From Lawvere to Brandenburger-Keisler: interactive forms of diagonalization and self-reference

Samson Abramsky and Jonathan Zvesper

Department of Computer Science, University of Oxford

1 Introduction

Diagonal arguments lie at the root of many fundamental phenomena in the foundations of logic and mathematics. Recently, a striking form of diagonal argument has appeared in the foundations of epistemic game theory, in a paper by Adam Brandenburger and H. Jerome Keisler [11]. The core Brandenburger-Keisler result can be seen, as they observe, as a two-person or interactive version of Russell's Paradox. This raises a number of fascinating questions at the interface of epistemic game theory, logic and theoretical computer science:

1. Is the Brandenburger-Keisler argument (henceforth: 'BK argument') just one example of a more general phenomenon, whereby mathematical structures and arguments can be generalized from a familiar 'one-person' form to a two- or multi-agent interactive form?
2. To address this question, a sharper understanding of the BK argument is needed. The argument hinges on a statement involving the modalities **believes** and **assumes**. The statement has the form

   Ann **believes** that Bob **assumes** that ...

   which is not familiar as it stands. Where does this **believes-assumes** pattern come from? How exactly does it relate to the more familiar arguments in the one-person case? In particular, can it be reduced to a one-person argument?
3. Is there a natural multi-agent generalization of the BK argument? In particular, does it have a *compositional structure*, which allows a smooth generalization to any number of agents?
4. The main formal consequence of the BK argument is that there can be no belief model which is 'assumption-complete' with respect to a collection of predicates including those definable in the first-order language of the model. Brandenburger and Keisler also give a positive result, a construction of a topological model which is assumption-complete with respect to the positive fragment of first-order logic extended with the **believes** and **assumes** modalities. They raise the question of a more general perspective on the availability of such models.

We shall provide substantial answers to questions (2)–(4) above in the present paper. These results also suggest that the Brandenburger-Keisler 'paradox' does offer a good point of entry for considering the more general question (1).

The starting point for our approach is a classic paper by F. William Lawvere from 1969 [17], in which he gave a simple form of the (one-person) diagonal argument as a *fixpoint lemma* in a very general setting. This lemma lies at the basis of a remarkable range of results. Lawvere's ideas were amplified and given a very attractive presentation in a recent paper by Noson Yanofsky [25].

Our contributions can be summarized as follows:

- We reformulate the core BK argument as a *fixpoint lemma*. This immediately puts it in the general genre of diagonal arguments, and in particular of the Lawvere fixpoint lemma.
- The BK argument applies to (belief) *relations*, while the Lawvere argument applies to *functions* (actually, abstractly to arrows in a category). To put them on common ground, we give a novel relational reformulation of the Lawvere argument.
- We analyze the exact logical resources required for our fixpoint version of the BK argument, and show that it can be carried out in *regular logic*, the fragment of first-order logic comprising sequents $\phi \vdash \psi$, where ϕ and ψ are built from atomic formulas by conjunction and existential quantification. Regular logic can be interpreted in any *regular category*, which covers a wide range of types of mathematical structure. The Lawvere argument can also be carried out in (a fragment of) regular logic. We can now recognize the Lawvere argument as exactly the one-person version of the BK argument, and interpret the key BK lemma as a reduction to the one-person Lawvere argument.
- This analysis leads in turn to a smooth generalization of the BK argument to multi-agent belief models. The content of the **believes-assumes** pattern, or more generally the **believes*****-assumes** pattern:

  $$A_1 \textbf{ believes that } A_2 \textbf{ believes that } \ldots A_n \textbf{ believes that } B \textbf{ assumes that}$$
  $$\ldots$$

  is that the Lawvere hypothesis of *weak point surjectivity* is propagated back along *belief chains*.
- We furthermore give a *compositional analysis* of the **believes-assumes** pattern, which *characterizes* what we call 'belief-complete' relations in terms of this propagation property. This gives a rather definitive analysis for why the BK argument takes the form it does.
- We then turn to the issue of the construction of assumption complete models. The categorical perspective allows us to apply general techniques from coalgebra and domain theory to the construction of such models.

The further contents of this paper are as follows. In Section 2, we review the setting for the BK argument, and give our formulation of it as a fixpoint lemma. In Section 3, we show how it can be formalized in regular logic. In Section 4, we introduce the Lawvere fixpoint lemma. In Section 5 we bring BK and

Lawvere together, giving a relational reformulation of the Lawvere lemma, and showing how to reduce BK to this version of Lawvere, *i.e.* the two-person to the one-person argument. In Section 6, we give the multi-agent generalization, and in Section 7 the compositional analysis of belief-completeness, and hence of the **believes-assumes** pattern. In Section 8, we show how general functorial methods lead to the construction of assumption-complete models. Section 9 concludes with some further directions.

2 The Brandenburger-Keisler Argument

A (two-person) *belief structure* has the form (U_a, U_b, R_a, R_b) where

$$R_a \subseteq U_a \times U_b, \qquad R_b \subseteq U_b \times U_a.$$

In the context of epistemic game theory, we think of U_a and U_b as *type spaces* for Alice and Bob:

- Elements of U_a represent possible epistemic states of Alice in which she holds beliefs about Bob, Bob's beliefs, etc. Symmetrically, elements of U_b represent possible epistemic states of Bob.
- The relations $R_a \subseteq U_a \times U_b$, $R_b \subseteq U_b \times U_a$ specify these beliefs. Thus $R_a(x, y)$ expresses that in state x, Alice believes that state y is possible for Bob.
- We say that a state $x \in U_a$ *believes* $P \subseteq U_b$ if $R_a(x) \subseteq P$. Modal logic provides a useful perspective on these notions, as shown by Eric Pacuit [20] (see also [11]). Modally, 'x believes P' is just $x \models \Box_a P$ where $\Box_a$ is the usual necessity operator defined with respect to the relation R_a:

$$x \models \Box_a \phi \;\equiv\; \forall y.\, R_a(x, y) \Rightarrow\; y \models \phi.$$

- We say that x *assumes* P if $R_a(x) = P$. This is $x \models \boxplus_a P$, where $\boxplus_a$ is the modality defined by

$$x \models \boxplus_a \phi \;\equiv\; \forall y.\, R_a(x, y) \Leftrightarrow\; y \models \phi.$$

A belief structure (U_a, U_b, R_a, R_b) is *assumption-complete* [11] with respect to a collection of predicates on U_a and U_b if for every predicate P on U_b in the collection, there is a state x on U_a such that x assumes P; and similarly for the predicates on U_a. (A *predicate* on a set U is just a subset of U.)[1]

Brandenburger and Keisler show in [11] that this hypothesis, in the case where the predicates include those definable in the first-order language of this structure, leads to a contradiction. (They also show the existence of assumption complete models for some other cases.)

Our aim is to understand the general structures underlying this argument. Our first step is to recast their result as a *positive* one — a fixpoint lemma.

[1] Related forms of completeness assumption are used in the analysis of various solution concepts in games in [9, 10].

2.1 The BK Fixpoint Lemma

We are given a belief structure (U_a, U_b, R_a, R_b). We assume that for 'all' (in some 'definable' class of) predicates p on U_a there is x_0 such that:

$$R_a(x_0) \subseteq \{y \mid R_b(y) = \{x \mid p(x)\}\}. \tag{1}$$

$$\exists y.\, R_a(x_0, y). \tag{2}$$

Modally, these assumptions can be expressed as follows:

$$x_0 \models \Box_a \boxplus_b p \ \wedge \ \Diamond_a \top.$$

Remark We can read (1) as saying: 'x_0 **believes** that (y **assumes** that p)', in the terminology of Brandenburger and Keisler.

Lemma 1 (Basic Lemma). *From (1) and (2) we have:*

$$p(x_0) \iff \exists y.[R_a(x_0, y) \ \wedge \ R_b(y, x_0)].$$

Proof Suppose $p(x_0)$. Take y as in (2), so $R_a(x_0, y)$. Then by (1), $R_b(y, x_0)$. Now consider y satisfying $R_a(x_0, y) \wedge R_b(y, x_0)$. By (1), from $R_a(x_0, y)$ we have that $R_b(y) = \{x \mid p(x)\}$. Hence from $R_b(y, x_0)$ we have that $p(x_0)$. □

Lemma 2 (BK Fixpoint Lemma). *Under our assumptions, every unary propositional operator O has a fixpoint.*

Proof Since p was arbitrary, we can define

$$q(x) \equiv \exists y.[R_a(x, y) \ \wedge \ R_b(y, x)] \tag{3}$$

$$p(x) \equiv O(q(x)). \tag{4}$$

(N.B. It is important that p is defined *without reference to* x_0 to avoid circularity.) These definitions combined with the equivalence given by the Basic Lemma immediately yield:

$$O(q(x_0)) \overset{(4)}{\equiv} p(x_0) \iff \exists y.[R_a(x_0, y) \ \wedge \ R_b(y, x_0) \overset{(3)}{\equiv} q(x_0),$$

so $q(x_0)$ is a fixpoint for the operator O, as required. □

Remarks Taking $O \equiv \neg$ yields the BK 'paradox'. (In fact $\neg q(x)$ is equivalent to their 'diagonal formula' D in [11]).

In general, since our assumptions (1) and (2) are relative to a class of predicates, this argument relies on $q(x)$ and $p(x)$ being in this class. Note that $q(x)$ only involves conjunction and existential quantification. This leads to our analysis of the logical resources needed to carry out the BK argument.

3 Formalizing BK in Regular Logic

We recall that *regular logic* is the fragment of (many-sorted) first-order logic comprising sequents of the form

$$\phi \vdash_X \psi$$

where ϕ and ψ are built from atomic formulas by conjunction (including the empty conjunction $\top$) and existential quantification; and X is a finite set of variables which includes all those occurring free in ϕ and ψ. The intended meaning of such a sequent is

$$\forall x_1 \cdots \forall x_n [\phi \Rightarrow \psi]$$

where $X = \{x_1, \ldots, x_n\}$. This is a common fragment of intuitionistic and classical logic. It plays a core rôle in categorical logic. A convenient summary of regular logic can be found in [12].

We shall write $\vdash_X \psi$ for the sequent $\top \vdash_X \psi$, and $\phi \vdash \psi$ for $\phi \vdash_\varnothing \psi$.

We shall assume a logical vocabulary containing the sorts U_a and U_b, and binary relation symbols $R_a : U_a \times U_b$ and $R_b : U_b \times U_a$, together with a constant $c : U_a$ which will correspond to x_0 in the informal argument given in the previous section. Thus c is associated with the given predicate p, which will be represented by a formula in one free variable of sort U_a.

The assumptions given in the informal argument can be expressed as regular sequents as follows.

$$\begin{array}{l} (A1)\ R_a(c,y)\ \wedge\ R_b(y,x) \vdash_{\{x,y\}} p(x) \\ (A2)\ R_a(c,y)\ \wedge\ p(x) \vdash_{\{x,y\}} R_b(y,x) \\ (A3) \vdash \exists y.\, R_a(c,y) \end{array}$$

Here (A1) and (A2) correspond to assumption (1) in the informal argument, while (A3) corresponds to assumption (2).

The formal version of Lemma 1 is as follows:

Lemma 3. *From (A1)–(A3) we can infer the following sequents:*

$$\begin{array}{l} (F1)\ p(c) \vdash q(c) \\ (F2)\ q(c) \vdash p(c) \end{array}$$

where

$$q(x)\ \equiv\ \exists y.[R_a(x,y)\ \wedge\ R_b(y,x)].$$

A *definable unary propositional operator* will be represented by a formula context $O[\cdot]$, which is a closed formula built from atomic formulas, plus a 'hole' $[\cdot]$. We obtain a formula $O[\phi]$ by replacing every occurrence of the hole by a formula ϕ.

The formal version of the Fixpoint Lemma is now stated as follows:

Lemma 4. *Under the assumptions (A1)–(A3), every definable unary propositional operator $O[\cdot]$ has a fixpoint, i.e. a sentence S such that*

$$S \vdash O[S], \qquad O[S] \vdash S.$$

This is obtained directly from the previous lemma, taking $p(x) \equiv O[q(x)]$. The required sentence S is then $q(c)$.

Remarks

- Regular logic can be interpreted in any *regular category* [23, 12]: well-powered with finite limits and images, which are stable under pullbacks.These are exactly the categories which support a good calculus of relations.
- The BK fixpoint lemma is valid in any such category. Regular categories are abundant — they include all (pre)toposes, all abelian categories, all equational varieties of algebras, compact Hausdorff spaces, and categories of Q-sets for right quantales Q.
- If the propositional operator O is fixpoint-free, the result must be read contrapositively, as showing that the assumptions (A1)–(A3) lead to a contradiction. This will of course be the case if $O = \neg[\cdot]$ in either classical or intuitionistic logic. This yields exactly the BK argument.
- In other contexts, this need not be the case. For example if the propositions (in categorical terms, the subobjects of the terminal object) form a complete lattice, and O is *monotone*, then by the Tarski-Knaster theorem there will indeed be a fixpoint. This offers a general setting for understanding why *positive logics*, in which all definable propositional operators are monotone, allow the paradoxes to be circumvented.

4 The Lawvere Fixpoint Lemma

We start off concretely working in **Set**. Suppose we have a function

$$g : X \to \mathcal{V}^X$$

or equivalently, by cartesian closure:

$$\hat{g} : X \times X \to \mathcal{V}$$

Think of $\mathcal{V}$ as a set of 'truth values': $\mathcal{V}^X$ is the set of '$\mathcal{V}$-valued predicates'. Then g is showing how predicates on X can be represented by elements of X. In terms of $\hat{g}$: a predicate $p : X \to \mathcal{V}$ is representable by $x \in X$ if for all $y \in X$:

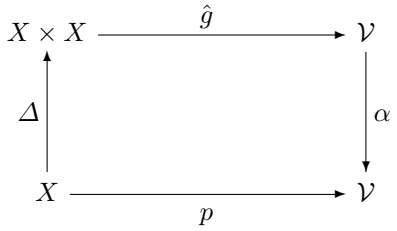
$$p(y) = \hat{g}(x, y)$$

Note that, if predicates 'talk about' X, then representable predicates allow X to 'talk about itself'.

If g is *surjective*, then *every* predicate on X is representable in X. When can this happen?

Proposition 1 (Lawvere Fixpoint Lemma). *Suppose that $g : X \to \mathcal{V}^X$ is surjective. Then* every *function $\alpha : \mathcal{V} \to \mathcal{V}$ has a* fixpoint*: $v \in \mathcal{V}$ such that $\alpha(v) = v$.*

Proof Define a predicate p by



There is $x \in X$ which represents p: then

$$p(x) = \alpha(\hat{g}(\Delta(x))) = \alpha(\hat{g}(x,x)) = \alpha(p(x))$$

so $p(x)$ is a fixpoint of α. □

Remarks on the proof Note firstly that the proof is constructive. The crucial idea is that it uses *two descriptions of p* — one from its definition, one from its representation via $\hat{g}$. And since x represents p, $p(x)$ is (indirect) *self-application.*

But does this make sense? Say that X has the *fixpoint property* if every endofunction on X has a fixpoint. Of course, *no set with more than one element has the fixpoint property!*

Basic example: $\mathbf{2} = \{0,1\}$. The negation

$$\neg 0 = 1, \qquad \neg 1 = 0$$

does not have a fixpoint. So the meaning of the theorem in **Set** must be taken *contrapositively*:

For all sets X, $\mathcal{V}$ where $\mathcal{V}$ has more than one element, there is no surjective map

$$X \to \mathcal{V}^X$$

Two Applications

Cantor's Theorem Take $\mathcal{V} = \mathbf{2}$. There is no surjective map $X \to \mathbf{2}^X$ and hence $|\mathbf{P}(X)| \not\leq |X|$.

We can apply the fixpoint lemma to any putative such map, with $\alpha = \neg$, to get the usual 'diagonalization argument'.

Russell's Paradox Let $\mathcal{S}$ be a 'universe' (set) of sets. Let $\hat{g} : \mathcal{S} \times \mathcal{S} \to \mathbf{2}$ define the membership relation:

$$\hat{g}(x,y) \Leftrightarrow y \in x$$

Then there is a predicate which can be defined on $\mathcal{S}$, and which is not representable by any element of $\mathcal{S}$.

Such a predicate is given by the standard Russell set, which arises by applying the fixpoint lemma with $\alpha = \neg$.

4.1 Abstract Version of the Basic Lemma

Lawvere's argument was in the setting of cartesian (closed) categories. Amazingly, it only needs finite products.[2]

Let $\mathcal{C}$ be a category with finite products. The terminal object (empty product) is written as $\mathbf{1}$. In **Set** it is any one-point set.

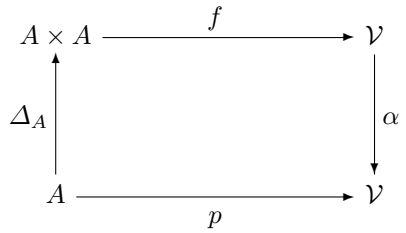
Definition 1 (Lawvere). *An arrow $f : A \times A \to \mathcal{V}$ is* weakly point surjective *(wps) if for every $p : A \to \mathcal{V}$ there is an $x : \mathbf{1} \to A$ such that, for all $y : \mathbf{1} \to A$:*

$$p \circ y = f \circ \langle x, y \rangle : \mathbf{1} \to \mathcal{V}$$

In this case, we say that *p is represented by x.*

Proposition 2 (Abstract Fixpoint Lemma). *Let $\mathcal{C}$ be a category with finite products. If $f : A \times A \to \mathcal{V}$ is weakly point surjective, then every endomorphism $\alpha : \mathcal{V} \to \mathcal{V}$ has a fixpoint $v : \mathbf{1} \to \mathcal{V}$ such that $\alpha \circ v = v$.*

Proof Define $p : A \to \mathcal{V}$ by



Suppose p is represented by $x : \mathbf{1} \to A$. Then

$$\begin{aligned} p \circ x &= \alpha \circ f \circ \Delta_A \circ x && \text{def of } p \\ &= \alpha \circ f \circ \langle x, x \rangle && \text{diagonal} \\ &= \alpha \circ p \circ x && x \text{ represents } p. \end{aligned}$$

So $p \circ x$ is a fixpoint of α. □

In [17], the Fixpoint Lemma is used to derive Gödel's First Incompleteness Theorem. Yanofsky's paper covers many more applications: semantic paradozes (Liar, Berry, Richard), the Halting Problem, existence of an oracle B such that $\mathbf{P}^B \neq \mathbf{NP}^B$, Parikh sentences, Löb's paradox, the Recursion theorem, Rice's theorem, von Neumann's self-reproducing automata, ...

All of these are 'one-person' results. The question of applying this argument to a two-person scenario such as the BK paradox has remained open.

[2] In fact, even less suffices: just monoidal structure and a 'diagonal' satisfying only point naturality and monoidality.

5 Reducing BK to Lawvere

How do we relate Lawvere to BK? As we have seen, the BK argument is valid in any regular category. This is pretty general. Nevertheless, BK needs a richer setting than Lawvere. To find common ground between them, we reformulate Lawvere, replacing *maps* by *relations*.

The rules of regular logic (just standard rules for this fragment of first-order logic) are sound in any regular category, and thus *we can use logic to reason about relations in a wide variety of mathematical contexts*. For further details, see [12].

We shall write $\mathsf{Sub} : \mathcal{C}^{\mathsf{op}} \to \mathbf{Set}$ for the subobject functor, which can be defined on any regular catgory. It sends an object A to the set of subobjects $\mathsf{Sub}(A)$, and acts by pullback on morphisms: notation is $f \mapsto f^*$.

Given a formula ϕ of regular logic, with free variables $X = x_1 : A_1, \ldots, x_n : A_n$, where each A_i is interpreted as an object of a regular category $\mathcal{C}$, the standard categorical semantics [12] assigns a subobject $[\![\phi]\!] \in \mathsf{Sub}(A_1 \times \cdots \times A_n)$ as the intepretation of ϕ.

5.1 Relational Reformulation of Lawvere

As a first step, we reformulate Lawvere's notion of *weak point surjectivity* in relational terms.

To see how to do this, imagine the Lawvere wps situation

$$\hat{g} : X \times X \to \Omega$$

is happening in a *topos*, and Ω is the subobject classifier. In the case of **Set**, Ω is just $\mathbf{2}$, and we are appealing to the familiar identification $\mathcal{P}(X) = \mathbf{2}^X$ of subsets with characteristic functions.

Then this map $\hat{g}$ corresponds to a *relation*

$$R \rightarrowtail X \times X$$

Such a relation is *weakly point surjective* (wps) if for every subobject $p \rightarrowtail X$ there is $x : \mathbf{1} \to X$ such that, for all $y : \mathbf{1} \to A$:

$$[\![R(x, y)]\!] = [\![p(y)]\!]$$

or in logical terms

$$R(x, y) \iff p(y).$$

In fact, a weaker notion suffices to prove the Fixpoint Lemma (cf. [22]). We say that R is *very weakly point surjective* (vwps) if for every subobject $p \rightarrowtail X$ there is $x : \mathbf{1} \to X$ such that:

$$[\![R(x, x)]\!] = [\![p(x)]\!].$$

5.2 What is a 'propositional operator'?

To find the right 'objective' — *i.e.* language independent — notion, once again we consider the topos case, and translate out of that into something which makes sense much more widely.

In a topos, a propositional operator is an endomorphism of the subobject classifier

$$\alpha : \Omega \to \Omega$$

(In more familiar terms: an operator on the lattice of truth values, as e.g. in Boolean Algebras with Operators.) This corresponds to the endomorphism of $\mathcal{V}$ in Lawvere's original formulation.

Note that by Yoneda, since $\mathsf{Sub} \cong \mathcal{C}(-, \Omega)$, such endomorphisms of Ω correspond bijectively with *endomorphisms of the subobject functor* — *i.e.* natural transformations

$$\tau : \mathsf{Sub} \Longrightarrow \mathsf{Sub}.$$

Thus this is the right semantic notion of 'propositional operator' in general. Naturality corresponds to *commuting with substitution.*

5.3 The Relational Lawvere Lemma

Lemma 5 (Relational Lawvere fixpoint lemma). *If R is a vwps relation on X in a regular category*[3]*, then every endomorphism of the subobject functor*

$$\tau : \mathsf{Sub} \Longrightarrow \mathsf{Sub}$$

has a fixpoint.

Note that a fixpoint $\mathbf{K1} \Longrightarrow \mathsf{Sub}$ from the constant functor valued at the terminal object is determined by its value at $\mathsf{Sub}(\mathbf{1})$.

Proof We define a predicate $P(x) \equiv \tau(R(x, x))$, so $[\![P]\!] = \tau_X(\Delta_X^*(R))$. By vwps, there is $c : \mathbf{1} \to X$ such that:

$$[\![P(c)]\!] = c^*([\![P]\!]) = \langle c, c\rangle^*(R) = [\![R(c, c)]\!].$$

Then

$$\begin{aligned}[\![P(c)]\!] &= c^*([\![P]\!]) = c^*(\tau_X(\Delta_X^*(R)) = \tau_{\mathbf{1}}(c^* \circ \Delta_X^*(R)) \\ &= \tau_{\mathbf{1}}((\Delta_X \circ c)^*(R)) = \tau_{\mathbf{1}}(\langle c, c\rangle^*(R)) \\ &= \tau_{\mathbf{1}}(c^*([\![P]\!])) = \tau_{\mathbf{1}}([\![P(c)]\!]).\end{aligned}$$

□

[3] In fact, it suffices to assume that the category is well-powered and has finite limits.

5.4 From BK to Lawvere

Now given relations

$$R_a \rightarrowtail A \times B, \qquad R_b \rightarrowtail B \times A$$

we can form their relational composition $R \rightarrowtail A \times A$:

$$[\![R(x_1, x_2)]\!] \equiv [\![\exists y.\, [R_a(x_1, y) \ \wedge \ R_b(y, x_2)]]\!]$$

Our Basic Lemma can now be restated as follows:

Lemma 6. *If R_a and R_b satisfy the BK assumptions (A1)–(A3), then R is vwps.*

Hence the relational Lawvere fixpoint lemma applies! As an immediate Corollary, we obtain:

Lemma 7 (BK Fixpoint Lemma). *If R_a and R_b satisfy the BK assumptions (A1)–(A3), then every endomorphism of the subobject functor has a fixpoint.*

6 Multi-Agent Generalization of BK

A *multiagent belief structure* in a regular category is

$$(\{A_i\}_{i \in I}, \{R_{ij}\}_{(i,j) \in I \times I})$$

where

$$R_{ij} \rightarrowtail A_i \times A_j.$$

A *belief cycle* in such a structure is

$$A \xrightarrow{R_1} A_1 \xrightarrow{R_2} \cdots \xrightarrow{R_n} A_n \xrightarrow{R_{n+1}} A$$

where we write $R : B \relbar\!\!\!\rightarrow C$ if R is a relation of the indicated type, *i.e.* a subobject of $B \times C$.

We now formulate *Generalized BK Assumptions* for such a belief cycle:

For each subobject $p \rightarrowtail A$, there is some $c : \mathbf{1} \to A$ such that

$$\begin{array}{l} c \models \Box_1 \cdots \Box_n \boxplus_{n+1} p \\ \quad \wedge \\ \quad \Diamond_1 \top \ \wedge \ \Box_1 \Diamond_2 \top \ \wedge \ \cdots \ \wedge \ \Box_1 \cdots \Box_{n-1} \Diamond_n \top \end{array}$$

These assumptions can be written straightforwardly as regular sequents.

Multiagent BK Fixpoint Lemma We can define the relation $R = R_1; \cdots; R_{n+1} : A \mathrel{-\!\!\!\rightarrowtail} A$.

Lemma 8 (Generalized Basic Lemma). *Under the Generalized BK assumptions, R is vwps.*

Hence the Relational Fixpoint Lemma applies. Note that in the one-person case $n = 0$, *assumption completeness coincides with weak point surjectivity.*

In modal terms:

$$c \models \boxplus p \;\equiv\; \forall x.\, R(c,x) \Leftrightarrow p(x).$$

One-person BK *is* (relational) Lawvere! The force of the BK argument is that the (very) wps property propagates back along *belief chains.*

In particular, this produces the '**believes-assumes**' construction of BK, or the generalized version **believes*****-assumes**, in which 'believes' is iterated n times followed by an 'assumes'.

7 Compositional Analysis

We shall briefly consider the issue of *compositional gluing* of belief relations with given completeness properties. For simplicity, we shall conduct our discussion concretely, in terms of relations on sets. To incorporate the idea of relativization to a set of predicates, we shall assume that each set A is given together with a set $\mathbf{P}(A) \subseteq \mathcal{P}(A) \setminus \{\varnothing\}$ of (non-empty) predicates on A.

Suppose we are given a relation $R : A \mathrel{-\!\!\!\rightarrowtail} B$. We say that R is *assumption-complete* (with respect to $\mathbf{P}$) if for every $p \in \mathbf{P}(B)$, for some $x \in A$, for all $y \in B$:

$$R(x,y) \;\Leftrightarrow\; p(y).$$

This is just wps again, of course.

We say that it is *belief-complete* if for all $y \in B$:

$$R(x,y) \;\Rightarrow\; p(y).$$

and also $\exists y.R(x,y)$. Modally, this corresponds to

$$x \models \Box p \;\wedge\; \Diamond\top.$$

Now suppose we have relations

$$R_{ab} : A \mathrel{-\!\!\!\rightarrowtail} B, \qquad R_{bc} : B \mathrel{-\!\!\!\rightarrowtail} C.$$

We define

$$\boxplus_{bc} p = \{y \in B \mid R_{bc}(y) = p\}.$$

Lemma 9 (Composition Lemma). *Suppose that:*

1. *R_{ab} is belief-complete with respect to $\mathbf{P}(B)$.*
2. *R_{bc} is assumption-complete with respect to $\mathbf{P}(C)$.*

3. *For each* $p \in \mathbf{P}(C)$, $\boxplus_{bc} p \in \mathbf{P}(B)$.

Then the composition $R_{ac} = R_{ab}; R_{bc} : A \rightarrowtail C$ *is assumption-complete with respect to* $\mathbf{P}(C)$.

Note the need for the *comprehension assumption* (3).

We now prove a kind of converse to the Composition Lemma, which *characterises* belief-completeness, and shows *why* the BK assumptions and the **believes-assumes** pattern arise in this context.

Theorem 1 (Compositional Characterization). *A relation* $R : A \rightarrowtail B$ *is belief complete with respect to* $\mathbf{P}(B)$ *if and only if, for every* $S : B \rightarrowtail C$ *such that*

1. S *is assumption complete with respect to* $\mathbf{P}(C)$
2. $\boxplus_S p \in \mathbf{P}(B)$ *for every* $p \in \mathbf{P}(C)$

the composition $R; S : A \rightarrowtail C$ *is assumption complete with respect to* $\mathbf{P}(C)$.

Proof The left to right implication is Lemma 9.

For the converse, we suppose that R is not belief-complete for some $p \in \mathbf{P}(B)$. We let $C = \{0, 1\}$, and define S to be the characteristic function of p. We take $\mathbf{P}(C) = \{q\}$, where $q = \{1\}$. Note that $\boxplus q = p$, and that S is assumption complete with respect to $\mathbf{P}(C)$ — indeed, any element of p, which by our general assumption on predicates is non-empty, assumes q.

We claim that $R; S$ is not assumption complete for q. Indeed, for any $x \in A$, if $R(x) = \varnothing$, then $R; S(x) = \varnothing$, and so x does not assume q. The only other possibility, since by assumption R is not belief complete with respect to p, is that for some $y \notin p$, $R(x, y)$. In this case, $R; S(x, 0)$, and so x does not assume q. □

Remark The proof of the Compositional Characterization Theorem assumes that we have the freedom to choose any collection of predicates we like on a given set. It would be useful to have a more general formulation and result.

8 Functorial Constructions of Assumption-Complete Models

We now turn to the question of constructing belief models which are assumption complete with respect to a natural class of predicates. The categorical perspective is well-suited to this task. Indeed, leaving aside model-theoretic subtleties, we can identify the problem as essentially one of finding fixpoints for certain 'powerset-like' functors. This 'recursion in the large' at the level of types, to support 'recursion in the small' at the level of programs, is a familiar theme in Theoretical Computer Science [5]. If we think of recursion as enabling self-reference, in formulas rather than programs, we see the link to the ideas being

considered here. Powerful general methods are available for finding such fixpoints, as solutions of domain equations [5] or final coalgebras [21].

The problem can be phrased as follows, in the setting of the *strategy-based belief models* of [11]. We are given strategy sets S_a, S_b for Alice and Bob respectively. We want to find sets of types T_a and T_b such that

$$T_a \cong \mathbf{P}(U_b), \qquad T_b \cong \mathbf{P}(U_a) \tag{5}$$

where $U_a = S_a \times T_a$ and $U_b = S_b \times T_b$ are the sets of states for Alice and Bob. Naively, $\mathbf{P}$ is powerset, but in fact it must be a restricted set of subsets (extensions of predicates) defined in some more subtle way, or such a structure would be impossible by mere cardinality considerations.

Thus a state for Alice is a pair (s,t) where s is a strategy from her strategy-set and t is a type. Given an isomorphism $\alpha : T_a \xrightarrow{\cong} \mathbf{P}(U_b)$, we can define a relation $R_a : U_a \rightarrowtail U_b$ by:

$$R_a((s,t),(s',t')) \;\equiv\; (s',t') \in \alpha(t).$$

Note that (s,t) **assumes** $\alpha(t)$. Because α is an isomorphism, the belief model (U_a, U_b, R_a, R_b) is automatically assumption complete with respect to $\mathbf{P}(U_a)$ and $\mathbf{P}(U_b)$.

In fact, having isomorphisms $\alpha : T_a \xrightarrow{\cong} \mathbf{P}(U_b)$, $\beta : T_b \xrightarrow{\cong} \mathbf{P}(U_a)$ is more than is strictly required for assumption completeness. It would be sufficient to have retractions

$$T_a \rhd \mathbf{P}(U_b), \qquad T_b \rhd \mathbf{P}(U_a)$$

i.e. maps

$$r_a : T_a \to \mathbf{P}(U_b), \qquad s_a : \mathbf{P}(U_b) \to T_a$$

such that $r_a \circ s_a = \mathsf{id}_{\mathbf{P}(U_b)}$, and similarly for T_b and $\mathbf{P}(U_a)$.[4] However, we shall emphasize the situation where we do have isomorphisms, where we can really speak of *canonical solutions*.

We shall now generalize this situation so as to clarify what the mathematical form of the problem is. Suppose that we have a category $\mathcal{C}$, which we assume to have finite products, and a functor $\mathbf{P} : \mathcal{C} \to \mathcal{C}$. We are given objects S_a and S_b in $\mathcal{C}$. Hence we can define functors $F_a, F_b : \mathcal{C} \to \mathcal{C}$:

$$F_a(Y) = \mathbf{P}(S_b \times Y), \qquad F_b(X) = \mathbf{P}(S_a \times X).$$

Intuitively, F_a provides one level of beliefs which Alice may hold about states which combine strategies for Bob with 'types' from the 'parameter space' Y; and symmetrically for F_b.

Now we define a functor $F : \mathcal{C} \times \mathcal{C} \to \mathcal{C} \times \mathcal{C}$ on the product category:

$$F(X,Y) = (F_a(Y), F_b(X)).$$

[4] Brandenburger and Keisler ask only for surjections, but they are working in a setting where surjections can always be split.

To ask for a pair of isomorphisms as in (5) is to ask for a *fixpoint* of the functor F: an object of $\mathcal{C} \times \mathcal{C}$ (hence a pair of objects of $\mathcal{C}$, (T_a, T_b)) such that

$$(T_a, T_b) \cong F(T_a, T_b).$$

This situation has been extensively studied in Category Theory and Theoretical Computer Science [21, 8, 5]. In particular, the notion of *final coalgebra* provides a canonical form of solution. Once again, previous work has focussed on 'one-person' situations, although the tools needed for two- or multi-agent forms of solution — essentially the ability to solve simultaneous equations — are already in hand. Indeed, final coalgebras subsume what are known as *terminal models* in the game-theoretic literature on type spaces. What amount to terminal sequence constructions of final coalgebras have been used in the literature on *Harsanyi type spaces* [14] to construct what are known as *universal models*. Heifetz and Samet gave the first construction of a universal type space in the category of measurable spaces [15], following other work in more restricted contexts. Subsequently, Moss and Viglizzo made explicit use of coalgebraic ideas in [19].

Thus these well-developed methods from Theoretical Computer Science can be used to address the following question raised by Brandenburger and Keisler:

> We end by noting that, to the best of our knowledge, no general treatment exists of the relationship between universal, complete, and terminal models (absent specific structure). Such a treatment would be very useful.

Our contribution here is to set the discussion in a wider context, emphasizing the construction of interactive belief models which are assumption complete.

The topic deserves a fuller treatment than is possible here. We shall content ourselves with giving some examples where known results on the existence of final coalgebras can be applied to yield assumption complete models.

8.1 Application to Assumption-Complete Models

We begin by noting that standard results allow us to lift one-person to two- (or multi-)agent constructions. Suppose we have endofunctors $G_1, G_2 : \mathcal{C} \to \mathcal{C}$. We can define a functor

$$G : \mathcal{C} \times \mathcal{C} \to \mathcal{C} \times \mathcal{C} \ :: \ G(X, Y) = (G_1(Y), G_2(X)).$$

Note that this directly generalizes our definition of F from F_a and F_b. We have $G = (G_1 \times G_2) \circ \mathsf{twist}$. It is standard that if G_1 and G_2 satisfy continuity or accessibility hypotheses which guarantee that they have final coalgebras, so will G.

Note that the final sequence for G will have the form

$$(\mathbf{1}, \mathbf{1}) \leftarrow (G_1(\mathbf{1}), G_2(\mathbf{1})) \leftarrow (G_1(G_2(\mathbf{1})), G_2(G_1(\mathbf{1})) \leftarrow$$

$$\cdots \leftarrow ((G_1 \circ G_2)^k(\mathbf{1}), (G_2 \circ G_1)^k(\mathbf{1})) \leftarrow \cdots$$

This 'symmetric feedback' is directly analogous to constructions which arise in Geometry of Interaction and the Int construction [4, 1, 6]. It is suggestive of a compositional structure for interactive belief models.

We shall now consider three specific settings where the general machinery we have described can be applied to construct assumption complete models as final coalgebras. In each case we must specify the ambient category $\mathcal{C}$, and the functor $\mathbf{P}$.

Sets We firstly consider **Set**, the category of sets and functions. Our candidate for $\mathbf{P}$ is a variant of the powerset functor. We take $\mathbf{P}(X) = \mathcal{P}_\kappa(X)$, the collection of all subsets of X of cardinality less than κ, where κ is an inaccessible cardinal.[5] It is standard that, for any sets S_a, S_b, the functors F_a and F_b are accessible, and hence so is the functor $F = (F_a \times F_b) \circ \mathsf{twist} : \mathcal{C} \times \mathcal{C} \to \mathcal{C} \times \mathcal{C}$. Hence we get a final coalgebra

$$\gamma : (T_a, T_b) \xrightarrow{\cong} (\mathcal{P}_\kappa(S_a \times T_b), \mathcal{P}_\kappa(S_b \times T_a)).$$

This yields an assumption complete belief model, as previously discussed.

Note that the terminal sequence for this functor is always transfinite, as analyzed in detail in [24]. Even in the case $\kappa = \omega$ (finite subsets), $\omega + \omega$ stages are required for convergence to the final coalgebra.

Stone Spaces Another convenient setting for final coalgebra is the category of *Stone spaces*, *i.e.* totally disconnected compact Haussdorff spaces [2, 16]. By Stone duality, this category is dual to the category of Boolean algebras. Our candidate for $\mathbf{P}$ here is the *Vietoris powerspace construction* [18]. In [2], one can find essentially a treatment of the one-person case of the situation being considered here. The final coalgebra constructed here is closely related to the model built in a more concrete fashion in [11]. We get stronger properties (isomorphism rather than surjection) and a clearer relation to general theory.

In this case, the final coalgebra is reached after ω stages of the terminal sequence, because of continuity properties of the functor.

Algebraic Lattices As a final example, we venture into the realm of Domain theory [13, 5]. We work in the category of algebraic lattices and Scott-continuous maps (those preserving directed joins). We have two convenient choices for $\mathbf{P}$: the *lower* and *upper powerdomain* constructions, both well-studied in Domain theory. In the first case, we take the lattice of Scott-closed subsets of an algebraic lattice, ordered by inclusion. In the second, we take the subsets which are compact in the Scott topology, and upwards closed in the partial ordering, ordered by reverse inclusion. In either case, we obtain a continuous functor, which converges to the final coalgebra in ω stages of the terminal sequence.

[5] Alternatively, and essentially equivalently, we can follow Peter Aczel [7], and work over the ('superlarge') category of classes, taking $\mathbf{P}(A)$ to be the class of sub-*sets* of a class A.

Closure under logical constructions We have constructed models which are assumption complete in a semantic sense, with respect to the predicates specified by the functor **P**. A further issue is how expressive these collections of predicates are; this can be made precise in terms of which logical constructions they are closed under, and hence which logics can be interpreted. Brandenburger and Keisler show that their topological belief model is closed under conjunction, disjunction, existential and universal quantification, and constructions corresponding to the **assumes** and **believes** modalities. The same arguments show that our model in Stone spaces is closed under these constructions. Similar arguments show that the model in **Set** is also closed under these constructions. In this case, closure under the **believes** modality requires that if a set S has cardinality less than κ, so does its powerset. This follows from the inaccessibility of κ. Finally, the models in algebraic lattices are also closed under these constructions, with the proviso that appropriate order-theoretic saturation (upwards or downwards closure) must be applied in some cases.

These models also allow for various forms of recursive definition. We leave a detailed account to an extended version of this paper.

9 Further Directions

There are a number of natural directions to be pursued. One is to a more comprehensive account of the construction of belief models and type spaces, taking full advantage of the use of categorical methods, and of developments in coalgebraic logic. Another is to a finer analysis of the use of completeness hypotheses in justifying solution concepts for games. Finally, we would like to pursue the broader agenda of understanding the mathematical structure of interaction, and the scope of interactive versions of logical and mathematical phenomena which have previously only been studied in 'one-person' versions.

Wider perspective: applied coalgebra We also wish to put this work in a wider perspective, emphasizing in particular the rôle which we believe that coalgebra can play, as a wide-ranging theory of systems encompassing reflexive forms of behaviour, going far beyond the direct applications in computer science which have mainly been studied to date. Coalgebraic representations of physical systems are discussed in [3]. Potential further applications in biology and economics are currently under investigation.

10 Acknowledgements

This research was supported by the EPSRC grant EP/F067607/1 and by ONR. An extended abstract of an earlier version of this paper was presented at the LOFT 2010 conference. Discussions with Adam Brandenburger and Viktor Winschel are gratefully acknowledged.

References

1. S. Abramsky. Retracing some paths in process algebra. In U. Montanari and V. Sassone, editors, *CONCUR '96: Concurrency Theory, 7th International Conference*, pages 1–17. Springer-Verlag, 1996.
2. S. Abramsky. A Cook's tour of the finitary non-well-founded sets. In Sergei Artemov, Howard Barringer, Artur d'Avila Garcez, Luis C. Lamb, and John Woods, editors, *We Will Show Them: Essays in honour of Dov Gabbay*, volume 1, pages 1–18. College Publications, 2005.
3. S. Abramsky. Coalgebras, Chu spaces, and representations of physical systems. In *Logic in Computer Science (LICS), 2010 25th Annual IEEE Symposium on*, pages 411–420. IEEE, 2010.
4. S. Abramsky and R. Jagadeesan. New foundations for the geometry of interaction. In *Information and Computation, 111(1)*, pages 53–119, 1994.
5. S. Abramsky and A. Jung. Domain theory. In S. Abramsky, D. Gabbay, and T. S. E. Maibaum, editors, *Handbook of Logic in Computer Science*, pages 1–168. Oxford University Press, 1994.
6. S. Abramsky and P.-A. Melliés. Concurrent games and full completeness. In *Proceedings of the Fourteenth International Symposium on Logic in Computer Science*, pages 431–442. Computer Society Press of the IEEE, 1999.
7. Peter Aczel and Nax Paul Mendler. A final coalgebra theorem. In David H. Pitt, David E. Rydeheard, Peter Dybjer, Andrew M. Pitts, and Axel Poigné, editors, *Category Theory and Computer Science*, volume 389 of *Lecture Notes in Computer Science*, pages 357–365. Springer, 1989.
8. Michael Barr. Terminal coalgebras in well-founded set theory. *Theor. Comput. Sci.*, 114(2):299–315, 1993.
9. P. Battigalli and M. Siniscalchi. Strong belief and forward-induction reasoning. *Journal of Economic Theory*, 106:356–391, 2002.
10. A. Brandenburger, A. Friedenberg, and H.J. Keisler. Admissibility in games. *Econometrica*, 76:307–352, 2008.
11. Adam Brandenburger and H. Jerome Keisler. An impossibility theorem on beliefs in games. *Studia Logica*, 84(2):211–240, November 2006.
12. Carsten Butz. Regular categories and regular logic. Technical Report LS-98-2, BRICS, October 1998.
13. G. Gierz, K. H. Hofmann, K. Keimel, J. D. Lawson, M. Mislove, and D. S. Scott. *Continuous Lattices and Domains*. Number 93 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2003.
14. John C. Harsanyi. Games with incomplete information played by "Bayesian" players, I–III. Part I. The basic model. *Management Science*, 14(3), 1967.
15. A. Heifetz and D. Samet. Topology-free typology of beliefs. *Journal of Economic Theory*, 82:324–381, 1998.
16. Clemens Kupke, Alexander Kurz, and Yde Venema. Stone coalgebras. *Theor. Comput. Sci.*, 327(1-2):109–134, 2004.
17. F. William Lawvere. Diagonal arguments and cartesian closed categories. *Lecture Notes in Mathematics*, 92:134–145, 1969.
18. E. Michael. Topologies on spaces of subsets. *Trans. Amer. Math. Soc.*, 71:152–182, 1951.
19. Lawrence S. Moss and Ignacio D. Viglizzo. Final coalgebras for functors on measurable spaces. *Inf. Comput.*, 204(4):610–636, 2006.

20. Eric Pacuit. Understanding the Brandenburger-Keisler paradox. *Studia Logica*, 86(3):435–454, 2007.
21. Jan J. M. M. Rutten. Universal coalgebra: a theory of systems. *Theor. Comput. Sci.*, 249(1):3–80, 2000.
22. J. Soto-Andrade and F. J. Varela. Self-reference and fixed points: a discussion and an extension of Lawvere's theorem. *Acta Applicandae Mathematicae*, 2:1–19, 1984.
23. Jaap van Oosten. Basic category theory. Technical Report LS-95-1, BRICS, January 1995.
24. James Worrell. Terminal sequences for accessible endofunctors. *Electr. Notes Theor. Comput. Sci.*, 19, 1999.
25. Noson S. Yanofsky. A universal approach to self-referential paradoxes and fixed points. *Bulletin of Symbolic Logic*, 9(3):362–386, 2003.