Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2014-013 June 9, 2014

Possibilistic Beliefs and Higher-Level Rationality

Jing Chen, Silvio Micali, and Rafael Pass

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Possibilistic Beliefs and Higher-Level Rationality

Jing Chen
CS Dept., Stony Brook University
Stony Brook, NY 11794
jingchen@cs.stonybrook.edu

Silvio Micali
CSAIL, MIT
Cambridge, MA 02139
silvio@csail.mit.edu

Rafael Pass
Dept. of Computer Science
Cornell University
rafael@cs.cornell.edu

June 10, 2014

1 Quick Summary

We consider rationality and rationalizability for normal-form games of incomplete information in which the players have *possibilistic* beliefs about their opponents. In this setting, we prove that the strategies compatible with the players being level-k rational coincide with the strategies surviving a natural k-step iterated elimination procedure. We view the latter strategies as the (level-k) rationalizable ones in our possibilistic setting.

Rationalizability was defined by Pearce [23] and Bernheim [12] for complete-information settings. Our iterated elimination procedure is similar to that proposed by Dekel, Fudenberg, and Morris [14] in a Bayesian setting. For other iterated elimination procedures and corresponding notions of rationalizability in Bayesian settings, see Brandenburger and Dekel [9], Tan and Werlang [24], Battigalli and Siniscalchi [8], Ely and Peski [15], Weinstein and Yildiz [25], and Halpern and Pass [19].

2 The Epistemic Framework

2.1 Possibilistic Structures and Rationality Models

Given an n-player normal-form game Γ, let S_i be the set of pure actions of player i in Γ and $S = S_1 \times \cdots \times S_n$. To model the players' uncertainty about each other's utility and action in Γ, we consider a possibilistic version of Harsanyi's type structure [20].

Definition 1. *A* possibilistic structure $\mathcal{G}$ *for* Γ *is a tuple of profiles,* $\mathcal{G} = (T, u, B, \boldsymbol{s})$*, where for each player* i*,*

- T_i *is a finite set of* i*'s* types*;*

- $u_i : S \times T \to \mathbb{R}$ *is i's* utility function*;*
- $B_i : T_i \to 2^{T_{-i}}$ *is i's* belief correspondence*; and*
- $\boldsymbol{s}_i : T_i \to S_i$ *is i's* strategy function*.*

A possibilistic structure does not impose any consistency requirements among the beliefs of different players. Indeed, a player may have totally wrong beliefs about another player's beliefs. For instance, in a single-good auction, player i may believe that player j's valuation for the good is greater than 100, whereas player j may believe that player i believes that j's valuation is less than 10. Moreover, each utility function u_i has domain $S \times T$ rather than $S \times T_i$. This enables us to deal with interdependent-type settings as well.

Below we define the players' rationality, higher-level rationality and common belief of rationality, in the same way as Aumann [5].

Definition 2. *Let $\mathcal{G} = (T, u, B, \boldsymbol{s})$ be a possibilistic structure for Γ and t be a type profile in T. Player i is* rational *at t_i if for every action s'_i of i, there exists $t'_{-i} \in B_i(t_i)$ such that*

$$u_i((\boldsymbol{s}_i(t_i), \boldsymbol{s}_{-i}(t'_{-i})), (t_i, t'_{-i})) \geq u_i((s'_i, \boldsymbol{s}_{-i}(t'_{-i})), (t_i, t'_{-i})).$$

Player i is rational at t *if he is rational at t_i.*

Based on this definition we define the following events.

- Let $RAT_i = \{t \in T | \ i \text{ is rational at } t\}$ be the event that player i is rational.
- For any event $E \subseteq T$, let $\mathbf{B}_i(E) = \{t \in T | \ (t_i, t'_{-i}) \in E \ \forall t'_{-i} \in B_i(t_i)\}$ be the event that player i believes that E occurs.
- Let $RAT_i^0 = T$ be the event that player i is level-0 rational (namely, irrational), and for any $k \geq 1$, let $RAT_i^k = RAT_i \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1})$ be the event that player i is level-k rational.
  Clearly, $RAT_i^1 = RAT_i \cap \mathbf{B}_i(\cap_{j\neq i} RAT_j^0) = RAT_i \cap \mathbf{B}_i(T) = RAT_i \cap T = RAT_i$. That is, being level-1 rational is equivalent to being rational.
- For any $k \geq 0$ let $RAT^k = \cap_i RAT_i^k$ be the event that every player is level-k rational, and let $RAT = RAT^1$ be the event that every player is rational.
- For any event $E \subseteq T$, let $\mathbf{EB}^0(E) = E$, $\mathbf{EB}^1(E) = \mathbf{EB}(E) = \cap_i \mathbf{B}_i(E)$ be the event that every player believes that E occurs, and $\mathbf{EB}^k(E) = \mathbf{EB}(\mathbf{EB}^{k-1}(E))$ for any $k \geq 2$.
- Let $\mathbf{CB}(RAT) = \cap_{k\geq 0} \mathbf{EB}^k(RAT)$ be the event that the players have comment belief of rationality.

Definition 3. *For any $t \in T$ and $k \geq 0$, player i is* level-k rational at t *if $t \in RAT_i^k$. For any $t_i \in T_i$, player i is* level-k rational at t_i *if there exists $t_{-i} \in T_{-i}$ such that i is level-k rational at (t_i, t_{-i}). For any $t \in T$, the players have* common belief of rationality at t *if $t \in \mathbf{CB}(RAT)$.*

Notice that whether player i is level-k rational or not at t solely depends on t_i and player i's belief hierarchy at t_i, and does not depend on t_{-i} at all. Thus it is immediately clear that

$(*)$ *Player i is level-k rational at t_i if and only if*
for all $t_{-i} \in T_{-i}$ player i is level-k rational at (t_i, t_{-i}).

2.2 Basic Properties of Our Model

The following six properties (proved in Section 4) help understanding our model.

Property 1. *For any player i, $RAT_i = \mathbf{B}_i(RAT_i)$.*

That is, a rational player believes that he is rational.

Property 2. *For all players i and all $E \subseteq T$, $\mathbf{B}_i(E) = \mathbf{B}_i(\mathbf{B}_i(E))$.*

That is, if a player believes (the occurrence of) event E, then he believes that he believes E.

Property 3. *For any player i and any $k \geq 0$, $RAT_i^k = \mathbf{B}_i(RAT_i^k)$.*

That is, a level-k rational player believes that he is level-k rational.

Property 4. *For any player i and any $k \geq 0$, $RAT_i^{k+1} \subseteq RAT_i^k$.*

The following two properties provide alternative definitions for level-k rationality and common belief of rationality.

Property 5. *For any player i and any $k \geq 1$, $RAT_i^k = RAT_i \cap \mathbf{B}_i(\cap_j RAT_j^{k-1})$.*

That is, for $k \geq 1$, being level-k rational is equivalent to being rational and believing that *every* player is level-$(k-1)$ rational.

Property 6. $\mathbf{CB}(RAT) = \cap_{k\geq 0} \cap_{i\in[n]} RAT_i^k$.

2.3 Type Structures and Iterated Elimination of Strictly Dominated Actions

In many scenarios the players' beliefs about each other's (payoff) types are given exogenously, and they reason about each other's actions based on their beliefs about types. To model this kind of information structure and reasoning procedure we define *type structures*: a type structure $\mathcal{T}$ for Γ is a tuple of profiles, $\mathcal{T} = (T, u, B)$, where T, u, B are as defined in a possibilistic structure for Γ. Thus a type structure can be considered as a possibilistic structure with the strategy function removed.

Definition 4. *A possibilistic structure $\mathcal{G} = (T, u, B, \boldsymbol{s})$ for Γ is* consistent *with a type structure $\mathcal{T}' = (T', u', B')$ for Γ if there exists a profile of functions ψ with $\psi_i : T_i \to T_i'$ $\forall i$ such that,*

- *$\forall i$ and $\forall t \in T$, $u_i(\cdot; t) = u_i'(\cdot; \psi(t))$; and*
- *$\forall i$ and $\forall t_i \in T_i$, $\psi_{-i}(B_i(t_i)) = B_i'(\psi_i(t_i))$.*

We refer to such a ψ as a consistency mapping.

The notion of consistency captures that, introducing actions into the picture does not cause the players to change their beliefs about *types*, but causes them to form additional beliefs about *actions*.

Illustratively, both possibilistic structures and type structures can be represented by directed graphs, with nodes corresponding to the players' types and edges corresponding to their beliefs. The only difference is that in a possibilistic structure each node is also associated with an action.

Example Consider a revised version of the BoS game, where player 1 has a unique type t_1 and player 2 has two types t_2 and t_2' —whether he wants to meet or avoid player 1. The players' utilities are specified in Figure 1.

	B	S
B	2,1	0,0
S	0,0	1,2

(a) Utilities under type profile (t_1, t_2)

	B	S
B	2,0	0,2
S	0,1	1,0

(b) Utilities under type profile (t_1, t_2')

Figure 1: A revised BoS game

Figure 2a provides an elementary type structure $\mathcal{T}'$ for the revised BoS game, where player 1 believes that player 2's type can be either t_2 or t_2' and player 2 believes that player 1's (unique) type is t_1. Figure 2b provides an elementary possibilistic structure $\mathcal{G}$ consistent with $\mathcal{T}'$. Here player 1's two types t_{11} and t_{12} induce the same utility function but different actions for him, and under both types player 1 believes that player 2 will use action B under type t_2 and S under t_2'. The type structure $\mathcal{T}$ obtained from $\mathcal{G}$ by removing the actions is then illustrated in Figure 2c. It is immediate to see that the consistency mapping $\psi = (\psi_1, \psi_2)$ is such that ψ_1 maps both t_{11} and t_{12} to t_1, and ψ_2 maps t_2 to t_2 and t_2' to t_2'. Indeed, under such mapping the utilities are preserved and "the belief correspondence and ψ commute."

$$\begin{tikzcd}
t_1 \arrow[r, shift left] \arrow[dr, shift left] & t_2 \arrow[l, shift left] \\
& t_2' \arrow[ul, shift left]
\end{tikzcd}$$

(a) Type structure $\mathcal{T}'$

$$\begin{tikzcd}
(t_{11}, B) \arrow[r, shift left] \arrow[dr] & (t_2, B) \arrow[l, shift left] \\
(t_{12}, S) \arrow[r, shift left] \arrow[ur] & (t_2', S) \arrow[l, shift left]
\end{tikzcd}$$

(b) Possibilistic structure $\mathcal{G}$

$$\begin{tikzcd}
t_{11} \arrow[r, shift left] \arrow[dr] & t_2 \arrow[l, shift left] \\
t_{12} \arrow[r, shift left] \arrow[ur] & t_2' \arrow[l, shift left]
\end{tikzcd}$$

(c) Type structure $\mathcal{T}$

Figure 2: A type structure and a consistent possibilistic structure

We now define rationality for type structures.

Definition 5. *Given a type structure $\mathcal{T} = (T, u, B)$ for Γ, for any player i, type $t_i \in T_i$, action s_i and integer $k \geq 0$, s_i is* consistent with level-k rationality for t_i *if, there exists a possibilistic structure $\mathcal{G} = (T', u', B', \boldsymbol{s})$ and a type $t_i' \in T_i'$, such that $\mathcal{G}$ is consistent with $\mathcal{T}$ under a consistency mapping ψ, $\psi_i(t_i') = t_i$, $\boldsymbol{s}_i(t_i') = s_i$ and i is level-k rational at t_i'.*

Action s_i is consistent with common belief of rationality for t_i *if, there exists a possibilistic structure $\mathcal{G} = (T', u', B', \boldsymbol{s})$ and a type profile $t' \in T'$, such that $\mathcal{G}$ is consistent with $\mathcal{T}$ under a consistency mapping ψ, $\psi_i(t_i') = t_i$, $\boldsymbol{s}_i(t_i') = s_i$ and the players have common belief of rationality at t'.*

Notice that our concept of consistency with level-k rationality or common belief of rationality is called rationalizability in other studies, see [8]. Next we define an iterated elimination procedure for refining the players' actions, and use it to characterize actions that are consistent with level-k rationality or common belief of rationality.

Definition 6. *Let $\mathcal{T} = (T, u, B)$ be a type structure for Γ. For each player i, type $t_i \in T_i$ and integer $k \geq 0$, we define $NSD_i^k(t_i)$, the set of actions surviving k-round elimination of strictly dominated actions for t_i, inductively as follows:*

- *$NSD_i^0(t_i) = S_i$.*
- *For each $k \geq 1$ and each $s_i \in NSD_i^{k-1}(t_i)$, $s_i \in NSD_i^k(t_i)$ if there does not exist an alternative action $s_i' \in NSD_i^{k-1}(t_i)$ such that $\forall t_{-i} \in B_i(t_i)$ and $\forall s_{-i} \in NSD_{-i}^{k-1}(t_{-i})$,*

$$u_i((s_i', s_{-i}), (t_i, t_{-i})) > u_i((s_i, s_{-i}), (t_i, t_{-i})),$$

where $NSD_{-i}^{k-1}(t_{-i}) = \times_{j \neq i} NSD_j^{k-1}(t_j)$.

In the definition for $NSD_i^k(t_i)$, if the required action s_i' does exist, we say that s_i is strictly dominated (by s_i') for t_i over level-$(k-1)$ surviving actions. It is easy to see that defining $NSD_i^k(t_i)$ by eliminating strictly dominated actions from $NSD_i^{k-1}(t_i)$ is the same as defining it by eliminating strictly dominated actions from S_i:

- For each $k \geq 1$ and each $s_i \in S_i$, $s_i \in NSD_i^k(t_i)$ if and only if there does not exist an alternative action $s_i' \in S_i$ such that $\forall t_{-i} \in B_i(t_i)$ and $\forall s_{-i} \in NSD_{-i}^{k-1}(t_{-i})$,

$$u_i((s_i', s_{-i}), (t_i, t_{-i})) > u_i((s_i, s_{-i}), (t_i, t_{-i})).$$

Given player i's knowledge about $\mathcal{T}$, he can iteratively compute $NSD_i^k(t_i)$ for any t_i and k. Since the game Γ is finite, the elimination procedure ends at some K when no action is strictly dominated over level-$(K-1)$ surviving actions. Letting $NSD_i(t_i) = \cap_{k \geq 0} NSD_i^k(t_i)$, we have $NSD_i(t_i) = NSD_i^K(t_i) \neq \emptyset$. We say that an action s_i survives iterated elimination of strictly dominated actions for t_i if $s_i \in NSD_i(t_i)$. Following [8] we refer to $NSD_i^k(t_i)$ as the set of *level-k rationalizable* actions for t_i, and to $NSD_i(t_i)$ as the set of *rationalizable* actions for t_i.

An immediate consequence of Definition 6 is the following lemma, stated without proof.

Lemma 1. *Action $s_i \in S_i$ survives k-round elimination of strictly dominated actions for t_i if and only if there exists $B_i' \subseteq B_i(t_i)$ and $Z_{-i}(t_{-i}) \subseteq NSD_{-i}^{k-1}(t_{-i})$ for each $t_{-i} \in B_i'$, such that for each $s_i' \in S_i$ there exists $t_{-i} \in B_i'$ and $s_{-i} \in Z_{-i}(t_{-i})$ with*

$$u_i((s_i, s_{-i}), (t_i, t_{-i})) \geq u_i((s_i', s_{-i}), (t_i, t_{-i})).$$

Intuitively, s_i survives k-round elimination if, given i's belief that other players' types are among (some subset of) $B_i(t_i)$ and they use (some subset of) actions that survive $(k-1)$-round elimination, no other action *according to i's belief* can lead to higher utility than what he gets by using s_i. Lemma 1 is a possibilistic analog of Pearce's lemma [23] which, in probabilistic models, relates best responses and rationalizability to strict dominance. Note that whereas in the possibilistic case (which is what we consider) the proof is trivial, Pearce's original lemma for the probabilistic case requires additional work.

3 Characterizing Level-k Rationality and Common Belief of Rationality

Theorem 1. *Given a type structure $\mathcal{T} = (T, u, B)$ for Γ, for any player i, type t_i, action s_i and integer $k \geq 0$, s_i is consistent with level-k rationality for t_i if and only if $s_i \in NSD_i^k(t_i)$.*

Proof. We first prove the "only if" direction. Assuming s_i is consistent with level-k rationality for t_i, we prove $s_i \in NSD_i^k(t_i)$ by induction on k. For $k = 0$, the property trivially holds since $NSD_i^0(t_i) = S_i$ by definition.

For $k > 0$, by Definition 5 there exists a possibilistic structure $\mathcal{G} = (T', u', B', \boldsymbol{s})$ and a type $t'_i \in T'_i$, such that $\mathcal{G}$ is consistent with $\mathcal{T}$ under a consistency mapping ψ, $\psi_i(t'_i) = t_i$, $\boldsymbol{s}_i(t'_i) = s_i$ and i is level-k rational at t'_i.

By Definition 3 and Property $(*)$, player i being level-k rational at t'_i implies: (a) i is rational at t'_i; and (b) for each type subprofile $t'_{-i} \in B'_i(t'_i)$ we have $(t'_i, t'_{-i}) \in \cap_{j\neq i} RAT_j^{k-1}$. According to (a) and Definition 2, for each action $s'_i \in S_i$ there exists $t'_{-i} \in B'_i(t'_i)$ such that

$$u'_i((s_i, \boldsymbol{s}_{-i}(t'_{-i})), (t'_i, t'_{-i})) \geq u'_i((s'_i, \boldsymbol{s}_{-i}(t'_{-i})), (t'_i, t'_{-i})). \tag{1}$$

According to (b), for each $t'_{-i} \in B'_i(t'_i)$ and each $j \neq i$, player j is level-$(k-1)$ rational at t'_j. By Definition 5, $\boldsymbol{s}_j(t'_j)$ is consistent with level-$(k-1)$ rationality for $\psi_j(t'_j)$ and thus, by the induction hypothesis,

$$\boldsymbol{s}_j(t'_j) \in NSD_j^{k-1}(\psi_j(t'_j)). \tag{2}$$

For each $t_{-i} \in B_i(t_i)$, let $Z_{-i}(t_{-i}) = \boldsymbol{s}_{-i}(\psi_{-i}^{-1}(t_{-i}))$. Because $\psi_{-i}(B'_i(t'_i)) = B_i(t_i)$, $Z_{-i}(t_{-i}) \neq \emptyset$. By Equation 2,

$$Z_{-i}(t_{-i}) \subseteq NSD_{-i}^{k-1}(t_{-i}).$$

For each $s'_i \in S_i$, let $t'_{-i} \in B'_i(t'_i)$ be such that Equation 1 holds, $t_{-i} = \psi_{-i}(t'_{-i})$ and $s_{-i} = \boldsymbol{s}_{-i}(t'_{-i})$. Accordingly, $s_{-i} \in Z_{-i}(t_{-i})$. Since $u_i(\cdot; (t_i, t_{-i})) = u'_i(\cdot; (t'_i, t'_{-i}))$, Equation 1 implies

$$u_i((s_i, s_{-i}), (t_i, t_{-i})) \geq u_i((s'_i, s_{-i}), (t_i, t_{-i})).$$

By Lemma 1 we have $s_i \in NSD_i^k(t_i)$, concluding the proof of the "only if" direction.

Now we prove the "if" direction. By definition, proving this direction is equivalent to proving that, if $s_i \in NSD_i^k(t_i)$ then there exists a possibilistic structure $\mathcal{G} = (T', u', B', \boldsymbol{s})$ for Γ and a type $t'_i \in T'_i$ such that, $\mathcal{G}$ is consistent with $\mathcal{T}$ under a consistency mapping ψ, $\psi_i(t'_i) = t_i$, $\boldsymbol{s}_i(t'_i) = s_i$ and i is level-k rational at t'_i. Notice that $\mathcal{G}$, t'_i and ψ may depend on k, i, t_i and s_i.

In fact, we shall prove a stronger statement. Namely, for each k, there exists a *universal* possibilistic structure $\mathcal{G} = (T', u', B', \boldsymbol{s})$ for Γ, consistent with $\mathcal{T}$ under a consistency mapping ψ, such that for *every* player i, type $t_i \in T_i$, action s_i and non-negative integer $k' \leq k$,

if $s_i \in NSD_i^{k'}(t_i)$ then there exists a type $t'_i \in T'_i$ such that

$$\psi_i(t'_i) = t_i, \quad \boldsymbol{s}_i(t'_i) = s_i \quad \text{and} \quad i \text{ is level-}k' \text{ rational at } t'_i, \tag{3}$$

which implies that s_i is consistent with level-k' rationality for t'_i.

We define $\mathcal{G}$ as follows: for each player i,

- $T'_i = \big\{(t_i, k', s_i) : t_i \in T_i, k' \in \{0, \ldots, k\}, s_i \in NSD_i^{k'}(t_i)\big\}$;
- for each type profile $t' \in T'$, letting $t \in T$ be the type profile obtained by projecting each t'_j to its first component, $u'_i(\cdot; t') = u_i(\cdot; t)$;
- for each type $t'_i = (t_i, k', s_i)$, $\boldsymbol{s}_i(t'_i) = s_i$; and

- for each type $t'_i = (t_i, k', s_i)$ and type profile $t'_{-i} \in T'_{-i}$, $t'_{-i} \in B'_i(t'_i)$ if and only if there exists $t_{-i} \in B_i(t_i)$ and $s_{-i} \in NSD_{-i}^{\max\{k'-1,0\}}(t_{-i})$ such that $t'_j = (t_j, \max\{k'-1, 0\}, s_j)$ for all $j \neq i$.

It is easy to check that $\mathcal{G}$ is consistent with $\mathcal{T}$ under the consistency mapping ψ where $\psi_i(t_i, k', s_i) = t_i$ for each player i and type $(t_i, k', s_i) \in T'_i$.

We now prove by induction on k' that for any $i, t_i \in T_i$ and $s_i \in NSD_i^{k'}(t_i)$, player i is level-k' rational at $t'_i = (t_i, k', s_i)$. For $k' = 0$, since $RAT_i^0 = T$ by definition, it trivially holds that player i is level-0 rational at t'_i.

For $k' > 0$, for any $t'_{-i} = (t_j, k'-1, s_j)_{j \neq i} \in B'_i(t'_i)$, by construction we have $t_{-i} \in B_i(t_i)$ and $s_{-i} \in NSD_{-i}^{k'-1}(t_{-i})$. By the hypothesis induction, for any player $j \neq i$, j is level-$(k'-1)$ rational at t'_j and thus at (t'_i, t'_{-i}). Therefore

$$(t'_i, t'_{-i}) \in \cap_{j \neq i} RAT_j^{k'-1}.$$

Since this is true for any $t'_{-i} \in B'_i(t'_i)$, we have

$$(t'_i, t'_{-i}) \in \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k'-1})$$

for any $t'_{-i} \in B'_i(t'_i)$, as again whether player i believes some event or not only depends on t'_i and not t'_{-i}.

Since $s_i \in NSD_i^{k'}(t_i)$, by definition for any $s'_i \in NSD_i^{k'-1}(t_i)$, there exists $t_{-i} \in B_i(t_i)$ and $s_{-i} \in NSD_{-i}^{k'-1}(t_{-i})$ such that $u_i((s_i, s_{-i}), (t_i, t_{-i})) \geq u_i((s'_i, s_{-i}), (t_i, t_{-i}))$. Since any strategy s'_i that does not survive $(k'-1)$-round elimination for t_i is strictly dominated by some action in $NSD_i^{k'-1}(t_i)$ for t_i over level-$(k'-1)$ surviving actions, we further have that for any $s'_i \in S_i$, there exists $t_{-i} \in B_i(t_i)$ and $s_{-i} \in NSD_{-i}^{k'-1}(t_{-i})$ such that

$$u_i((s_i, s_{-i}), (t_i, t_{-i})) \geq u_i((s'_i, s_{-i}), (t_i, t_{-i})).$$

Letting $t'_{-i} = (t_j, k'-1, s_j)_{j \neq i}$, we have $t'_{-i} \in B'_i(t'_i)$, $\psi(t'_i, t'_{-i}) = (t_i, t_{-i})$, $\boldsymbol{s}_i(t'_i) = s_i$ and $\boldsymbol{s}_{-i}(t'_{-i}) = s_{-i}$. Thus

$$u'_i((\boldsymbol{s}_i(t'_i), \boldsymbol{s}_{-i}(t'_{-i})), (t'_i, t'_{-i})) \geq u'_i((s'_i, \boldsymbol{s}_{-i}(t'_{-i})), (t'_i, t'_{-i})).$$

Accordingly, player i is rational at t'_i and $(t'_i, t'_{-i}) \in RAT_i$ for any $t'_{-i} \in B'_i(t'_i)$. By definition, $(t'_i, t'_{-i}) \in RAT_i \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k'-1})$ for any $t'_{-i} \in B'_i(t'_i)$, and thus i is level-k' rational at t'_i. This concludes the induction step and the proof of Statement (3). Therefore the "if" direction holds, concluding the proof of Theorem 1. ■

Similarly, we characterize common belief of rationality in our model by the following theorem.

Theorem 2. *Given a type structure $\mathcal{T} = (T, u, B)$ for Γ, for any player i, type t_i and action s_i, s_i is consistent with common belief of rationality for t_i if and only if $s_i \in NSD_i(t_i)$.*

The proof of Theorem 2 uses Property 6, but is otherwise similar to that of Theorem 1 and is omitted.

4 Proofs of the Basic Properties of Our Model

Property 1. *For any player i, $RAT_i = \mathbf{B}_i(RAT_i)$.*

Proof. By definition, for any $t \in RAT_i$, player i is rational at t_i. Thus for any $t'_{-i} \in B_i(t_i)$, i is rational at (t_i, t'_{-i}), implying $t \in \mathbf{B}_i(RAT_i)$.

On the other hand, for any $t \in \mathbf{B}_i(RAT_i)$, for any $t'_{-i} \in B_i(t_i)$, i is rational at (t_i, t'_{-i}), which implies that i is rational at t_i. Thus i is rational at t, namely, $t \in RAT_i$. ■

Property 2. *For any player i and any $E \subseteq T$, $\mathbf{B}_i(E) = \mathbf{B}_i(\mathbf{B}_i(E))$.*

Proof. By definition, for any $t \in \mathbf{B}_i(E)$, for any $t'_{-i} \in B_i(t_i)$, we have $(t_i, t'_{-i}) \in E$. Thus for any $t''_{-i} \in B_i(t_i)$, (t_i, t''_{-i}) is such that for any $t'_{-i} \in B_i(t_i)$, $(t_i, t'_{-i}) \in E$. Accordingly, $(t_i, t''_{-i}) \in \mathbf{B}_i(E)$, which implies $(t_i, t_{-i}) \in \mathbf{B}_i(\mathbf{B}_i(E))$.

On the other hand, for any $t \in \mathbf{B}_i(\mathbf{B}_i(E))$, for any $t'_{-i} \in B_i(t_i)$, we have $(t_i, t'_{-i}) \in \mathbf{B}_i(E)$, which implies that for any $t''_{-i} \in B_i(t_i)$, $(t_i, t''_{-i}) \in E$. Accordingly, $(t_i, t_{-i}) \in \mathbf{B}_i(E)$. ■

Property 3. *For any player i and any $k \geq 0$, $RAT_i^k = \mathbf{B}_i(RAT_i^k)$.*

Proof. For $k = 0$, $RAT_i^0 = T$ and $\mathbf{B}_i(RAT_i^0) = \mathbf{B}_i(T) = T$, as desired. For $k = 1$, since $RAT_i^1 = RAT_i$, the desired equality follows from Property 1.

For any $k \geq 2$,

$$\begin{aligned} RAT_i^k &= RAT_i \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1}) = \mathbf{B}_i(RAT_i) \cap \mathbf{B}_i(\mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1})) \\ &= \mathbf{B}_i(RAT_i \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1})) = \mathbf{B}_i(RAT_i^k) \end{aligned}$$

as desired, where the first equality is by definition and the second is by Properties 1 and 2. ■

Property 4. *For any player i and any $k \geq 0$, $RAT_i^{k+1} \subseteq RAT_i^k$.*

Proof. By induction on k. For $k = 0$, $RAT_i^1 \subseteq T = RAT_i^0$. For $k > 0$, by the induction hypothesis we have $RAT_j^k \subseteq RAT_j^{k-1}$ for each j, thus $\mathbf{B}_i(\cap_{j \neq i} RAT_j^k) \subseteq \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1})$. Accordingly, $RAT_i^{k+1} = RAT_i \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^k) \subseteq RAT_i \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1}) = RAT_i^k$, as desired. ■

Property 5. *For any player i and any $k \geq 1$, $RAT_i^k = RAT_i \cap \mathbf{B}_i(\cap_j RAT_j^{k-1})$.*

Proof. For $k = 1$ we have $RAT_i^1 = RAT_i = RAT_i \cap T = RAT_i \cap \mathbf{B}_i(T) = RAT_i \cap \mathbf{B}_i(\cap_j RAT_j^0)$, as desired. For $k \geq 2$, by the properties above we have

$$\begin{aligned} RAT_i^k &= RAT_i \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1}) = RAT_i \cap \mathbf{B}_i(\cap_{j \neq i}(RAT_j^{k-1} \cap RAT_j^{k-2})) \\ &= RAT_i \cap \mathbf{B}_i((\cap_{j \neq i} RAT_j^{k-2}) \cap (\cap_{j \neq i} RAT_j^{k-1})) \\ &= RAT_i \cap \big(RAT_i \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-2})\big) \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1}) \\ &= RAT_i \cap RAT_i^{k-1} \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1}) = RAT_i \cap \mathbf{B}_i(RAT_i^{k-1}) \cap \mathbf{B}_i(\cap_{j \neq i} RAT_j^{k-1}) \\ &= RAT_i \cap \mathbf{B}_i(RAT_i^{k-1} \cap (\cap_{j \neq i} RAT_j^{k-1})) = RAT_i \cap \mathbf{B}_i(\cap_j RAT_j^{k-1}), \end{aligned}$$

where the second equality is by Property 4 and the sixth is by Property 3. ■

Property 6. $\mathbf{CB}(RAT) = \cap_{k \geq 0} \cap_{i \in [n]} RAT_i^k$.

Proof. We show by induction that for any $k \geq 1$, $\cap_i RAT_i^k = \mathbf{EB}^{k-1}(RAT)$. For $k = 1$, $\cap_i RAT_i^1 = RAT^1 = RAT = \mathbf{EB}^0(RAT)$ as desired. For $k > 1$,

$$\begin{aligned}
\cap_i RAT_i^k &= \cap_i \left(RAT_i \cap \mathbf{B}_i(\cap_j RAT_j^{k-1})\right) = \cap_i \left(\mathbf{B}_i(RAT_i) \cap \mathbf{B}_i(\cap_j RAT_j^{k-1})\right) \\
&= \cap_i \left(\mathbf{B}_i((RAT_i^1 \cap RAT_i^{k-1}) \cap (\cap_{j \neq i} RAT_j^{k-1}))\right) \\
&= \cap_i \mathbf{B}_i(RAT_i^{k-1} \cap (\cap_{j \neq i} RAT_j^{k-1})) = \cap_i \mathbf{B}_i(\cap_j RAT_j^{k-1}) \\
&= \mathbf{EB}(\cap_j RAT_j^{k-1}) = \mathbf{EB}(\mathbf{EB}^{k-2}(RAT)) = \mathbf{EB}^{k-1}(RAT),
\end{aligned}$$

where the first equality is by Property 5, the second by Property 1, the fourth by Property 4, and the seventh by the induction hypothesis. Since $\cap_i RAT_i^0 = T$, we have

$$\cap_{k \geq 0} \cap_i RAT_i^k = \cap_{k \geq 1} \cap_i RAT_i^k = \cap_{k \geq 0} \mathbf{EB}^k(RAT) = \mathbf{CB}(RAT),$$

as desired. ■

Acknowledgements

The first two authors thank Shafi Goldwasser, Andrew Lo, and Ron Rivest for their comments. The third author wishes to thank Joseph Halpern for introducing him to the area of epistemic game theory, and for hours and hours of enlightening discussions about it. This work has been partially supported by ONR Grant No. N00014-09-1-0597.

References

[1] D. Abreu and H. Matsushima. Virtual Implementation in Iteratively Undominated Actions: Complete Information. *Econometrica*, Vol. 60, No. 5, pp. 993-1008, 1992.

[2] M. Allais. Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica*, Vol. 21, No. 4, pp. 503-546, 1953.

[3] G. B. Asheim, M. Voorneveld, and J. W. Weibull. Epistemically stable action sets. Working paper, 2009.

[4] R. Aumann. Agreeing to Disagree. *Annals of Statistics 4*, pp. 1236-1239, 1976.

[5] R. Aumann. Backwards Induction and Common Knowledge of Rationality. *Games and Economic Behavior*, Vol. 8, pp. 6-19, 1995.

[6] R. Aumann and A. Brandenburger. Epistemic Conditions for Nash Equilibrium. *Econometrica*, Vol. 63, No. 5, pp. 1161-1180, 1995.

[7] K. Basu and J.W. Weibull. Action subsets closed under rational behavior. *Economics Letters*, Vol. 36, pp. 141-146, 1991.

[8] P. Battigali and M. Siniscalchi. Rationalization and Incomplete Information. The B.E. Journal of Theoretical Economics, Volume 3, Issue 1, Article 3, 2003.

[9] A. Brandenburger and E. Dekel. Rationalizability and correlated equilibria. *Econometrica*. Vol. 55, pp. 1391-1402, 1987.

[10] D. Bergemann and S. Morris. Robust mechanism design. *Econometrica*, Vol. 73, No. 6, pp. 1771-1813, 2005.

[11] D. Bergemann and S. Morris. Robust Mechanism Design: An Introduction. In D. Bergemann and S. Morris, Robust Mechanism Design, World Scientific Press, 2012.

[12] B. Bernheim. Rationalizable Strategic Behavior. *Econometrica*, Vol. 52, No. 4, pp. 1007-1028, 1984.

[13] J. Chen and S. Micali. Mechanism Design with Set-Theoretic Beliefs. *Symposium on Foundations of Computer Science (FOCS)*, pp. 87-96, 2011.

[14] E. Dekel, D. Fudenberg, S. Morris. Interim correlated rationalizability. *Theoretical Economics*, Vol. 2, pp. 15-40, 2007.

[15] J. C. Ely and M. Peski. Hierarchies of belief and interim rationalizability. *Theoretical Economics*, Vol. 1, pp. 19-65, 2006.

[16] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. Reasoning About Knowledge. MIT Press, 2003.

[17] J. Glazer and M. Perry. Virtual Implementation in Backwards Induction. *Games and Economic Behavior*, Vol.15, pp. 27-32, 1996.

[18] J. Halpern and R. Pass. A Logical Characterization of Iterated Admissibility. *Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pp. 146-155, 2009.

[19] J. Halpern and R. Pass. Conservative belief and rationality. *Games and Economic Behavior*, Vol. 80, pp. 186-192, 2013.

[20] J. Harsanyi. Games with Incomplete Information Played by "Bayesian" Players, I-III. Part I. The Basic Model. *Management Science*, 14(3) Theory Series: 159-182, 1967.

[21] M. Jackson. Implementation in Undominated Actions: A Look at Bounded Mechanisms. *The Review of Economic Studies*, 59(4): 757-775, 1992.

[22] S. Kripke. Semantical analysis of modal logic I: normal modal propositional calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, Vol. 9, pp. 67-96, 1963.

[23] D. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, Vol. 52, No. 4, pp. 1029-1050, 1984.

[24] T. Tan and S. Werlang. The Bayesian foundation of solution concepts of games. *Journal of Economic Theory*, Vol 45, pp. 370-391, 1988.

[25] J. Weinstein and M. Yildiz. A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements. *Econometrica*, 75(2), pp. 365-400, 2007.