

# Proof Theory

## The Fixed Point Theorem for the Logic of Provability

S. J. van Gool

April 28, 2008

I declare that this essay is work done as part of the Part III Examination. I have read and understood the *Statement on Plagiarism for Part III and Graduate Courses* issued by the Faculty of Mathematics, and have abided by it. This essay is the result of my own work, and except where explicitly stated otherwise, only includes material undertaken since the publication of the list of essay titles, and includes nothing which was performed in collaboration. No part of this essay has been submitted, or is concurrently being submitted, for any degree, diploma or similar qualification at any university or similar institution.

Signed ..... Date .....

S. J. van Gool, King's College, University of Cambridge

# Contents

<b>1</b>	<b>A Logic for Provability</b>	<b>2</b>
1.1	The Modal Logic $\mathbf{G}$ . . . . .	3
1.2	Semantics for $\mathbf{G}$ . . . . .	7
<b>2</b>	<b>The Fixed Point Theorem for <math>\mathbf{G}</math></b>	<b>12</b>
2.1	Fixed points . . . . .	12
2.2	Roadmap to the Fixed Point Theorem . . . . .	15
2.3	Implicit definability of fixed points . . . . .	18
2.4	Gleit and Goldfarb's reduction . . . . .	20
<b>3</b>	<b>Craig Interpolation Theorem</b>	<b>23</b>
3.1	Separability . . . . .	24
3.2	The MIS Model . . . . .	25
3.3	Discussion of the proof . . . . .	31
<b>4</b>	<b>Conclusion</b>	<b>33</b>

# Chapter 1

## A Logic for Provability

The purpose of this essay is to give a clear exposition of the proof of the Fixed Point Theorem for provability logic, combining the ideas from the proofs found in George Boolos' [1] and Craig Smoryński's [10].

Proof theory is the systematic study of mathematical proofs. The starting point of proof theory is usually placed around the year 1931, when Kurt Gödel published his infamous paper [5]. In this paper, Gödel proved the First Incompleteness Theorem and outlined the proof of a Second Incompleteness Theorem, an explicit proof of which was later given by Paul Bernays in [6].

In his proof, Gödel constructs sentences in the language of arithmetic which can be interpreted as making assertions<sup>1</sup> about the *provability* of certain other sentences.

Gödel then realized that these sentences can just as well make assertions about *their own* provability. This self-referentiality of sentences then leads to many (apparent) paradoxes; for instance, there is a sentence which asserts its own unprovability. Such a sentence is an example of a fixed point, also see Example 2.2.

Fixed points were discovered right at the start of the study of provability. The Fixed Point Theorem, which I will discuss here, was not proved in its

---

<sup>1</sup>Here and hereafter, whenever I use the anthropomorphic terminology 'the sentence asserts  $p$ ', I mean 'the sentence is equivalent to  $p$ '. For an account of anthropomorphism in science, see [3].

full generality until 1976 by Giovanni Sambin.

The appropriate language to study the properties of the provability predicate is the language of modal logic, interpreting the  $\Box$  operator as ‘it is provable in arithmetic that’.

The axiomatic system which corresponds to this interpretation is  $\mathbf{G}$ , also known as  $\mathbf{GL}$ ,  $\mathbf{PRL}$ ,  $\mathbf{L}$ , or  $\mathbf{K4W}$  [8].

In section 1.1, I will define the language and axioms of the modal logic  $\mathbf{G}$ . Then, in section 1.2, I will review the possible world semantics appropriate to  $\mathbf{G}$ . To avoid having several pages containing only definitions, I will also prove some properties of the syntax and semantics for  $\mathbf{G}$  as I go along.

Then, in Chapter 2, I will prove the fixed point theorem for  $\mathbf{G}$ , which shows the existence of fixed points in  $\mathbf{G}$  for a large class of sentences. This proof relies heavily on the Craig Interpolation Theorem, which I will prove in Chapter 3. Chapter 4 contains a discussion of the obtained results and possible directions for further research.

I would like to thank Nathan Bowler and Olivia Caramello for organising the Part III seminar, enabling me to give a talk on part of this essay at an early stage. Thanks also to Stuart H. Smith for proof-reading a first version of the essay. Finally, I am very grateful to Martin Hyland for suggesting the topic of this essay and offering advice and guidance throughout the process of writing it.

## 1.1 The Modal Logic $\mathbf{G}$

In this section, I will define the language of modal logic and the system  $\mathbf{G}$ .

**Definition 1.1.** The **language** of modal logic consists of sentence letters  $p, q, r, \dots$ , connectives  $\rightarrow, \neg$ , a symbol  $\perp$  (‘falsum’) and one modal operator  $\Box$ .

I will make use of the following common abbreviations:

$$p \vee q := (\neg p \rightarrow q), p \wedge q := \neg(p \rightarrow \neg q), \Box p := \Box p \wedge p.$$

A **sentence**  $A$  in this language is defined to be one of the following

- (i).  $A = \perp$  or  $A = p$  for some letter  $p$ ;
- (ii).  $A = \neg B$  or  $A = \Box B$  for some sentence  $B$ ;
- (iii).  $A = B \rightarrow C$  for some sentences  $B, C$ .

A **construction of**  $A$  is a list of sentences  $(A_1, \dots, A_n)$  in which every  $A_i$  either satisfies (i) above, or satisfies (ii) or (iii) above for  $B = A_j$ ,  $C = A_k$  with  $j < i$ ,  $k < i$ , and  $A_n = A$ .

A sentence  $D$  is a **subsentence** of  $A$  if there exists a construction of  $A$  containing  $D$ . The set of subsentences of  $A$  is denoted by  $\text{Sub}(A)$ .

A letter  $p$  is said to **occur** in  $A$  if  $p \in \text{Sub}(A)$ .

If  $X$  is a set of sentences, it will be convenient to write  $\neg[X] := \{\neg A : A \in X\}$  and similarly  $\Box[X] := \{\Box A : A \in X\}$ , as well as  $\Box^{-1}[X] := \{G : \Box G \in X\}$ .

The symbols  $\vee$ ,  $\wedge$ ,  $\leftrightarrow$ ,  $\top$  are defined as abbreviations in the usual way.

Whenever I write  $A(p)$ , I am referring to a sentence  $A$  in which the letter  $p$  occurs. In that case, if  $B$  is a sentence, then  $A(B)$  (or sometimes, to avoid ambiguity,  $A_p(B)$ ) denotes the sentence in which every occurrence of  $p$  is replaced by  $B$ .

I will work in the system **G** for the rest of this essay. Following Reidhaar-Olson [8] and Boolos [1] closely, I define **G** as follows.

The **axiom schemes for G** are, for all sentences  $A$  and  $B$ , and propositional tautologies  $T$ :

- (A1).  $T$ ;
- (A2).  $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$ , the *distribution axiom*;
- (A3).  $\Box(\Box A \rightarrow A) \rightarrow \Box A$ , the *Löb axiom*.

The **rules of inference for G** are

(R1). From  $A$  and  $A \rightarrow B$ , infer  $B$ , the *modus ponens rule*.

(R2). From  $A$  infer  $\Box A$ , the *necessitation rule*.

**Definition 1.2.** A **proof in  $\mathbf{G}$**  is a list of sentences  $\langle A_1, \dots, A_n \rangle$ , such that for each  $i$ , either  $A_i$  is an instance of one of the axiom schemes (A1)-(A3), or  $A_i$  follows from earlier sentences in the list by one of the rules (R1)-(R2).

A sentence  $A$  is a **theorem of  $\mathbf{G}$**  (notation:  $\vdash_{\mathbf{G}} A$ ) if there exists a proof  $\langle A_1, \dots, A_n \rangle$  with  $A_n = A$ .

I will say a few words to clarify these definitions, which might look a bit intimidating at first sight. The underlying reason for taking these axioms and rules for  $\mathbf{G}$ , is that all of the axioms are translations of theorems of arithmetic which involve the ‘Bew’-predicate (for German ‘Beweisbar’, provable). Taking these axioms and rules, one expects to get a modal logic which effectively simulates provability. It takes quite some effort to prove that  $\mathbf{G}$  really is the right way to translate ‘Bew’ into  $\Box$ ; this is the content of Solovay’s Completeness Theorem, which I will not need in this essay. The interested reader is referred to [1], Chapter 9.

Note that if  $\vdash_{\mathbf{G}} A(p)$  for some sentence  $A(p)$ , then  $\vdash_{\mathbf{G}} A_p(H)$  for any sentence  $H$ : given a proof in  $\mathbf{G}$  of  $A(p)$ , transform it into a proof of  $A_p(H)$  by replacing all  $p$ ’s in the original proof by  $H$ . This gives a proof of  $A_p(H)$ .

The  $\Box$  operator interacts nicely with  $\wedge$  in  $\mathbf{G}$ , as one would expect from the ‘provability’ interpretation.

**Proposition 1.3.** *Let  $A, B$  be sentences. Then  $\vdash_{\mathbf{G}} \Box(A \wedge B) \leftrightarrow (\Box A \wedge \Box B)$ .*

*Proof.* Observe that  $\vdash_{\mathbf{G}} B \rightarrow (A \rightarrow (A \wedge B))$  since that sentence is a tautology. Necessitation and distribution then yield

$$\vdash_{\mathbf{G}} \Box B \rightarrow \Box(A \rightarrow (A \wedge B)).$$

Also

$$\vdash_{\mathbf{G}} \Box(A \rightarrow (A \wedge B)) \wedge \Box A \rightarrow \Box(A \wedge B),$$

since this sentence is equivalent to a distribution axiom. From these two it follows that  $\vdash_{\mathbf{G}} (\Box A \wedge \Box B) \rightarrow \Box(A \wedge B)$ .

For the other direction, use the tautology  $(A \wedge B) \rightarrow A$  and a distribution axiom to conclude that  $\vdash_{\mathbf{G}} \Box(A \wedge B) \rightarrow \Box A$ . Doing the same thing for the tautology  $(A \wedge B) \rightarrow B$  and combining the two implications gives the desired result.  $\square$

Until now, there has been no mention of the axiom scheme ' $\Box A \rightarrow \Box\Box A$ ', which is part of a large class of systems of modal logic called K4. The following basic fact will provide some constraint on the behaviour of the  $\Box$ .

**Proposition 1.4.** *For any sentence  $A$ ,*

$$\vdash_{\mathbf{G}} \Box A \rightarrow \Box\Box A.$$

*Proof.* I will repeatedly use Proposition 1.3 and a combined rule of necessitation and distribution:

(RND). If  $\vdash_{\mathbf{G}} B \rightarrow C$ , then  $\vdash_{\mathbf{G}} \Box B \rightarrow \Box C$ .

By this rule,  $\vdash_{\mathbf{G}} \Box\Box A \rightarrow (\Box\Box A \wedge \Box A)$ , so that  $\vdash_{\mathbf{G}} A \rightarrow (\Box\Box A \rightarrow \Box A)$  is a logical consequence of the tautology  $A \rightarrow ((\Box\Box A \wedge \Box A) \rightarrow \Box A)$ . Using (RND) again,  $\vdash_{\mathbf{G}} \Box A \rightarrow \Box(\Box\Box A \rightarrow \Box A)$ . Apply the Löb axiom to  $\Box A$  to obtain  $\vdash_{\mathbf{G}} \Box A \rightarrow \Box\Box A$ . By (RND),  $\vdash_{\mathbf{G}} \Box\Box A \rightarrow \Box\Box A$ , from which the proposition now follows.  $\square$

One good reason to include the last proof is to make clear that although syntactic proofs in  $\mathbf{G}$  are possible, their aesthetic appeal is very small. To see an extreme example, cf. Boolos' [1] p. 9, a page which contains 173  $\Box/\Diamond$ -symbols, and only 73 words. A much smoother way of proving theorems in  $\mathbf{G}$  is given by the semantics which I will discuss in the next section.

## 1.2 Semantics for $\mathbf{G}$

In this section, I will briefly review the possible worlds semantics for  $\mathbf{G}$ . Most of this material is from Boolos [1], Chapter 4, and Smorynski [10], Chapter 2. The aim of this section is to collect exactly the parts I will need to be able to prove the Fixed Point Theorem.

I will first define the concept of a model for  $\mathbf{G}$ .

**Definition 1.5.** A **model  $M$  for  $\mathbf{G}$**  or **FIT model** is a triple  $M = \langle W, R, V \rangle$ , where  $W$  is a **Finite** set (the *set of worlds*),  $R$  is an **Irreflexive, Transitive** relation on  $W$  (the *accessibility relation*), and  $V$  is a relation (the *valuation*) between worlds and sentence letters.

If  $M$  is a model,  $w \in W$ , and  $A$  a modal sentence, then  $w$  **believes**  $A$  ( $\langle M, w \models A \rangle$ ) is defined inductively by

- $M, w \models p$  if and only if  $wVp$ , for  $p$  a sentence letter;
- $M, w \models B \rightarrow C$  if and only if  $M, w \not\models B$  or  $M, w \models C$ ;
- $M, w \models \neg B$  if and only if  $M, w \not\models B$ ;
- $M, w \models \Box B$  if and only if  $M, x \models B$  whenever  $wRx$ .

A model  $M$  **believes**  $A$  ( $\langle M \models A \rangle$ ) if and only if  $M, w \models A$  for all  $w \in W$ .

As one would certainly hope, FIT models believe all theorems of  $\mathbf{G}$ :

**Theorem 1.6** (Soundness Theorem). *Suppose  $\vdash_{\mathbf{G}} A$ . Whenever  $M$  is a FIT model,  $M \models A$ .*

*Proof.* It is elementary, but tedious, to check that all the axioms and rules of  $\mathbf{G}$  hold in FIT models, see Boolos [1], Chapter 4. From this and the definition of  $\mathbf{G}$ , the theorem is clear.  $\square$

In particular, this gives an easy way of proving that  $\not\vdash_{\mathbf{G}} A$ : simply exhibit a FIT model  $M$  such that  $M \not\models A$ .

From how I defined models in Definition 1.5, it is intuitively clear that worlds ‘never look back’, or, in the language of the following definition, that the decision of a world  $w$  whether to believe something or not will depend only on the submodel generated by  $w$ , since that submodel contains everything  $w$  knows.

**Definition 1.7** (Generated Submodels). Let  $M = \langle W, R, V \rangle$  be a model,  $w \in W$ .

The **accessible worlds from**  $w$  are worlds  $x \in W$  such that  $xRw$ . The set of accessible worlds from  $w$  is denoted by  $R[w]$ .

The **submodel generated by**  $w$  is the model  $\langle R[w] \cup \{w\}, R', V' \rangle$ , where  $R'$  and  $V'$  are the restrictions of the relations  $R$  and  $V$  to the domain  $R[w] \cup \{w\}$ .

The Lemma which makes the idea of worlds being ‘non-retrospective’ rigorous is the Continuity Lemma.

**Lemma 1.8** (Continuity Lemma). *Let  $M = \langle W, R, V \rangle, N = \langle W, S, T \rangle$  be models on the same domain, and  $w \in W$ . Suppose that  $M[w] = N[w]$ . Then, for any formula  $A$ :  $M, w \models A$  if and only if  $N, w \models A$ .*

*Proof.* By induction on the number of connectives in  $A$ .

The propositional cases are all immediate. The only mildly interesting case is  $A = \Box B$  for some sentence  $B$ .

If  $wRx$ , then a submodel generated by  $x$  is contained in a submodel generated by  $w$ . Therefore, from the hypothesis  $M[w] = N[w]$  it follows that  $M[x] = N[x]$ , and I may apply the induction hypothesis to the submodels generated by  $x$ . This gives  $M, x \models B$  if and only if  $N, x \models B$ .

Now, since  $x$  was arbitrary, this proves that  $M, w \models \Box B$  if and only if  $N, w \models \Box B$ , as claimed.  $\square$

I briefly remarked on my disliking syntactically involved proofs at the end of Section 1.1. This is not just a matter of personal taste: many introductory logic texts, even when discussing ‘easy’ propositional calculus, start by

proving that truth tables provide a sound and complete way to determine the syntactic validity of a sentence, enabling one to prove semantic entailment instead of syntactic entailment. In modal logic, where the extra operator  $\Box$  at the same time renders truth tables obsolete and makes the syntactic calculations more involved, the need for such a ‘semantic entailment implies syntactic entailment’-principle is even more pressing.

The assertion that FIT models provide such semantics is validated by the following theorem, first proved by Segerberg in 1971 ([9] Theorem II.2.2, pp. 86-88).

**Theorem 1.9** (Completeness Theorem). *Suppose that, whenever  $M$  is a FIT model,  $M \models A$ . Then  $\vdash_{\mathbf{G}} A$ .*

*Proof (sketch).* Assuming that  $\not\vdash_{\mathbf{G}} A$ , a FIT model  $M_A$  is constructed which does not believe  $A$ : the worlds of  $M_A$  are maximal consistent sets of sentences in the language of  $A$ .

The trick in the proof is to choose an irreflexive transitive accessibility relation and valuation in such a way that a world in the model precisely believes those sentences which are its members. At this point in the proof, the syntactic description (axioms and rules of inference) of  $\mathbf{G}$  is once and for all ‘translated’ into a semantic description, the language of FIT models.

To finish the proof, one notes that  $\{\neg A\}$  is consistent and therefore contained in some world of the model, which then believes  $\neg A$ , so that  $M_A \not\models A$ .

The method of this proof is similar to the construction I will use to prove the Craig interpolation theorem given in Chapter 3, also see Remark 3.10 at the end of Chapter 3.

For a detailed account of the proof, see for example Boolos [1], Chapter 5. □

Whenever  $\not\vdash_{\mathbf{G}} A$  for some sentence  $A$ , it is possible to extract some more useful information about the worlds in the countermodel  $M$  at which  $A$  fails; these worlds cannot be scattered over  $M$  in a completely random way. It

turns out there must always be a last world at which  $A$  fails. This is the content of the following lemma.

**Lemma 1.10** (Last Resisting World). *Suppose  $M = \langle W, R, V \rangle$  is a FIT model such that  $M \not\models A$ . Then there exists  $x \in W$  such that both  $x \not\models A$  and for any  $y \in R[x]$ ,  $y \models A$ , or equivalently,  $x \not\models A$  and  $x \models \Box A$ .*

*Proof.* Since  $M \not\models A$ , choose  $x_0 \in W$  such that  $x_0 \not\models A$ . There are two possibilities:

- (a)  $x_0 \models \Box A$ , in which case I am done setting  $x := x_0$ .
- (b)  $x_0 \not\models \Box A$ . Then there is  $x_1$  such that  $x_0 R x_1$  and  $x_1 \not\models A$ . Note that  $x_1 \in W \setminus \{x_0\}$  since  $R$  is irreflexive.

Repeating this argument, if  $x_n \not\models \Box A$ , there is  $x_{n+1}$  such that  $x_n R x_{n+1}$  and  $x_{n+1} \not\models A$ . By induction and the transitivity of  $R$ ,  $x_i R x_{n+1}$  for all  $i \leq n$ , so  $x_{n+1} \in W \setminus \{x_0, \dots, x_n\}$ .

Since  $W$  is finite, this process must halt before reaching stage  $|W|$ . Hence, there must be  $m$  such that  $x_m \models \Box A$ , while  $x_m \not\models A$ . Setting  $w := x_m$  proves the lemma.  $\square$

The next lemma is an easy and often useful application of Proposition 1.4. It states, intuitively, that boxed sentences which a world  $w$  believes are believed by any world that  $w$  sees. In other words,  $w$  ‘transmits’ all boxed sentences it believes to all the worlds it can see.

**Lemma 1.11** (Transmission of boxed sentences). *Let  $M = \langle W, R, V \rangle$  be a FIT model, and suppose  $w \models \Box A$  and  $w R x$ . Then  $x \models \Box A$ .*

*Proof.* Let  $M, w, x, A$  be as in the assumptions. By Proposition 1.4,  $w \models \Box \Box A$  too, so  $x \models (\Box A \wedge A) = \Box A$ .  $\square$

I will need the following consequence of this lemma later.

**Lemma 1.12.** *Suppose  $\vdash_{\mathbf{G}} \Box A \rightarrow B$ . Then  $\vdash_{\mathbf{G}} \Box A \rightarrow \Box B$ .*

*Proof.* Let  $M = \langle W, R, V \rangle$  be a FIT model and  $w \in W$  such that  $w \models \Box A$ . By the Completeness Theorem 1.9, it suffices to show that  $w \models \Box B \wedge B$ .

By hypothesis,  $w \models \Box A \rightarrow B$ , so by modus ponens,  $w \models B$ .

By the Box Transmission Lemma 1.11, for each  $x \in R[w]$ ,  $x \models \Box A$ . Again, by hypothesis,  $x \models \Box A \rightarrow B$ . So  $x \models B$ . Since this holds for all  $x \in R[w]$ , I have now proved that  $w \models \Box B$ , as claimed.  $\square$

The semantics for  $\mathbf{G}$  make it easier to prove some of the deeper properties of  $\mathbf{G}$ , such as the following, which reflects the substitution principle within the formal system  $\mathbf{G}$ .

**Theorem 1.13** (Substitution Theorem). *Let  $A(p)$  be a sentence. Then*

$$\vdash_{\mathbf{G}} \Box(p \leftrightarrow q) \rightarrow A(p) \leftrightarrow A(q).$$

*Proof (sketch).* Following L. Reidhaar-Olson [8], I prove this by induction on the complexity of  $A(p)$ . The interesting case is when  $A(p) = \Box D(p)$ . Then, if  $M$  is a FIT model, any world  $w$  believing the premise  $\Box(p \leftrightarrow q)$  only sees worlds believing the premise (Lemma 1.11), so by the induction hypothesis all the worlds it sees believe  $D(p) \leftrightarrow D(q)$ . Therefore  $M, w \models \Box(D(p) \leftrightarrow D(q))$ .

The last step is to show that  $\vdash_{\mathbf{G}} \Box(D(p) \leftrightarrow D(q)) \rightarrow (\Box D(p) \leftrightarrow \Box D(q))$ , a consequence of the distribution axiom, so that  $M, w \models A(p) \leftrightarrow A(q)$ . This concludes the proof, by the Completeness Theorem 1.9.  $\square$

# Chapter 2

## The Fixed Point Theorem for G

### 2.1 Fixed points

The objects of interest in this essay are fixed points of sentences of the modal language.

Given an operation  $U$  acting on a space of objects  $X$ , a **fixed point for  $U$**  is an object  $x \in X$  which gets sent to itself by  $U$ , i.e.  $U(x) = x$ .

I will now specify this general notion of fixed point defining a fixed point for a *sentence* in a language of logic.

Let  $A(p)$  be a sentence in which the letter  $p$  occurs. Analogously to the general notion given above, the sentence  $A$  acts in a natural way on the ‘space’ of all sentences  $B$  not containing  $p$ , by defining  $A(B)$  to be the sentence where every instance of ‘ $p$ ’ is substituted by ‘ $B$ ’.

How does the general notion of fixed point translate to this more concrete situation? What does ‘ $A(B) = B$ ’ mean here? In logic, ‘equality’ between two sentences must be interpreted as ‘logical equivalence’. This gives the following definition.

**Definition 2.1.** Let  $A(p)$  be a sentence of modal logic. A **fixed point for**

$A(p)$  (in  $\mathbf{G}^1$ ) is a sentence  $H$  not containing  $p$  such that

$$\vdash_{\mathbf{G}} H \leftrightarrow A(H).$$

Fixed points for sentences are connected to self-referentiality: a fixed point  $H$  for a given sentence  $A(p)$  can be regarded as a sentence which makes a statement about itself. More specifically,  $H$  ‘states’ that  $A(H)$  holds. This may lead to paradoxical sentences. Consider the following famous example.

**Example 2.2.** Let  $A = \neg\Box p$ . A fixed point for  $A$  is a sentence  $H$  which ‘asserts’ its own unprovability: it satisfies

$$\vdash_{\mathbf{G}} \neg\Box H \leftrightarrow H.$$

In this case, a fixed point for  $A(p)$  is easily exhibited, as I will show now.

**Claim.**  $H = \neg\Box\perp$  is a fixed point for  $A(p)$ .

*Proof.* I may show that

$$\vdash_{\mathbf{G}} \Box\neg\Box\perp \leftrightarrow \Box\perp.$$

Note that  $\Box(\Box\perp \rightarrow \perp) \rightarrow \Box\perp$  is an axiom, so it only remains to show the other implication.

This can be done by a straight-forward syntactic calculation:

$$\begin{aligned} &\vdash_{\mathbf{G}} \perp \rightarrow (\Box\perp \rightarrow \perp) \\ &\vdash_{\mathbf{G}} \Box(\perp \rightarrow (\Box\perp \rightarrow \perp)) \text{ (necessitation),} \\ &\vdash_{\mathbf{G}} \Box\perp \rightarrow \Box(\Box\perp \rightarrow \perp) \text{ (distribution),} \end{aligned}$$

as claimed. □

Even though some sentences are easily seen to have fixed points, it is not at all *prima facie* clear why a large class of sentences should have fixed points.

---

<sup>1</sup>In this essay I will only consider fixed points in  $\mathbf{G}$ . Whenever I write ‘fixed point’, this is to be understood as ‘fixed point in  $\mathbf{G}$ ’.

In this essay, I will prove the existence of ‘fixed points’ in a stronger sense than defined above:

**Definition 2.3.** A **strong fixed point for**  $A(p)$  is a sentence  $H$  not containing  $p$  such that

$$\vdash_{\mathbf{G}} \Box(p \leftrightarrow A(p)) \leftrightarrow \Box(p \leftrightarrow H).$$

Strong fixed points, as the name suggests, really are stronger than the fixed points defined in Definition 2.1. Intuitively, a strong fixed point ‘defines’ what it means to be a demonstrable fixed point for  $A(p)$ ; if  $H$  is a strong fixed point, then the only way to be a provable fixed point for  $A(p)$  is to be provably equivalent to  $H$ .

*Remark 2.4.* Not every sentence can be expected to have a fixed point. For instance, if the sentence  $A(p) = \neg p$  had a fixed point, then there would be a sentence  $H$  satisfying

$$\vdash_{\mathbf{G}} H \leftrightarrow \neg H,$$

so that  $\mathbf{G}$  would be inconsistent.

However, the following turns out to be a sufficient condition for a sentence to have fixed points:

**Definition 2.5.** A sentence  $A(p)$  is called **modalized in**  $p$  if  $A$  is a truth-functional compound of sentence letters other than  $p$  and sentences of the form  $\Box D$ .

This simply means that any occurrence of the letter  $p$  is within the scope of a  $\Box$ , or, equivalently, that  $A(p) = B(\Box C_1(p), \dots, \Box C_n(p))$  where  $B(q_1, \dots, q_n)$  is propositional and  $p$  does not occur in  $B(q_1, \dots, q_n)$ .

**Example 2.6.** Consider the following two sentences:

(a)  $\Box p \wedge q,$

(b)  $\Box p \vee p.$

Sentence (a) is modalized in  $p$ , but not in  $q$ . Indeed, letting  $B(q_1) := q_1 \wedge q$  and  $C_1(p) := p$  shows that (a) has the desired form.

Sentence (b) is not modalized in  $p$ , but it is (trivially) modalized in  $q$ . It is also not hard to prove that (b) is not equivalent to a sentence which is modalized in  $p$ , so that one could call (b) a ‘strongly non-modalized’ sentence.  $\square$

Establishing the following theorem will be the aim of this essay.

**Theorem 2.7** (Fixed Point Theorem for  $\mathbf{G}$ ). *Let  $A(p)$  be a sentence modalized in  $p$ . Then there exists a strong fixed point  $H$  for  $A(p)$ .*

*Remark 2.8.* The condition of being modalized in  $p$  is not necessary for having a fixed point: the sentence  $\Box p \vee p$  has a fixed point,  $\top$ , as can be proved using arguments very similar to those in sections 2.3 and 2.4. However, as I mentioned in Example 2.6(b), this sentence is ‘strongly not-modalized’ in  $p$ : it is not equivalent to any sentence which is modalized in  $p$ .  $\square$

## 2.2 Roadmap to the Fixed Point Theorem

I will now describe a ‘roadmap’ to the proof of the Fixed Point Theorem 2.7. In this roadmap, I will first establish the global strategy of the proof, omitting many details and referring the reader to later sections, where the proofs of the main lemmata can be found.

The aim of the proof is to establish the existence of a strong fixed point. A powerful tool in proving the existence of some kind of sentence is the Beth definability theorem. Intuitively, this theorem states that whenever we are confronted with a ‘nice enough’ property of sentences, there must exist a sentence which ‘explicitly defines’ this property.

It is important to specify the two predicates in quotation marks: (i) when is a sentence nice enough, and (ii) when does a sentence explicitly define a property?

**Definition 2.9** (Definability). Let  $D(p)$  be a sentence.

(i) The letter  $p$  is **implicitly definable from**  $D(p)$  if

$$\vdash_{\mathbf{G}} (D(p) \wedge D(q)) \rightarrow (p \leftrightarrow q).$$

(ii) The letter  $p$  is **explicitly definable from**  $D(p)$  if there exists a sentence  $H$  containing only letters from  $D(p)$  other than  $p$ , such that

$$\vdash_{\mathbf{G}} D(p) \rightarrow (p \leftrightarrow H).$$

$H$  is called an **explicit definition** of  $p$  from  $D(p)$ .

**Theorem 2.10** (Beth Definability Theorem). *Suppose  $p$  is implicitly definable from  $D(p)$ . Then  $p$  is explicitly definable from  $D(p)$ .*

*Proof.* This is a consequence of the Craig Interpolation Theorem<sup>2</sup>.

**Theorem 2.11** (Craig Interpolation Theorem). *Let  $A$  and  $C$  be sentences such that  $\vdash_{\mathbf{G}} A \rightarrow C$ . There exists a sentence  $B$ , containing only letters which are in both  $A$  and  $C$ , such that both of the following hold:*

$$\begin{cases} \vdash_{\mathbf{G}} A \rightarrow B, \\ \vdash_{\mathbf{G}} B \rightarrow C. \end{cases}$$

This ‘inverse to modus ponens’ is by far the most technical result needed for this proof of the Fixed Point Theorem. The proof of Theorem 2.11 will be given in Chapter 3.

To derive Beth’s Definability Theorem from this theorem, let  $D(p)$  be a sentence which implicitly defines  $p$ .

Set  $A(p) := D(p) \wedge p$  and  $C(q) := D(q) \rightarrow q$ . From the assumption that  $D(p)$  implicitly defines  $p$ , an easy propositional calculations shows that

---

<sup>2</sup>Usually called ‘Craig Interpolation Lemma’, but, as Boolos remarks, the label Theorem “would be preferable” ([1], ch. 8, note 4)

$\vdash_{\mathbf{G}} A(p) \rightarrow C(q)$ . Hence, by Craig's Theorem 2.11, there is a sentence  $H$  such that both

$$\vdash_{\mathbf{G}} ((D(p) \wedge p) \rightarrow H), \quad (2.1)$$

$$\vdash_{\mathbf{G}} (H \rightarrow (D(q) \rightarrow q)) \quad (2.2)$$

hold, and  $H$  only contains letters in both  $A(p)$  and  $C(q)$ . In particular,  $H$  only contains letters from  $D$ , and not the letter  $p$ .

Substituting  $p$  for  $q$  in (2.2) and again performing a little syntactic manipulation, (2.1) and (2.2) can be transformed into

$$\vdash_{\mathbf{G}} D(p) \rightarrow (p \leftrightarrow H),$$

as claimed. □

The obvious way to proceed after establishing Theorem 2.10 is to apply it to an appropriately chosen  $D(p)$ , in order to get a strong fixed point.

Setting  $D(p) := \Box(p \leftrightarrow A(p))$ , an explicit definition  $H$  of  $p$  from  $D(p)$  is not exactly the same as a strong fixed point for  $A$ . However,  $H$  does satisfy the '**light fixed point**' property:

$$\vdash_{\mathbf{G}} \Box(p \leftrightarrow A(p)) \rightarrow (p \leftrightarrow H). \quad (2.3)$$

Now, the following crucial lemma justifies the use of Beth's Theorem (2.10).

**Lemma 2.12** (Implicit definability of fixed points). *Let  $A(p)$  be a sentence modalized in  $p$ . Then  $p$  is implicitly definable from  $\Box(p \leftrightarrow A(p))$ .*

*Proof.* This key result was proved independently by Bernardi, de Jongh and Sambin. I will give the model-theoretic proof, due to de Jongh, in section 2.3. □

Hence, I obtain a sentence  $H$  which is a **light fixed point**, in the sense that it satisfies (2.3).

Now that I have a light fixed point, how can I ever get to the desired strong fixed point? The following lemma gives the (surprising) solution.

**Lemma 2.13** (Light Fixed Points are Strong). *If  $A(p)$  is modalized in  $p$  and  $H$  is a light fixed point for  $A(p)$ , then it is a strong fixed point.*

*Proof.* This result first appeared in Gleit and Goldfarb's [4] and will be proved in section 2.4. □

This important last shortcut then concludes the proof the Fixed Point Theorem 2.7.

The three key ingredients to the proof remain to be shown: the implicit definability of fixed points (Lemma 2.12) will be the topic of section 2.3, the reduction to light fixed points (Lemma 2.13) is done in section 2.4, and the next chapter is devoted to the proof of Craig's Interpolation Theorem (Theorem 2.11).

## 2.3 Implicit definability of fixed points

This section will be devoted to proving Lemma 2.12, which states

**Lemma.** *Let  $A(p)$  be a sentence modalized in  $p$ . Then  $p$  is implicitly definable from  $\Box(p \leftrightarrow A(p))$ .*

*Proof.* (De Jongh) To spare ink, write

$$\text{ImpDef} := \Box(p \leftrightarrow A(p)) \wedge \Box(q \leftrightarrow A(q)) \rightarrow (p \leftrightarrow q).$$

Let  $M = \langle W, R, V \rangle$  be an arbitrary FIT model. I will prove that  $M \models \text{ImpDef}$ , which implies the result by the Completeness Theorem 1.9.

To do this, I make the following

**Claim.** For any world  $w$  with  $w \not\models \text{ImpDef}$  there is a world  $y$  such that  $wRy$  and  $y \not\models \text{ImpDef}$ .

*Proof of Claim.* Suppose  $w \in W$  disbelieves ImpDef, that is to say

$$w \not\models p \leftrightarrow q, \quad (2.4)$$

$$w \models \Box(p \leftrightarrow A(p)) \wedge \Box(q \leftrightarrow A(q)). \quad (2.5)$$

First note that  $w \not\models A(p) \leftrightarrow A(q)$ , for otherwise  $w$  would certainly believe  $p \leftrightarrow q$ , since  $w \models (p \leftrightarrow A(p)) \wedge (q \leftrightarrow A(q))$ .

Now use the assumption that  $A(p)$  is modalized in  $p$  to write

$$A(p) = B(\Box C_1(p), \dots, \Box C_n(p)),$$

where  $B(q_1, \dots, q_n)$  does not contain boxes, nor the letter  $p$ .

How can it happen that  $w$  does not believe  $A(p) \leftrightarrow A(q)$ ? Since  $B$  is propositional, the trouble must be in one of the  $C_i$ 's: for some  $i$ , it must be the case that  $w \not\models \Box C_i(p) \leftrightarrow \Box C_i(q)$ . So, by distribution of  $\Box$  over  $\leftrightarrow$ ,  $w$  sees some world  $x$  such that  $x \not\models C_i(p) \leftrightarrow C_i(q)$ .

Now, the Substitution Theorem 1.13 gives

$$\vdash_{\mathbf{G}} \Box(p \leftrightarrow q) \rightarrow (C_i(p) \leftrightarrow C_i(q)),$$

so  $x$  can not believe  $\Box(p \leftrightarrow q)$  either. This means that there must be  $y \in R[x] \cup \{x\}$  such that  $y \not\models p \leftrightarrow q$ . Note that, by transitivity,  $wRy$ .

However, from (2.5) it follows that  $w \models \Box(p \leftrightarrow A(p)) \wedge \Box(q \leftrightarrow A(q))$ . Hence, by Lemma 1.11, transmit this boxed sentence to world  $y$  to get  $y \models \Box(p \leftrightarrow A(p)) \wedge \Box(q \leftrightarrow A(q))$ .

So  $y \not\models \text{ImpDef}$ , as claimed.  $\square$

If there existed a world  $w$  disbelieving ImpDef, then the Claim would contradict the Last Resisting World Lemma (1.10). So there can not be such a world, so  $M \models \text{ImpDef}$ , as claimed.  $\square$

## 2.4 Gleit and Goldfarb's reduction

The important step in the proof which shows that it suffices to prove the existence of light fixed points first appears in a joint paper by Zachary Gleit and Warren Goldfarb [4].

Recall that we want to prove Lemma 2.13:

**Lemma** (Light Fixed Points are Strong). *If  $A(p)$  is modalized in  $p$  and  $H$  is a light fixed point for  $A(p)$ , then it is a strong fixed point.*

Recall that this lemma reduces the problem of finding a strong fixed point to finding a light fixed point, which can be accomplished using the Beth Definability Theorem.

This is a surprising result, since it allows one to get to a sentence of the form  $\Box A \leftrightarrow \Box B$  assuming only  $\Box A \rightarrow B$ . This ‘reverse of implication’ does not hold in general, of course, as can be seen for example by setting  $A := p$  and  $B := p \vee q$ .

However, assuming that  $A(p)$  is modalized in  $p$  makes it possible to prove this ‘reverse of implication’ result in this particular case: semantically, the assumption that  $A(p)$  is modalized in  $p$  says that the opinion of any world regarding  $A(p)$  does not depend on its own opinion regarding  $p$ . It is therefore possible to transform an hypothetical countermodel for the reverse implication into a countermodel for the light fixed point property, by changing the opinion of a world regarding  $p$  and leaving everything else unchanged.

The following proof makes the preceding description rigorous.

*Proof.* Suppose  $H$  is a light fixed point for  $A(p)$ , i.e.

$$\vdash_{\mathbf{G}} \Box(p \leftrightarrow A(p)) \rightarrow (p \leftrightarrow H). \quad (2.6)$$

I need to establish the seemingly stronger

$$\vdash_{\mathbf{G}} \Box(p \leftrightarrow A(p)) \leftrightarrow \Box(p \leftrightarrow H). \quad (2.7)$$

By a consequence of Box Transmission (Lemma 1.12), (2.6) readily implies  $\vdash_{\mathbf{G}} \Box(p \leftrightarrow A(p)) \rightarrow \Box(p \leftrightarrow H)$ . The harder part is the converse direction.

Suppose, to obtain a contradiction, that  $\not\vdash_{\mathbf{G}} \Box(p \leftrightarrow H) \rightarrow \Box(p \leftrightarrow A(p))$ . Again by Lemma 1.12, this implies  $\not\vdash_{\mathbf{G}} \Box(p \leftrightarrow H) \rightarrow (p \leftrightarrow A(p))$ . Write  $\text{FixConverse} := \Box(p \leftrightarrow H) \rightarrow (p \leftrightarrow A(p))$ .

By the Completeness Theorem 1.9 and the Last Resisting World Lemma 1.10, there is a model  $M = \langle W, R, V \rangle$  and a world  $w \in W$  such that

$$M, w \not\models \text{FixConverse}, \quad (2.8)$$

$$M, w \models \Box \text{FixConverse}. \quad (2.9)$$

Now, the idea is to modify the model  $M$  slightly, to obtain a model  $N$  on the same frame and a world  $v$  in  $N$  which denies that  $H$  is a light fixed point (2.6). In symbols, the goal is to find  $N = \langle W, R, V' \rangle$  and  $v \in W$  such that:

$$N, v \models \Box(p \leftrightarrow A(p)) \quad (2.10)$$

$$N, v \not\models p \leftrightarrow H. \quad (2.11)$$

By (2.8),  $M, w \models p \leftrightarrow H$ . Since  $H$  does not contain the letter  $p$ , changing the valuation of  $p$  at  $w$  will not change the valuation of  $H$  at  $w$ .

Therefore, let  $N := \langle W, R, V' \rangle$  be the model with valuation  $V'$  defined by  $wV'p$  iff **not**  $wVp$ , and  $xV'q$  iff  $xVq$  for all  $(x, q) \neq (w, p)$ . In words,  $N$  is a FIT model which is exactly like the model  $M$  except for the opinion of  $w$  with regard to  $p$ .

I now claim that by choosing  $v = w$ , both (2.10) and (2.11) hold.

Clearly, (2.11) holds, because  $N, w \models p$  iff  $M, w \not\models p$ , and  $M, w \models p \leftrightarrow H$ .

The remaining question is now: why does (2.10) hold for the world  $v = w$ ?

The key idea is the following. Since  $A(p)$  is modalized in  $p$ , the answer to a question of the kind “does  $z$  believe  $A(p)$ ?” will only depend on whether or not the worlds in  $R[z]$  believe  $p$ . By irreflexivity,  $z \notin R[z]$ , so the answer does **not** depend on whether or not  $z$  itself believes  $p$ .

To make this observation rigorous, write  $A(p) = B(\Box C_1(p), \dots, \Box C_n(p))$

as before. Note that for any  $x$  seen by  $w$ ,  $N[x] = M[x]$ . Therefore, by the Continuity Lemma 1.8,  $N, w \models \Box C_i(p)$  iff  $M, w \models \Box C_i(p)$ . Also, by definition,  $N, w \models q$  iff  $M, w \models q$  for all letters  $q$  other than  $p$ . Since  $B$  is propositional, it now follows that  $N, w \models A(p)$  if and only if  $M, w \models A(p)$ .

How does (2.10) follow from this idea? By (2.8),  $M, w \not\models p \leftrightarrow A(p)$ . By the previous paragraph and the definition of  $N$ , in going from  $M$  to  $N$ ,  $w$  changes its mind regarding  $p$  but retains its opinion regarding  $A(p)$ . So I can conclude that  $N, w \models p \leftrightarrow A(p)$ .

Also, by (2.9), if  $x \in R[w]$ , then  $M, x \models p \leftrightarrow A(p)$ , so by the Continuity Lemma 1.8 again,  $N, x \models p \leftrightarrow A(p)$ . Since  $x$  was arbitrary in  $R[w]$ , this establishes that  $N, w \models \Box(p \leftrightarrow A(p))$ , as claimed in (2.10).

To summarize: starting from the assumption that the right-to-left direction of (2.7) does not hold, I have constructed a model  $N$  in which  $\Box(p \leftrightarrow A(p)) \rightarrow (p \leftrightarrow H)$  is not valid, which contradicts the hypothesis that  $H$  is a light fixed point, by the Soundness Theorem 1.6.  $\square$

There is one gap left in the proof of the Fixed Point Theorem for  $\mathbf{G}$ : the Craig Interpolation Theorem 2.11. This will be the goal of the next chapter.

# Chapter 3

## Craig Interpolation Theorem

This chapter will be devoted to proving the Craig Interpolation Theorem<sup>1</sup>, the most technical result used to prove the Fixed Point Theorem in Chapter 2. I will restate it, modifying notation slightly.

**Theorem 3.1** (Craig Interpolation Theorem). *Let  $A_1$  and  $A_2$  be sentences. If  $\vdash_{\mathbf{G}} A_1 \rightarrow A_2$ , then there exists an interpolating sentence  $A_0$ , containing only letters which are in both  $A_1$  and  $A_2$ , such that both of the following hold:*

$$\begin{cases} \vdash_{\mathbf{G}} A_1 \rightarrow A_0, \\ \vdash_{\mathbf{G}} A_0 \rightarrow A_2. \end{cases}$$

Before I embark on the proof of this theorem, I will describe the notion of separability, which is a new concept needed for the proof, in Section 3.1. Using some basic properties of separability and relying on the small amount of general theory of  $\mathbf{G}$  developed in Chapter 1, I will do the hard work for the proof of the theorem in Section 3.2.

Finally, Section 3.3 gives an overview of the Proof of the Theorem. Readers who are interested in the general idea rather than the details of the proof are encouraged to start by reading that section, working backwards from there.

---

<sup>1</sup>The exposition given here is based on Boolos' exposition found in [1], Chapter 8, pp. 119-121, clarifying some more details.

### 3.1 Separability

The crucial part in the statement of the Craig Interpolation theorem is the requirement that  $A_0$  contains only letters which are in **both**  $A_1$  and  $A_2$ . Without this requirement, the statement is easily satisfied by setting  $A_0 := A_1 \vee A_2$ .

It is therefore natural to introduce the following notation. Denote by  $P_i$  the set of sentence letters occurring in  $A_i$  ( $i = 0, 1, 2$ ). In this notation, the aim of the proof is to find an interpolating sentence  $A_0$  with  $P_0 \subseteq P_1 \cap P_2$ .

For a set of sentence letters  $P$ , denote by  $\mathcal{L}_P$  the set of sentences containing only letters from  $P$ .

Then the conclusion of the Craig Interpolation Theorem is equivalent to the assertion that there exists a sentence  $A_0 \in \mathcal{L}_{P_1 \cap P_2}$  such that both  $\vdash_{\mathbf{G}} A_1 \rightarrow A_0$  and  $\vdash_{\mathbf{G}} \neg A_2 \rightarrow \neg A_0$ . In the terminology of the following definition, the set  $\{A_1, \neg A_2\}$  is *separable*.

**Definition 3.2.** Let  $P$  and  $Q$  be sets of sentence letters.

Let  $S$  be a finite set of sentences in  $\mathcal{L}_P \cup \mathcal{L}_Q$ . A sentence  $B$  **separates**  $S$  if  $B \in \mathcal{L}_{P \cap Q}$ , and both of the following hold:

$$\begin{aligned} \vdash_{\mathbf{G}} \bigwedge (S \cap \mathcal{L}_P) \rightarrow B; \\ \vdash_{\mathbf{G}} \bigwedge (S \cap \mathcal{L}_Q) \rightarrow \neg B. \end{aligned}$$

A finite set of sentences is **separable** if there exists a separating sentence. Otherwise the set is called **inseparable**.

The notion of inseparability of  $S$  is, in a way, similar to the notion of **consistency** of  $S$ , that is,  $\not\vdash_{\mathbf{G}} \bigwedge S \rightarrow \perp$ . I will discuss this similarity a bit further in Section 3.3, but the following easy proposition gives a preliminary connection between the two notions.

**Proposition 3.3.** *Let  $S \subseteq \mathcal{L}_P \cup \mathcal{L}_Q$  be finite. Then*

- (i). *If  $S$  is consistent, then  $S$  is inseparable.*

(ii). If  $S$  is inseparable, then  $S \cap \mathcal{L}_P$  and  $S \cap \mathcal{L}_Q$  are consistent.

*Proof.* (i). To prove the contrapositive, suppose that  $S$  is separable, and let  $B$  be a separating sentence for  $S$ . Then, in particular,

$$\vdash_{\mathbf{G}} \left( \bigwedge (S \cap \mathcal{L}_P) \wedge \bigwedge (S \cap \mathcal{L}_Q) \right) \rightarrow (B \wedge \neg B).$$

Since  $S = (S \cap \mathcal{L}_P) \cup (S \cap \mathcal{L}_Q)$ , from this it follows that  $\vdash_{\mathbf{G}} \bigwedge S \rightarrow \perp$ , so that  $S$  is inconsistent.

(ii). If  $S \cap \mathcal{L}_P$  is inconsistent, then  $\perp$  is a separating sentence for  $S$ , since  $\bigwedge (S \cap \mathcal{L}_Q) \rightarrow \top$  is a tautology.

Similarly, if  $S \cap \mathcal{L}_Q$  is inconsistent,  $\top$  is a separating sentence for  $S$ .  $\square$

To prove the Craig Interpolation Theorem, assume  $A_1$  and  $A_2$  are sentences for which the conclusion of the theorem fails. Then, as I remarked above,  $\{A_1, \neg A_2\}$  is an inseparable set.

The idea of the proof of Craig's Theorem is to build a model 'on' this inseparable set which does not believe  $A_1 \rightarrow A_2$ . Then, by soundness,  $\not\vdash_{\mathbf{G}} A_1 \rightarrow A_2$ , as required.

How does one build a model 'on' a given inseparable set? This question will be answered in the next section.

## 3.2 The MIS Model

Assume for the rest of this section that  $A_1$  and  $A_2$  are sentences for which the conclusion of the Craig Interpolation Theorem fails. Then, as I showed in the last section,  $\{A_1, \neg A_2\}$  is an inseparable set.

The idea is to build a model 'on' this inseparable set by consecutively adding subsentences to this set, maintaining inseparability, until I get to a 'maximal' inseparable set, in a sense to be made precise below.

A general definition of Maximal Inseparable Set runs as one would expect:

**Definition 3.4.** Let  $Z$  be a set of sentences. A subset of sentences  $w \subseteq Z$  is called a  **$Z$ -maximal inseparable set** ( $Z$ -MIS) if it is inseparable, and for every  $D \in Z$ , either  $D \in w$  or  $\neg D \in w$ .

Specifying this with respect to the current situation, let  $X := \text{Sub}(A_1) \cup \text{Sub}(\neg A_2)$ , the set of all subsentences of  $A_1$  together with the subsentences of  $\neg A_2$ . Let  $Y := \neg[X]$ , that is,  $Y$  is the set of negations of sentences in  $X$ . I will consider  $X \cup Y$ -maximal inseparable sets.

Observe immediately that to prove that a set  $w$  is  $X \cup Y$ -maximal it suffices to check that for every  $D \in X$  the sentence itself or its negation is in  $w$ .

As the notation of Definition 3.4 suggests, the model I will build will have the maximal inseparable sets as its ‘worlds’.

The following lemma will make it possible to build maximal inseparable sets in the first place.

**Lemma 3.5.** *Let  $S \subseteq X \cup Y$  be an inseparable set and  $D \in X \setminus S$ . Then at least one of  $S \cup \{D\}$  and  $S \cup \{\neg D\}$  is inseparable.*

*Proof.* Suppose on the contrary that both  $S^+ := S \cup \{D\}$  and  $S^- := S \cup \{\neg D\}$  are separable. Choose sentences  $B, C \in \mathcal{L}_{P_1} \cap \mathcal{L}_{P_2}$  separating  $S^+$  and  $S^-$ , respectively.

I distinguish three cases, according to which sentence letters occur in  $D$ . I will show in each case that the set  $S$  is separable, contrary to the assumption.

Case 1.  $D \in \mathcal{L}_{P_1} \cap \mathcal{L}_{P_2}$ . Then  $B^* := (D \rightarrow B) \wedge (\neg D \rightarrow C)$  is in  $\mathcal{L}_{P_1} \cap \mathcal{L}_{P_2}$  too.

Note that  $S^\pm \cap \mathcal{L}_{P_i} = (S \cap \mathcal{L}_{P_i}) \cup \{\pm D\}$ , so that

$$\begin{aligned} \vdash_{\mathbf{G}} \left( \bigwedge S \cap \mathcal{L}_{P_1} \right) &\rightarrow B^* \\ \vdash_{\mathbf{G}} \left( \bigwedge S \cap \mathcal{L}_{P_2} \right) &\rightarrow \neg B^*, \end{aligned}$$

directly from the separating properties of  $B$  and  $C$ . So  $B^*$  separates  $S$ .

Case 2.  $D \in \mathcal{L}_{P_1} \setminus \mathcal{L}_{P_2}$ . Let  $B^* := B \vee C$ . Now  $S^\pm \cap \mathcal{L}_{P_1} = (S \cap \mathcal{L}_{P_1}) \cup \{\pm D\}$ , but  $S^\pm \cap \mathcal{L}_{P_2} = S \cap \mathcal{L}_{P_2}$ .

Then, since  $D \vee \neg D$  is a tautology, using the separating properties of  $B$  and  $C$  it follows from a propositional calculation that  $B^*$  separates  $S$ .

Case 3.  $D \in \mathcal{L}_{P_2} \setminus \mathcal{L}_{P_1}$ . Applying essentially the dual of the trick used in Case 2, let  $B^* := B \wedge C$ . This is again a separating sentence for  $S$ .

□

Of course, the motivation for proving this Lemma is to get the following immediate Corollary:

**Corollary 3.6.** *Let  $S \subseteq X \cup Y$  be an inseparable set. Then there is a maximal inseparable set  $w$  which contains  $S$ .*

In particular, this corollary gives a way of building a model ‘on’ the inseparable set  $\{A_1, \neg A_2\}$ , in the following way.

**Definition 3.7** (MIS-model). Let  $W$  be the set of  $X \cup Y$ -maximal inseparable sets. For  $w, x \in W$ , define an accessibility relation  $R$  by

$$wRx \leftrightarrow \Box^{-1}[w] \subseteq x \cap \Box^{-1}[x] \text{ and } \Box^{-1}[x] \setminus \Box^{-1}[w] \neq \emptyset,$$

that is,  $w$  sees  $x$  if and only if  $x$  contains all sentences of the form  $E, \Box E$ , where  $\Box E \in w$ , and  $x$  contains a boxed sentence which is not in  $w$ . Define a valuation for  $w \in W$  and  $p \in P_1 \cup P_2$  by  $wVp$  if and only if  $p \in w$ . Let  $M = \langle W, R, V \rangle$ , the **Maximal Inseparable Set-model**.

I will first show that this model is FIT, so that it is appropriate to  $\mathbf{G}$ , and then proceed by showing that any world in this model believes precisely the sentences which are its elements. Then, by Corollary 3.6, there is a world  $w$  in this model which contains  $\{A_1, \neg A_2\}$ , so that  $M, w \not\models A_1 \rightarrow A_2$ , as I wanted to show.

**Lemma 3.8.** *The MIS-model is FIT.*

Using the notation  $\square^{-1}$ , this has become nothing more than a set-theoretic exercise using the definition of  $R$ .

*Proof.* Since  $X \cup Y$  is finite, there are only finitely many  $X \cup Y$ -maximal inseparable sets.

Clearly, if  $wRx$ , then  $w \neq x$ , since  $x$  contains a boxed sentence which is not in  $w$ . So  $R$  is irreflexive.

Suppose  $wRx$  and  $xRy$ . Then  $\square^{-1}[w] \subseteq \square^{-1}[x] \subseteq y \cap \square^{-1}[y]$ . Also,  $\square^{-1}[y] \setminus \square^{-1}[w] \supseteq \square^{-1}[y] \setminus \square^{-1}[x] \neq \emptyset$ . Hence,  $wRy$ , so that  $R$  is transitive.  $\square$

The following proposition shows that the – not so obvious – definition of the accessibility relation in the MIS-model is exactly the one that enables the proof to work.

**Proposition 3.9.** *Let  $M = \langle M, R, V \rangle$  be the MIS-model,  $w \in W$  and  $D \in X$ . Then  $M, w \models D$  if and only if  $D \in w$ .*

*Proof.* By induction on the construction of  $D$ .

Since  $D \in X$ ,  $D$  is by definition either a subsentence of  $A_1$  or  $\neg A_2$ . Therefore, fix  $a$  such that  $D \in \mathcal{L}_{P_a}$ . This  $a$  will be used in the rest of the proof, and the other element of  $\{1, 2\}$  will be denoted by  $b$ .

Writing  $w_i := w \cap \mathcal{L}_{P_i}$  for  $i \in \{1, 2\}$ , it is then clear that

$$(\text{Sub}(D) \cup \neg[\text{Sub}(D)]) \cap w = (\text{Sub}(D) \cup \neg[\text{Sub}(D)]) \cap w_a, \quad (3.1)$$

that is, for any (negation of a) subsentence of  $D$ , asking whether it is in  $w$  is the same as asking whether it is in  $w_a$ . This trivial observation will be crucial to the rest of the proof, combined with the fact that  $w_a$  is consistent by Proposition 3.3(ii). I will repeatedly make use of these two facts without referring to them each time I use them.

If  $D$  is a sentence letter, then the claim is true by definition of the valuation  $V$ .

Also note that  $\perp \notin w$  since  $w_a$  is consistent. More generally, if  $A$  is in  $X \cup Y$ , then exactly one of  $A$  and  $\neg A$  is in  $w_i$ , since each  $w_i$  is consistent and  $w$  is  $X \cup Y$ -maximal.

Now suppose  $D = E_1 \rightarrow E_2$ .

By the induction hypothesis,  $w \models E_j$  if and only if  $E_j \in w$ , if and only if  $E_j \in w_a$ , according to (3.1). Therefore, to establish the claim for this case, it suffices to show that  $D \in w_a$  if and only if  $\neg E_1 \in w_a$  or  $E_2 \in w_a$ .

If  $D \in w_a$ , then, since  $w_a$  is consistent, not both  $E_1$  and  $\neg E_2$  can be in  $w$ . So, one of  $\neg E_1$  and  $E_2$  must be in  $w$ .

Conversely, if one of  $\neg E_1$  and  $E_2$  is in  $w$ , then it is in  $w_a$ , which then can not contain  $\neg D$ , so that  $w$  contains  $D$ , again by maximality.

This leaves the case where  $D = \Box E$ .

One direction is easy: if  $\Box E \in w$ , then for any  $x \in R[w]$ ,  $E \in x$  by definition of  $R$ , whence  $x \models E$  by the induction hypothesis, so that  $w \models \Box E$ .

The converse direction is the point where the proof for  $\mathbf{G}$  diverges from the standard proofs of Craig Interpolation for other logics (such as first-order, propositional logic). Supposing that  $\Box E \notin w$ , I will now exhibit an inseparable set  $S$  contained in a maximal inseparable set  $x$  seen by  $w$ , but containing  $\neg E$ , so that  $E \notin x$ , and then by the induction hypothesis,  $x \not\models E$ . By the definition of models, this will prove that  $w \not\models \Box E$ , as required. This step would not be possible when working with consistent instead of inseparable sets, see also the discussion in Remark 3.10.

Suppose  $\Box E \notin w$ . By maximality,  $\neg \Box E \in w$ . The most obvious choice for the set  $S$  is the right one, just include everything that *has* to be in a set seen by  $w$ :

$$S := \Box^{-1}[w_a] \cup \Box[\Box^{-1}[w_a]] \cup \Box^{-1}[w_b] \cup \Box[\Box^{-1}[w_b]] \cup \{\neg E, \Box E\}.$$

Any maximal inseparable set  $x$  containing this set  $S$  is seen by  $w$ , by definition of the relation  $R$ .

It remains to show that  $S$  is inseparable. This will be done in a way reminiscent of the proof of Lemma 3.5: if  $S$  were separable, then  $w$  would be

separable too.

Suppose  $S$  is separated by a sentence  $B$ . I will distinguish two cases according to what sentence letters  $D$  contains. In both of these cases the idea is to show that there are two subsets of  $w_a$  and  $w_b$ , namely  $\Box[\Box^{-1}[w_a]]$  and  $\Box[\Box^{-1}[w_b]]$ , having a separating sentence. This immediately implies that  $w$  is separable, the desired contradiction.

Case 1.  $D \notin \mathcal{L}_{P_b}$ . Then the separating sentence  $B$  satisfies

$$\vdash_{\mathbf{G}} \bigwedge (\Box^{-1}[w_a] \cup \Box[\Box^{-1}[w_a]] \cup \{\neg B\}) \rightarrow (\Box E \rightarrow E), \quad (3.2)$$

$$\vdash_{\mathbf{G}} \bigwedge (\Box^{-1}[w_b] \cup \Box[\Box^{-1}[w_b]]) \rightarrow \neg B. \quad (3.3)$$

I now apply the necessitation rule to (3.2), use the well-behaved interaction of  $\Box$  and  $\bigwedge$  (Proposition 1.3) and subsequently cancel all sentences of the form  $\Box\Box A$  for which  $\Box A$  already occurs in the big conjunction on the left: this is allowed since each sentence of the form  $\Box A \rightarrow \Box\Box A$  is a theorem of  $\mathbf{G}$  (Proposition 1.4). This gives

$$\vdash_{\mathbf{G}} \bigwedge (\Box[\Box^{-1}[w_a]] \cup \{\Box\neg B\}) \rightarrow \Box(\Box E \rightarrow E).$$

Now the consequent of this implication is asking for an application of the Löb axiom. Setting  $B' := \Box\neg B$  and reordering propositions gives

$$\vdash_{\mathbf{G}} \bigwedge (\Box[\Box^{-1}[w_a]] \cup \{\neg\Box E\}) \rightarrow \neg B'.$$

Manipulating (3.3) in the same way, I also get

$$\vdash_{\mathbf{G}} \bigwedge \Box[\Box^{-1}[w_b]] \rightarrow B',$$

which shows that  $B'$  separates two subsets of  $w_a$  and  $w_b$ , as claimed.

Case 2.  $D \in \mathcal{L}_{P_b}$ . In this case,  $B$  satisfies

$$\vdash_{\mathbf{G}} \bigwedge (\Box^{-1}[w_a] \cup \Box[\Box^{-1}[w_a]]) \rightarrow (\neg B \rightarrow (\Box E \rightarrow E)), \quad (3.4)$$

$$\vdash_{\mathbf{G}} \bigwedge (\Box^{-1}[w_b] \cup \Box[\Box^{-1}[w_b]]) \rightarrow (B \rightarrow (\Box E \rightarrow E)). \quad (3.5)$$

Also note that  $\Box^{-1}(D) = E \in \mathcal{L}_{P_a} \cap \mathcal{L}_{P_b}$ , so set  $B' := \Box(B \rightarrow (\Box E \rightarrow E))$ .

Manipulations very similar to, but even more tedious than those in Case 1, show that

$$\vdash_{\mathbf{G}} \bigwedge (\Box^{-1}[w_a] \cup \Box[\Box^{-1}[w_a]] \cup \{\neg \Box E\}) \rightarrow \neg B',$$

$$\vdash_{\mathbf{G}} \bigwedge (\Box^{-1}[w_b] \cup \Box[\Box^{-1}[w_b]]) \rightarrow B',$$

so that, again,  $B'$  separates  $w_a$  and  $w_b$ .

□

### 3.3 Discussion of the proof

The key to the proof is Proposition 3.9, relating the elements of a world to the sentences it believes in the MIS-model. This Proposition concludes the proof of the Craig Interpolation Theorem. To summarize and to be able to put the proof in a greater perspective, let me first give an overview of the proof.

*Proof of Theorem 3.1.* Suppose the conclusion of the Theorem does not hold, that is to say,  $\{A_1, \neg A_2\}$  is an inseparable set. I will derive a contradiction.

Consider the MIS-model  $M = \langle W, R, V \rangle$ , consisting of the maximal inseparable subsets of  $X \cup Y = (\text{Sub}(A_1) \cup \text{Sub}(\neg A_2)) \cup \neg[\text{Sub}(A_1) \cup \text{Sub}(\neg A_2)]$ .

By Corollary 3.6, there is a maximal inseparable set  $w \in W$  containing the inseparable set  $\{A_1, \neg A_2\}$ .

Then, by Proposition 3.9, this world  $w$  actually believes  $A_1$  and  $\neg A_2$ . So  $M \not\models A_1 \rightarrow A_2$ , while  $M$  is a FIT model by Lemma 3.8.

Therefore, by the Soundness Theorem 1.6,  $\not\vdash_{\mathbf{G}} A_1 \rightarrow A_2$ , which is the desired contradiction with the assumption of Theorem 3.1.  $\square$

*Remark 3.10.* When phrased like this, it becomes clear that the proof is very similar to model-theoretic proofs of the Craig Interpolation Theorem for propositional and first-order logic, and to the proof of the Completeness Theorem for  $\mathbf{G}$  (Theorem 1.9).

However, for the proof of the Craig Interpolation Theorem, a model of maximal *consistent* sets does not suffice, and must be replaced by a model of maximal *inseparable* sets. As I showed in Proposition 3.3, the notions of inseparability and consistency are related, but not the same.

If one tries to literally copy the proof of the Craig Interpolation Theorem for propositional or first-order logic [7], it turns out that ‘consistency’ of a set of sentences in  $\mathbf{G}$  is not a strong enough condition anymore. Attempting to prove the analogue of Proposition 3.9, if one passes to a maximal *consistent* (instead of *inseparable*) set to obtain a countermodel, one loses so much information that it is impossible to show that the proof of the case  $D = \Box E$  in the proof of 3.9 still applies.

This problem seems to be related to the fact that  $\mathbf{G}$  has no Deduction Theorem: as Smorynski [10] puts it, the rule ‘from  $A \vdash_{\mathbf{G}} B$  infer  $\vdash_{\mathbf{G}} A \rightarrow B$ ’ fails, for example putting  $A := p$ ,  $B := \Box p$ . In particular, a set of sentences  $S$  can be consistent in the sense that  $\not\vdash_{\mathbf{G}} \bigwedge S \rightarrow \perp$ , while there does exist a proof of  $\perp$  in the system  $\mathbf{G} + S$  ( $\mathbf{G}$  with the sentences in  $S$  added as axioms). The deduction theorem is necessary when one is working with maximal consistent sets. In the proof for  $\mathbf{G}$ , inseparable sets provide a way around this.

In 2004, R. Diaconescu ‘lifted’ the proof of the Craig Interpolation Theorem to a so-called institution-independent level, using the language of category theory [2]. The possibility of applying this to modal logic is not mentioned there. If the proof of [2] applies to the modal logic  $\mathbf{G}$ , this might help shed more light on the question what property of  $\mathbf{G}$  it is that makes Craig Interpolation possible.

# Chapter 4

## Conclusion

In this essay, I outlined a proof of the Fixed Point Theorem for  $\mathbf{G}$ , the axiomatic system of modal logic designed for the study of provability in arithmetic. I will now discuss the obtained result and hint on some interesting questions for further research.

The Fixed Point Theorem gives a necessary condition for a sentence to have a strong fixed point, but, as I mentioned in Remark 2.8, this condition is not sufficient. It seems natural to look for an easy necessary and sufficient condition for a sentence to have a (strong) fixed point, but in a small survey of the available literature I have not been able to find any research in this direction.

Another possible kind of extension of the Theorem has been more widely researched: there are many generalisations of the Fixed Point Theorem to other systems of modal logic and bi-modal logic, some of which are discussed in [10] and [1]. As the Craig Interpolation Theorem is so essential to the proof (cf. Remark 3.10), it seems reasonable to try to prove this kind of generalisation in systems in which an analogue of Craig Interpolation holds.

In the Introduction, I mentioned Gödel's two incompleteness theorems. Both of these theorems rely heavily on the fact that the sentence  $A(p) = \neg\Box p$  has a fixed point, as discussed in Example 2.2. The existence of this fixed point is an immediate consequence of the Fixed Point Theorem.

However, the Fixed Point Theorem asserts only the *existence* of a fixed point for any modalized sentence, but does not guarantee that there is an *algorithm* to construct such a fixed point. It is not obvious that the long proof given in this essay is constructive.

Fortunately, there are much shorter proofs of the Fixed Point Theorem, which explicitly give an algorithm to construct a fixed point ([8], [10], [1]). The advantage of these proofs is that one can obtain information about the structure of fixed points from the algorithms used to construct them.

The reason I prefer the proof given in this essay is that it is much more conceptually clear: at each step in the proof, there is a natural, intuitively clear way to proceed, and this way often turns out to be the right way. This property of this particular proof makes it possible to give an account which is not only detailed and correct, but also heuristically and didactically presentable. To give such an account is what I have tried to do in this essay.

# Bibliography

- [1] G. Boolos, *The Logic of Provability*, Cambridge University Press (1995).
- [2] R. Diaconescu, An Institution-independent Proof of Craig Interpolation Theorem, *Studia Logica* **77**: 59-79 (2004).
- [3] E. W. Dijkstra, On anthropomorphism in science, *E. W. Dijkstra Archive* **936**, <http://www.cs.utexas.edu/users/EWD/>
- [4] Z. Gleit, W. Goldfarb, Characters and Fixed Points in Provability Logic, *Notre Dame Journal of Formal Logic* **31**:1 (1990), 26-36.
- [5] K. Gödel, Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme I, *Monatshefte für Mathematik und Physik* **38** (1931), 173-198.
- [6] D. Hilbert and P. Bernays, *Grundlagen der Mathematik*, Vol. II, 2nd edition, Berlin: Springer (1968).
- [7] H. J. Keisler, Fundamentals of Model Theory, in: *Handbook of Mathematical Logic*, ed. J. Barwise, North Holland (1977), p. 72.
- [8] L. Reidhaar-Olson, A New Proof of the Fixed-Point Theorem of Provability Logic, *Notre Dame Journal of Formal Logic* **31**:1 (1990), 37-43.
- [9] K. Segerberg, *An Essay in Classical Modal Logic*, Uppsala, Filosofiska Freningen och Filosofiska Institutionen vid Uppsala Universitet (1971).

- [10] C. Smoryński, *Self-reference and modal logic*, New York: Springer-Verlag (1985).
- [11] R. M. Smullyan, *Diagonalization and Self-reference*, Oxford; New York: Clarendon Press (1994).
- [12] Verbrugge, Rineke (L.C.), “Provability Logic”, *The Stanford Encyclopedia of Philosophy (Summer 2003 Edition)*, Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/sum2003/entries/logic-provability/>