Notre Dame Journal of Formal Logic
Volume 31, Number 1, Winter 1990

A New Proof of the Fixed-Point Theorem of Provability Logic

LISA REIDHAAR-OLSON*

Abstract In this paper we give a simple semantic proof of the fixed-point theorem of the modal system G (also known as GL, PRL, L, and K4W). This proof is modeled after a syntactic proof of Sambin's found in Sambin and Valentini [10], yet is simpler than his, due to our taking advantage of the Kripke semantics for G. Other semantic proofs of the theorem exist, e.g. in Gleit [6] and Goldfarb and Gleit [7]; however, the advantages of this particular version are that it is less complicated and the fixed-point so obtained has the same general "appearance" as the original formula.

Introduction In this paper we present a simple semantic proof of the fixed-point theorem for the modal system G (also referred to as GL, PRL, L, and K4W in the literature). G is a version of modal logic that characterizes the notion of provability in a formal system, whence the name "provability logic". It is the connection of provability logic with formal systems that has brought modal logic into the mainstream. For the definition of notions pertaining to G and its connection with formalized theories, see Boolos [4].

Many proofs of the fixed-point theorem exist; one obvious way to classify the previously existing proofs is according to whether the methods used are primarily syntactic or primarily model-theoretic. The existing syntactic proofs tend to be rather complicated and magical, yet they do provide reasonable algorithms for the explicit calculation of fixed-points. Another appealing aspect of this type of proof is that the fixed-point obtained resembles the formula from which it is derived. For example, the fixed-point of the formula $\neg\Box p$ corresponding to the Gödel sentence is the similar formula $\neg\Box\bot$, and that of the formula $\Box p$ corresponding to the Henkin sentence is $\Box\top$. The previously existing semantic proofs, on the other hand, tend to dispel the mystery concerning the reasons for

*The author wishes to thank George Boolos for guidance and helpful discussions.

Received June 20, 1988; revised January 1, 1989

the truth of the theorem, yet provide unreasonably cumbersome algorithms for computation. Also, the fixed-points do not tend to "look like" the formulas from which they are derived.

The new proof to be presented here is essentially a semantic version of one of the nicest existing syntactic proofs (see Sambin and Valentini [10]). Yet it is simpler than that proof because it takes advantage of the Kripke semantics for G. It provides a simple algorithm for calculation that produces a fixed-point having the same "appearance" as the original formula, and at the same time provides much motivation through the semantic approach.

There are a number of other methods of proving the theorem. One is a nonconstructive proof using the Beth definability theorem for G (see Smoryński [11] or Boolos [4]). This particular proof of the fixed-point theorem was not noticed until a few years after the theorem was first proved by de Jongh and Sambin, as discussed below. Another technique is that used by a group of Italian universal algebraists under the leadership of R. Magari. Rather than working with a modal system, they work with Boolean algebras augmented by a single extra operator τ corresponding to the box. It amounts to a modal approach when appropriately translated. It appears that the Magari school did not realize the connection between the universal algebraic and modal programs until the early 1970's. By 1975, when Boolos solved Friedman's thirty-fifth problem (see Boolos [3] and Friedman [5]), the connection was fully understood.

The fixed-point theorem was conjectured independently by several people. First, Bernardi [1] and Smoryński independently proved a special case of the theorem; Bernardi's approach was algebraic, while Smoryński obtained an attractive computational algorithm using Kripke semantics. Smoryński's algorithm first appeared in [12]. De Jongh and Sambin later independently proved the full result. De Jongh's proof was semantic and apparently so complicated that he neglected to publish it. Sambin's approach [9] was universal algebraic in character.

After these initial solutions, there have been many revisions and improvements. A syntactic version that is a simplification by de Jongh of a proof of Sambin's can be found in [13]. Something similar appears in [10]. In [4] a very nice adaptation to a truth-table method of Smoryński's computational algorithm for the special case is presented. In the same book, Boolos gives a proof of the full theorem that is related to the Lindenbaum algebra for G, having to do with the concept of an "n-character". Roughly speaking, an n-character, if consistent, is a maximally strong formula of G that is of modal degree n. (A sentence is of *modal degree n* if and only if n is the maximum number of nested modal operators that occur in it.) Gleit [6] and Goldfarb and Gleit [7] have presented proofs along these lines that are more fully semantic. The algorithm that these proofs provide is not really feasible for doing actual calculations of fixed-points, where the relevant formulas are at all complex. It does have the advantage, though, of giving the best possible bound on the nesting of modal operators in the fixed-points so obtained.

Technical background The system G is the modal logic whose axioms consist of:

(i) all tautologies (including those in which the box may occur),
(ii) all sentences of the form $\Box(A \to B) \to (\Box A \to \Box B)$, and
(iii) all sentences of the form $\Box(\Box A \to A) \to \Box A$.

Axioms of type (ii) are called *distribution axioms*, and axioms of type (iii) are called *Löb axioms*. G has two *inference rules*; they are

(i) from A and $(A \to B)$ infer B, and
(ii) from A infer $\Box A$.

Rule (i) is called *modus ponens* and rule (ii) is called *necessitation*. As usual, a sentence A is a *theorem* of G if and only if A is the last member of a finite sequence of sentences, each of which is either an axiom or follows from one of the earlier sentences in the sequence by one of the inference rules. In this case we write "$\vdash_G A$". Such a finite sequence of sentences is called a *proof of A in* G. The following simple lemmas are needed to obtain the results presented below. The proofs of Lemmas 1 and 2 are the usual ones.

Lemma 1 *Any substitution instance of a theorem of* G *is a theorem of* G.

Lemma 2 *For any sentences A and B,* $\vdash_G \Box(A \leftrightarrow B) \to (\Box A \leftrightarrow \Box B)$.

We shall now discuss a Kripke-style model theory for G with respect to which G is both sound and complete. A *model* $\mathfrak{M}$ is an ordered triple $\langle W, R, P\rangle$ where

(i) W is a nonempty finite set;
(ii) R is an irreflexive transitive binary relation on W;
(iii) P is a mapping which assigns to any pair consisting of a member of W and a sentence letter a truth value $\boldsymbol{t}$ or $\boldsymbol{f}$, that is, $P: W \times \{p_0, p_1, \ldots\} \to \{\boldsymbol{t}, \boldsymbol{f}\}$.

R is called the *accessibility relation* of $\mathfrak{M}$. If x and y are worlds such that xRy, then we say "y is accessible from x" or "x sees y". For any $w \in W$ define $\mathrm{acc}(w) = \{x \in W \mid wRx\}$; $\mathrm{acc}(w)$ is simply the collection of worlds seen by w.

Given a model $\mathfrak{M} = \langle W, R, P\rangle$, we define the *truth* of a sentence *at* a world inductively in the usual manner. As usual, a sentence $\Box A$ is true at a world w if and only if A is true at every world $x \in \mathrm{acc}(w)$. When A is true at w we write $\mathfrak{M} \vDash_w A$. When A is false at w we write $\mathfrak{M} \nvDash_w A$. When there is no risk of ambiguity, we suppress "$\mathfrak{M}$" and simply write $\vDash_w A$ or $\nvDash_w A$. Using this notation, $\vDash_w \Box A$ if and only if $\vDash_y A$ for all $y \in \mathrm{acc}(w)$. We shall find it convenient to use the *strong box* $\boxdot$; $\boxdot A$ is an abbreviation for $A \wedge \Box A$. Note that $\vDash_w \boxdot A$ if and only if $\vDash_y A$ for all $y \in \{w\} \cup \mathrm{acc}(w)$. A sentence is *valid in* a model $\mathfrak{M}$ if and only if it is true at all worlds in the universe of $\mathfrak{M}$. A sentence is *valid* if and only if it is valid in all models.

As an immediate consequence of the finiteness, transitivity, and irreflexivity of models, worlds lie in levels in the universe according to how many worlds they see. Formally, we observe the following:

Lemma 3 *Given any model $\mathfrak{M} = \langle W, R, P\rangle$, for every $w \in W$, if w sees at least one world, then there is a greatest positive integer n such that for some $w_0, \ldots, w_n$ in W, $w = w_n R w_{n-1} R \cdots w_1 R w_0$.*

In light of Lemma 3, given any model $\mathfrak{M} = \langle W, R, P \rangle$, we define for $w \in W$ the *rank of w in* $\mathfrak{M}$ to be the number n described in Lemma 3. (If $w \in W$ sees no other world, then the rank of w is defined to be zero.) Clearly, if x and y are worlds such that xRy, then the rank of x is greater than that of y.

The following two simple lemmas will be of great use to us in the remainder of this paper:

Lemma 4 *Given any model* $\mathfrak{M} = \langle W, R, P \rangle$, $w \in W$, *and sentence* A, *if* $\models_w \Box A$ *then* $\models_x \Box A$ *for every* $x \in \mathrm{acc}(w)$. *Furthermore, we actually have* $\models_x \boxdot A$ *for every* $x \in \mathrm{acc}(w)$.

Proof: Suppose that $\models_w \Box A$, w sees x, and x sees y. By transitivity w sees y, so $\models_y A$. Since y was chosen arbitrarily, $\models_x \Box A$. Also $\models_x A$, so $\models_x \boxdot A$ as well.

Lemma 5 *Given any model* $\mathfrak{M} = \langle W, R, P \rangle$, $w \in W$, *and sentence* A, *if* $\not\models_w \Box A$ *then there is a world* x *seen by* w *such that* $\models_x \Box A$ *and* $\not\models_x A$.

Proof: Suppose that $\not\models_w \Box A$. Then for some x seen by w, $\not\models_x A$. Choose such an x of least rank. Suppose that xRy. y is of lesser rank than x, so $\models_y A$ by leastness. Since y was chosen arbitrarily, $\models_x \Box A$. Thus, $\models_x \Box A$ and $\not\models_x A$.

G is sound and complete with respect to this model theory, hence the following:

Theorem 1 *For any sentence* A, A *is a theorem of* G *if and only if* A *is valid.*

For a proof of this, see either Reidhaar-Olson [8] or Boolos and Jeffrey [2].

The fixed-point theorem We now turn our attention to the new proof of the fixed-point theorem to be presented here. Integral to the proof is the following:

Semantic Substitution Lemma *For any sentences* A, B, *and* C, $\boxdot(B \leftrightarrow C) \rightarrow (A(B) \leftrightarrow A(C))$ *is valid.*

Here "$A(B)$" is intended to denote the result of replacing all occurrences of p in A with B; the meaning of "$A(C)$" is similar.

Proof: Fix B and C. We show by induction on the complexity of A that $\boxdot(B \leftrightarrow C) \rightarrow (A(B) \leftrightarrow A(C))$ is valid. The only case in the induction that makes any appeal to modal logic is the $\Box$-case, so that is the only case that will be discussed here.

Hence suppose that A is $\Box D$ where $\boxdot(B \leftrightarrow C) \rightarrow (D(B) \leftrightarrow D(C))$ is valid. Let $\mathfrak{M}$ be any model and w any world in its universe. Suppose that $\models_w \boxdot(B \leftrightarrow C)$. Let x be any world seen by w. Then $\models_x \boxdot(B \leftrightarrow C)$ by Lemma 4. Since $\boxdot(B \leftrightarrow C) \rightarrow (D(B) \leftrightarrow D(C))$ is valid, we have $\models_x D(B) \leftrightarrow D(C)$. Thus, since x was chosen arbitrarily, $\models_w \Box(D(B) \leftrightarrow D(C))$. Therefore by Lemma 2 and completeness, $\models_w \Box D(B) \leftrightarrow \Box D(C)$.

Before proving the fixed-point theorem, we need to make the following definitions:

Definition 1 A sentence A is *modalized in* p if and only if every occurrence of the sentence letter p in A occurs within the scope of $\Box$.

Definition 2 A sentence A is *n-decomposable* if and only if for some (possibly empty) sequence $q_1, \ldots, q_n$ of distinct sentence letters not occurring in A, some sentence $B(q_1, \ldots, q_n)$ not containing p but containing all $q_1, \ldots, q_n$ (and possibly other sentence letters as well), and some sequence of distinct sentences $C_1(p), \ldots, C_n(p)$, each containing p, $A = B(\Box C_1(p), \ldots, \Box C_n(p))$.

If A is modalized in p, then it is n-decomposable for some n. For example, let A be the sentence $\Box(\Box p \rightarrow q) \vee \Box\Box p$. Then A is 2-decomposable; let $C_1(p) = \Box p \rightarrow q$, $C_2(p) = \Box p$, and $B(q_1, q_2) = q_1 \vee q_2$. A is also 1-decomposable; let $C_1(p) = p$ and $B(q_1) = \Box(q_1 \rightarrow q) \vee \Box q_1$.

Fixed-point Theorem *If A is modalized in p, then there exists a sentence D in which the only sentence letters that occur are those other than p that occur in A, and such that $\vdash_G \boxdot(p \leftrightarrow A) \rightarrow (p \leftrightarrow D)$. The sentence D is called a* fixed-point *of A. (Since $\vdash_G D \leftrightarrow A(D)$, as we shall see below, the appellation "fixed-point" is appropriate.)*

Proof: We prove by induction on n that if A is n-decomposable then it has a fixed-point.

Suppose that A is 0-decomposable. Then p does not occur in A, so A itself is a suitable D.

Suppose that every sentence that is n-decomposable has a fixed-point. We shall show that every sentence that is $(n+1)$-decomposable has a fixed-point as well. To that end, suppose that

$A(p) = B(\Box C_1(p), \ldots, \Box C_{n+1}(p))$. For each i let
$A_i(p) = B(\Box C_1(p), \ldots, \Box C_{i-1}(p), \top, \Box C_{i+1}(p), \ldots, \Box C_{n+1}(p))$. For each i, $A_i(p)$ is n-decomposable, so it has a fixed-point D_i. Put
$D = B(\Box C_1(D_1), \ldots, \Box C_{n+1}(D_{n+1}))$. We shall show that D is a fixed-point of A.

Lemma *For each i, $\vdash_G \boxdot(p \leftrightarrow A) \rightarrow \boxdot(\Box C_i(p) \leftrightarrow \Box C_i(D_i))$.*

Proof: By completeness it suffices to show that for any model $\mathfrak{M}$ and any world w in its universe $\mathfrak{M} \vDash_w \boxdot(p \leftrightarrow A) \rightarrow \boxdot(\Box C_i(p) \leftrightarrow \Box C_i(D_i))$. So fix i, $\mathfrak{M}$, and w. Suppose that $\vDash_w \boxdot(p \leftrightarrow A)$. We must show that $\vDash_w \boxdot(\Box C_i(p) \leftrightarrow \Box C_i(D_i))$, or equivalently that $\vDash_y \Box C_i(p) \leftrightarrow \Box C_i(D_i)$ for all $y \in \{w\} \cup \mathrm{acc}(w)$. Let $y \in \{w\} \cup \mathrm{acc}(w)$. Suppose that $\vDash_y \Box C_i(p)$. Then $\vDash_y \Box C_i(p) \leftrightarrow \top$. For any $x \in \mathrm{acc}(y)$, $\vDash_x \Box C_i(p)$ too, by Lemma 4; hence $\vDash_x \Box C_i(p) \leftrightarrow \top$. Thus $\vDash_y \boxdot(\Box C_i(p) \leftrightarrow \top)$. By the semantic substitution lemma, $\vDash_y A_i \leftrightarrow A$. Since y was chosen arbitrarily, $\vDash_w \Box(A_i \leftrightarrow A)$, hence $\vDash_y \boxdot(A_i \leftrightarrow A)$ by Lemma 4. Also by Lemma 4, since $\vDash_w \boxdot(p \leftrightarrow A)$, we have $\vDash_y \boxdot(p \leftrightarrow A)$ too. Thus $\vDash_y \boxdot(p \leftrightarrow A_i)$. By the induction hypothesis and completeness $\vDash_y (p \leftrightarrow D_i)$, and hence, since y was chosen arbitrarily, $\vDash_w \Box(p \leftrightarrow D_i)$. Thus $\vDash_y \boxdot(p \leftrightarrow D_i)$ by Lemma 4. Therefore by the semantic substitution lemma

$$\vDash_y C_i(p) \leftrightarrow C_i(D_i) \tag{1}$$

and

$$\vDash_y \Box C_i(p) \leftrightarrow \Box C_i(D_i). \tag{2}$$

By (2), $\models_y \Box C_i(p) \rightarrow \Box C_i(D_i)$. (We shall use (1) later.) Notice that we have shown that (1) and (2) hold for any $y \in \{w\} \cup \text{acc}(w)$ such that $\models_y \Box C_i(p)$.

Now suppose that $\not\models_y \Box C_i(p)$. Then by Lemma 5, there is some world x seen by y such that $\not\models_x C_i(p)$ and $\models_x \Box C_i(p)$. Since $x \in \{w\} \cup \text{acc}(w)$, (1) holds for x, hence $\models_x C_i(p) \leftrightarrow C_i(D_i)$. Thus $\not\models_x C_i(D_i)$. It follows that $\not\models_y \Box C_i(D_i)$. Thus $\models_y \Box C_i(D_i) \rightarrow \Box C_i(p)$. Therefore $\models_y \Box C_i(p) \leftrightarrow \Box C_i(D_i)$. This completes the proof of the lemma.

Now to finish the proof of the theorem, suppose that $\mathfrak{M}$ is a model and w a world in its universe such that $\models_w \boxdot(p \leftrightarrow A)$. By the Lemma and completeness, $\models_w \boxdot(\Box C_i(p) \leftrightarrow \Box C_i(D_i))$. So, applying the semantic substitution lemma $n+1$ times, we obtain $\models_w B(\Box C_i(p), \ldots, \Box C_{n+1}(p)) \leftrightarrow B(\Box C_1(D_1), \ldots, \Box C_{n+1}(D_{n+1}))$; that is, $\models_w A \leftrightarrow D$. Since $\models_w p \leftrightarrow A$, we have $\models_w p \leftrightarrow D$. Thus $\models_w \boxdot(p \leftrightarrow A) \rightarrow (p \leftrightarrow D)$, and since $\mathfrak{M}$ and w were chosen arbitrarily, $\boxdot(p \leftrightarrow A) \rightarrow (p \leftrightarrow D)$ is valid. So by completeness, $\vdash_G \boxdot(p \leftrightarrow A) \rightarrow (p \leftrightarrow D)$.

To illustrate the use of the above algorithm, we present several examples of fixed-point calculations. The algorithm is relatively easy to use, especially when one simplifies along the way by substituting simpler equivalent formulas whenever possible.

Example 1 Let $A(p) = \Box \neg p$. We shall calculate the fixed-point of A. Let $C_1(p) = \neg p$ and $B(q_1) = q_1$. Then $A(p) = B(\Box C_1(p))$ and $A_1(p) = B(\top) = \top$. The fixed-point D_1 of A_1 is simply $\top$, since p does not occur in A_1. Thus the fixed-point D of A is $B(\Box C_1(\top)) = B(\Box \neg \top) = \Box \neg \top$. The equivalent formula $\Box \bot$ is also a suitable fixed-point.

Example 2 Let $A(p) = \Box p \rightarrow \Box \neg p$. To calculate this fixed-point, let $C_1(p) = p$, $C_2(p) = \neg p$, and $B(q_1, q_2) = q_1 \rightarrow q_2$. Then $A(p) = B(\Box C_1(p), \Box C_2(p))$, $A_1(p) = B(\top, \Box C_2(p)) = \top \rightarrow \Box \neg p$, and $A_2(p) = B(\Box C_1(p), \top) = \Box p \rightarrow \top$. A_1 is clearly equivalent to $\Box \neg p$, and A_2 to $\top$, hence they have fixed-points $D_1 = \Box \bot$ and $D_2 = \top$. Thus $D = B(\Box C_1(\Box \bot), \Box C_2(\top)) = \Box \Box \bot \rightarrow \Box \neg \top$. The equivalent formula $\Box \Box \bot \rightarrow \Box \bot$ is also a suitable fixed-point.

Example 3 Let $A(p) = \neg \Box \neg p \rightarrow (q \wedge \neg \Box (p \rightarrow q))$. The reader may enjoy showing that $\neg \Box \neg \top \rightarrow (q \wedge \neg \Box (\Box \bot \rightarrow q))$ is a fixed-point for A.

The theorem we have referred to as "the fixed-point theorem" admittedly looks more like a uniqueness theorem than an existence theorem. However the following result, which looks more like an existence theorem, actually follows from the fixed-point theorem:

Theorem 2 *Let $A(p)$ be modalized in p, and let D be a fixed-point of A. Then $\vdash_G \boxdot(p \leftrightarrow D) \rightarrow (p \leftrightarrow A)$.*

A semantic proof from the fixed-point theorem of this result can be found in Goldfarb and Gleit [7]. The following theorem, which justifies the use of the term "fixed-point", now follows easily:

Theorem 3 *Let $A(p)$ be modalized in p, and let D be a fixed-point of A. Then $\vdash_G D \leftrightarrow A(D)$.*

Proof: By Lemma 1 and Theorem 2, the result of substituting D for p in $\Box(p \leftrightarrow D) \rightarrow (p \leftrightarrow A)$ is a theorem of G. Thus $\vdash_G \Box(D \leftrightarrow D) \rightarrow (D \leftrightarrow A(D))$. Since $\Box(D \leftrightarrow D)$ is obviously a theorem of G, we have $\vdash_G D \leftrightarrow A(D)$.

REFERENCES

[1] Bernardi, C., "The fixed-point theorem for diagonalizable algebras," *Studia Logica*, vol. 34 (1975), pp. 239–251.

[2] Boolos, G. and R. Jeffrey, *Computability and Logic*, 2nd ed., Chapter 27, Cambridge University Press, Cambridge, 1980.

[3] Boolos, G., "On deciding the truth of certain statements involving the notion of consistency," *The Journal of Symbolic Logic*, vol. 41 (1976), pp. 779–781.

[4] Boolos, G., *The Unprovability of Consistency*, Cambridge University Press, Cambridge, 1979.

[5] Friedman, H., "One hundred and two problems in mathematical logic," *The Journal of Symbolic Logic*, vol. 40 (1975), pp. 113–129.

[6] Gleit, Z., *Box and Bew: Modal Logic and Provability*, Undergraduate Thesis, Philosophy and Mathematics Departments, Harvard University, 1987.

[7] Goldfarb, W. and Z. Gleit, "Characters and fixed points in provability logic," *Notre Dame Journal of Formal Logic*, vol. 31 (1990), pp. xx–xx.

[8] Reidhaar-Olson, L., *A New Proof of the Fixed-Point Theorem of Provability Logic*, Master's Thesis, Mathematics Department, Massachusetts Institute of Technology, 1988.

[9] Sambin, G., "An effective fixed-point theorem in intuitionistic diagonalizable algebras," *Studia Logica*, vol. 35 (1975), pp. 345–361.

[10] Sambin, G. and S. Valentini, "The modal logic of provability. The sequential approach," *Journal of Philosophical Logic*, vol. 11 (1982), pp. 311–342.

[11] Smoryński, C., "Beth's theorem and self-referential sentences," pp. 253–261 in *Logic Colloquium 77*, edited by A. Macintyre, L. Pacholski, and J. Paris, North Holland, Amsterdam, 1977.

[12] Smoryński, C., "Calculating self-referential statements, I: Explicit calculations," *Studia Logica*, vol. 38 (1979), pp. 17–36.

[13] Smoryński, C., *Self-Reference and Modal Logic*, Springer-Verlag, New York, 1985.

[14] Solovay, R., "Provability interpretations of modal logic," *Israel Journal of Mathematics*, vol. 25 (1976), pp. 287–304.

Department of Mathematics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139