline{w})$. By transitivity of R^*, since $\overline{w} \in R^*$(**a**), $R^*(\overline{w}) \subseteq R^*$(**a**). By definition of $\overline{w}$, R^*(**a**) $\subseteq \{w \in W : S(w) \in C^{\overline{k}}\}$. Hence $R_i(\overline{w}) \subseteq \{w \in W : S(w) \in C^{\overline{k}}\}$. It follows from this and (1) and (2) that

$$\sum_{y \in R_i(\overline{w})} P_{i,\overline{w}}(y)\, u_i(S(y)) < \sum_{y \in R_i(\overline{w})} P_{i,\overline{w}}(y)\, u_i(m_i, S_{-i}(y)),$$

[recall that, for all $y \in R_i(\overline{w})$, $S_i(y) = S_i(\overline{w})$] that is, player i is not rational at $\overline{w}$. Hence it cannot be common belief at **a** that player i is rational. ■

REMARK. As a corollary to the above theorem one obtains that, if at least one of the players does not have false beliefs at the true state, then S(**a**) $\in C^{\infty}$.[1]

References

Stalnaker, R. (1994), On the evaluation of solution concepts, *Theory and Decision*, 37, 49-74.

[1] A stronger version of this condition, namely that at the true state *every* player has correct beliefs, appears in a modified definition of strong rationalizability in Stalnaker (1996, p. 28; we are grateful to Philippe Mongin for bringing this paper to our attention). However, he does not comment on the validity of his results in the *Theory and Decision* paper (nor does he mention how to correct them). Indeed, later on (p. 37) he writes "We could drop the assumption that the players beliefs are all actually true, assuming not common knowledge of rationality, but only common belief in rationality and common belief that no one is in error about anything."

Stalnaker, R. (1996), Knowledge, belief and counterfactual reasoning in games, forthcoming in *Economics and Philosophy*.