Provability as a Modal Operator with the models of PA as the Worlds

by Marcello Herreshoff
marce110@stanford.edu

May 20, 2011

Abstract

This paper introduces a propositional modal model of Gödel-Löb Logic whose worlds are the models of Peano Arithemetic and whose box modality is equivalent to an operator satisfying the Hilbert Bernay's conditions (e.g. provability.) The semantics of this model is extended to public announcement logic, and it is shown that announcing a formula is semantically equivalent to adding it as a new axiom. The graph structure of the model is also explored, and it is shown that if the descendants of a world are well-founded then they have finite depth.

1 Introduction

1.1 Motivation

The work in this paper is greatly inspired by [Boolos], which discusses the connection between modal logic and proof theory, as introduced by Gödel [Gödel, pp.300-303] and developed by Solovay ([Solovay]), Artemov ([Artemov]), Boolos, and many others. While Boolos devotes an entire chapter [Boolos, pp.68-78] to discussing the semantics of modal logics, he does not produce a model of the logic of provability for Peano Arithmetic. This paper, while certainly not the first, creates such a modal model, in order to provide opportunities for concrete visualization, to potentially aid exploration of the philosophical view of formal systems as agents with states of knowledge, and to explore its properties.

1.2 Outline

- Section 2 of the paper reviews some preliminary notions from proof theory, model theory, and modal logic that are used later in the paper.
- Section 3 constructs the modal model of worlds are the models of Peano Arithmetic and shows that its modal operator corresponds to provability. The construction works for any operator which satisfies the first two of the Hilbert-Bernays provability conditions.
- Section 4 extends the semantics to public announcement logic.
- Section 5 shows that, if our operator satisfied all three of the Hilbert-Bernays provability conditions, the corresponding accessibility relation must be transitive and satisfy Gödel Löb logic, but not the D-axiom.
- Finally, in section 6, we explore the graph structure of the model; in particular, we prove that if the descendants of a world are well-founded then they have finite depth and that every finite depth is represented in the structure.

2 Preliminaries

A model of first order arithmetic is a tuple $\mathfrak{A} = \langle X, 0^{\mathfrak{A}}, S^{\mathfrak{A}}, +^{\mathfrak{A}}, *^{\mathfrak{A}} \rangle$ where $0^{\mathfrak{A}} \in X$, $S^{\mathfrak{A}}: X \to X$, $+^{\mathfrak{A}}: X \times X \to X$ and $*^{\mathfrak{A}}: X \times X \to X$. The standard model is the tuple $\langle \mathbb{N}, 0, S, +, * \rangle$, where 0 is just the number zero, and S, $+$, and $*$ are the successor, addition and multiplication functions respectively. We denote the standard model by $\mathbb{N}$.

The truth of a sentence of first order arithmetic is defined within such a model in the usual way. The notation "$\vDash_{\mathfrak{A}} \varphi$" means that φ is true in $\mathfrak{A}$ where $\mathfrak{A}$ is a model of first order arithmetic and φ is a formula of first order arithmetic. If $\mathfrak{S}$ is a set of models of first order arithmetic then $\vDash_{\mathfrak{S}} \varphi$ iff $\vDash_{\mathfrak{A}} \varphi$ for all $\mathfrak{A} \in \mathfrak{S}$.

If Γ is a set of formulas of first order arithmetic then $\mathrm{Mod}(\Gamma)$ denotes the class of all models of first order arithmetic in which all the formulas in Γ hold.[1] If Γ is a set of formulas of first order arithmetic, and φ is another formula of first order arithmetic, then $\Gamma \vdash \varphi$ iff φ is provable from Γ using only the usual axioms and rules of inference of first order logic. By the soundness and completeness of first order logic, $\Gamma \vdash \varphi$ iff $\vDash_{\mathrm{Mod}(\Gamma)} \varphi$.

PA denotes the set of axioms of Peano Arithmetic. We let $\ulcorner \varphi \urcorner$ denote the Gödel number of φ in the standard way. $\Box_{\mathrm{PA}}(x)$ denotes a formula which expresses that the formula with Gödel number x is provable in PA. We use $\Box_{\mathrm{PA}}\varphi$ as a shorthand for $\Box_{\mathrm{PA}}(\ulcorner \varphi \urcorner)$.

2.1 Modal Logic

2.1.1 Definitions

A (propositional) modal model $\mathcal{M}$ over a set of propositional variables P is a set of worlds W, a "valuation" relation $V \subseteq W \times P$, which assigns a truth value to each propositional variable at each world, and an "accessibility" relation $R \subseteq W \times W$. The set of formulae in the language of modal logic can be defined recursively as follows: $\psi = \bot | p | \neg\psi_1 | \psi_1 \wedge \psi_2 | \psi_1 \to \psi_2 | \Box\psi_1$, where $p \in P$ and ψ_1, ψ_2 are formulae in the language of modal logic. ($\psi_1 \vee \psi_2$ is taken to be short for $\neg(\neg\psi_1 \wedge \neg\psi_2)$ and $\Diamond\psi_1$ is taken to be short for $\neg\Box\neg\psi_1$.)[2]

The relation $\mathcal{M}, \mathfrak{A} \vDash \psi$ expresses that the formula ψ holds at world $\mathfrak{A}$ of the model $\mathcal{M}$ and is defined in a standard way.

A world $\mathfrak{A}$ is *irreflexive* iff $\langle \mathfrak{A}, \mathfrak{A} \rangle \notin R$.

A world $\mathfrak{A}$ in $\mathcal{M}$ is *deluded about* φ iff $\mathcal{M}, \mathfrak{A} \vDash \Box\varphi$ but $\mathcal{M}, \mathfrak{A} \nvDash \varphi$. It follows that if $\mathfrak{A}$ is deluded about any formula φ then $\mathfrak{A}$ is irreflexive.

If $\mathcal{M} = \langle W, R, V \rangle$, then $\mathfrak{A} \in W$ *is a root* iff $\langle \mathfrak{A}, \mathfrak{B} \rangle \in R$ for every $\mathfrak{B} \in W$.

$\mathfrak{A}$ is a *terminal world* iff $\langle \mathfrak{A}, \mathfrak{B} \rangle \notin R$ for every $\mathfrak{B} \in W$. Another way to define this is that $\mathfrak{A}$ is terminal iff $\mathcal{M}, \mathfrak{A} \vDash \Box\bot$.

$\mathcal{M}$ is *serial* iff it has no terminal worlds.

A set of worlds $S \subseteq W$ in $\mathcal{M}$ is *well-founded* iff there is no infinite chain of worlds $\mathfrak{A}_1, \mathfrak{A}_2, ..., \mathfrak{A}_n, ...$ such that $\mathfrak{A}_i \in S$ and $\langle \mathfrak{A}_i, \mathfrak{A}_{i+1} \rangle \in R$ for all $i \geqslant 1$. We say that $\mathfrak{A}$ is a well founded world if the set S is well founded, where $S = \bigcup_{i \in \mathbb{N}} S_i$ and $S_0 = \{\mathfrak{A}\}$ and $S_{i+1} = S_i \cup \{\mathfrak{B} : \langle \mathfrak{C}, \mathfrak{B} \rangle \in R, \mathfrak{C} \in S_i\}$.

We denote the subset of the worlds of $\mathcal{M}$ which are well-founded by $\mathrm{WF}(\mathcal{M})$.

The *rank* $\mathrm{rk}(\mathfrak{A})$ of a well-founded world $\mathfrak{A}$ is an ordinal defined by transfinite recursion by $\mathrm{rk}(\mathfrak{A}) = \sup\{\mathrm{rk}(\mathfrak{B}) + 1 : \langle \mathfrak{A}, \mathfrak{B} \rangle \in R\}$.

2.1.2 Proof Systems

- The *D-axiom* is defined as $\neg\Box\bot$
- The *negative introspection* schema is defined as $\neg\Box\psi \to \Box\neg\Box\psi$.
- The *K-axiom* schema is defined as $\Box(\psi_1 \to \psi_2) \to (\Box\psi_1 \to \Box\psi_2)$
- The *Löb schema* is $\Box(\Box\psi \to \psi) \to \Box\psi$
- Modus Ponens is the inference rule:

$$\frac{\psi_1;\ \psi_1 \to \psi_2}{\psi_2}$$

1. $\mathrm{Mod}(\Gamma)$ is a proper class, but we could also interpret it as the set of countable models using one particular set of objects X as their universe (which is a set) and all of the results and proofs in this paper would be preserved.

2. $\Box$ is not to be confused with $\Box_{\mathrm{PA}}$. $\Box$ will only appear in modal formulas and $\Box_{\mathrm{PA}}$ will only occur in first order formulas.

- Necessitation is the inference rule:

$$\frac{\psi}{\Box\psi}$$

- *Gödel Löb logic* (GL) is the closure of propositional tautologies, the K-axiom, and the Löb schema under the inference rules of modus ponens and necessitation.

3 The Correspondence Theorem

3.1 The Hilbert Bernays conditions

Suppose $B(x)$ is a formula of first order arithmetic with one free variable. Then B satisfies the Hilbert Bernays provability conditions iff, for aribitrary sentences α, β of first-order arithmetic, we have:

1. $PA \vdash \alpha$ implies $PA \vdash B\alpha$ (B has necessitation)
2. $PA \vdash B(\alpha \to \beta) \to (B\alpha \to B\beta)$ (B satisfies the K axiom)
3. $\mathrm{PA} \vdash B\alpha \to B\,B\alpha$ (B is transitive)

Here, we write $B\varphi$ as a shorthand for $B(\ulcorner\varphi\urcorner)$. For the rest of this paper, we will assume that B is an operator satisfying the Hilbert Bernays conditions.

3.2 Examples of Operators satisfying the Hilbert Bernays conditions

1. The motivating example: $B = \Box_{\mathrm{PA}}$.
2. $\Box_S$, defined in the usual way, where S is a superset of PA.
3. If, S is any system stronger than PA, we can define the n-provability predicate $[n]_S\varphi :=$ $\exists\ulcorner\psi\urcorner$: $\mathrm{True}_{\Pi_n}(\ulcorner\psi\urcorner) \to \Box_S(\psi \to \varphi)$, following [Beklemishev]. According to ([Beklemishev], **Proposition 2.10**, p. 17), this operator satisfies the Hilbert Bernays conditions.

3.3 The Model

Let $\mathcal{L}$ be a language for modal logic with one propositional variable for each sentence of first order arithmetic. That is, the formulas of $\mathcal{L}$ are $\psi = \bot | P_\varphi | \neg\psi_1 | \psi_1 \wedge \psi_2 | \psi_1 \to \psi_2 | \Box\psi_1$ where φ is a formula of first order arithmetic and ψ_1, ψ_2 are formulas of $\mathcal{L}$. These formulas have the usual semantics on modal models.

Let $\mathcal{M}_B = \langle W, R, V\rangle$, where $W = \mathrm{Mod(PA)}$, $V = \{\langle \mathfrak{A}, P_\varphi\rangle | \mathfrak{A} \in W, \vDash_{\mathfrak{A}} \varphi\}$ and finally,

$$R = \{\langle \mathfrak{A}, \mathfrak{B}\rangle | \mathfrak{A}, \mathfrak{B} \in W \texttt{ and for all } \varphi, \ \vDash_{\mathfrak{A}} B\varphi \texttt{ implies } \vDash_{\mathfrak{B}} \varphi\}$$

In other words, each model of PA is a world of our modal mode and the propositional variable P_φ is true in the world $\mathfrak{A}$ iff φ is a true statement about the model $\mathfrak{A}$.

One way to think of this definition of R is that statements of the form $B\varphi$ which are true in $\mathfrak{A}$ constitute our "beliefs" at the world $\mathfrak{A}$. Thus the world $\mathfrak{B}$ appears possible from world $\mathfrak{A}$ iff it conforms to all of $\mathfrak{A}$'s beliefs. In particular, $\mathcal{M}_{\Box_{\mathrm{PA}}}$ is the set of models of $\mathcal{P}\,\mathcal{A}$ considered as a graph, where we draw an arrow from model $\mathfrak{A}$ to model $\mathfrak{B}$ if everything that appears provable in $\mathfrak{A}$ is true in $\mathfrak{B}$.

3.4 The Correspondence between $\Box$ and B

Let the function T which maps formulas of $\mathcal{L}$ to formulas of first order arithmetic, defined inductively by $T(\bot) = \bot$, $T(P_\varphi) = \varphi$, $T(\neg\alpha) = \neg T(\alpha)$, $T(\alpha \wedge \beta) = T(\alpha) \wedge T(\beta)$, $T(\alpha \to \beta) = T(\alpha) \to T(\beta)$, $T(\Box\alpha) = B\,T(\alpha)$. That is, T replaces all variables with the formulas they correspond to, and all $\Box$'s with B's.

Main Theorem: For any formula ψ of $\mathcal{L}$, we have $\mathcal{M}_B, \mathfrak{A} \vDash \psi$ iff $\vDash_{\mathfrak{A}} T(\psi)$ for any world $\mathfrak{A} \in W$, provided that B satisfies the first two of the Hilbert Bernay's conditions.

Proof: We proceed by induction on ψ.

1. If $\psi = \bot$ then $\mathcal{M}_B, \mathfrak{A} \not\models \bot$ and $\not\models_{\mathfrak{A}} \bot$. Thus $\mathcal{M}_B, \mathfrak{A} \models \psi$ iff $\models_{\mathfrak{A}} \psi$ as desired.
2. Suppose $\psi = P_\varphi$. Then $\mathcal{M}_B, \mathfrak{A} \models P_\varphi$ iff $V(\mathfrak{A}, P_\varphi)$ iff $\models_{\mathfrak{A}} \varphi$. But $\varphi = T(P_\varphi)$, so $\mathcal{M}_B, \mathfrak{A} \models P_\varphi$ iff $\models_{\mathfrak{A}} T(P_\varphi)$, as desired.
3. Suppose $\psi = \neg\alpha$, and by hypothesis $\mathcal{M}, \mathfrak{A} \models \alpha$ iff $\models_{\mathfrak{A}} T(\alpha)$. Then $\mathcal{M}_B, \mathfrak{A} \models \neg\alpha$ iff $\mathcal{M}_\mathcal{B}$, $\mathfrak{A} \not\models \alpha$ iff $\not\models_{\mathfrak{A}} T(\alpha)$ iff $\models_{\mathfrak{A}} \neg T(\alpha)$ iff $\models_{\mathfrak{A}} T(\neg\alpha)$, as desired.
4. Suppose $\psi = \alpha \wedge \beta$ and by hypothesis $\mathcal{M}_B, \mathfrak{A} \models \alpha$ iff $\models_{\mathfrak{A}} T(\alpha)$ and $\mathcal{M}, \mathfrak{A} \models \beta$ iff $\models_{\mathfrak{A}} T(\beta)$. Then $\mathcal{M}_B, \mathfrak{A} \models \alpha \wedge \beta$ iff $\mathcal{M}_B, \mathfrak{A} \models \alpha$ and $\mathcal{M}_B, \mathfrak{A} \models \beta$ iff $\models_{\mathfrak{A}} T(\alpha)$ and $\models_{\mathfrak{A}} T(\beta)$ iff $\models_{\mathfrak{A}} T(\alpha) \wedge T(\beta)$ iff $\models_{\mathfrak{A}} T(\alpha \wedge \beta)$.
5. Suppose $\psi = \alpha \to \beta$ and by hypothesis $\mathcal{M}_B, \mathfrak{A} \models \alpha$ iff $\models_{\mathfrak{A}} T(\alpha)$ and $\mathcal{M}, \mathfrak{A} \models \beta$ iff $\models_{\mathfrak{A}} T(\beta)$. Then $\mathcal{M}_B, \mathfrak{A} \models \alpha \to \beta$ iff $\mathcal{M}, \mathfrak{A} \models \alpha$ implies $\mathcal{M}_B, \mathfrak{A} \models \beta$ iff $\models_{\mathfrak{A}} T(\alpha)$ implies $\models_{\mathfrak{A}} T(\beta)$ iff $\models_{\mathfrak{A}} T(\alpha) \to T(\beta)$ iff $\models_{\mathfrak{A}} T(\alpha \to \beta)$.
6. Suppose $\psi = \Box\alpha$ and and by hypothesis $\mathcal{M}_B, \mathfrak{B} \models \alpha$ iff $\models_{\mathfrak{B}} T(\alpha)$ for any world $\mathfrak{B} \in W$.
   a. Suppose that $\models_{\mathfrak{A}} T(\Box\alpha)$, that is $\models_{\mathfrak{A}} B\,T(\alpha)$. For any $\langle \mathfrak{A}, \mathfrak{B} \rangle \in R$, $\models_{\mathfrak{A}} B\,T(\alpha)$ yeilds $\models_{\mathfrak{B}} T(\alpha)$ and thus $\mathcal{M}, \mathfrak{B} \models \alpha$. Therefore $\mathcal{M}_B, \mathfrak{A} \models \Box\alpha$.
   b. Now we get to the interesting case:
      Suppose $\mathcal{M}_B, \mathfrak{A} \models \Box\alpha$. This means that $\mathcal{M}_B, \mathfrak{B} \models \alpha$ and thus $\models_{\mathfrak{B}} T(\alpha)$ for all $\mathfrak{B}$ such that $\langle \mathfrak{A}, \mathfrak{B} \rangle \in R$. Now the set of worlds $\{v\colon \langle \mathfrak{A}, \mathfrak{B} \rangle \in R\}$ is precisely $\mathrm{Mod}(PA \cup \{\varphi\colon \models_{\mathfrak{A}} B\varphi\})$. Because $T(\alpha)$ is true in all models of $PA \cup \{\varphi\colon \models_{\mathfrak{A}} B\varphi\}$, it follows by the completeness theorem for first order logic that $P\,A \cup \{\varphi\colon \models_{\mathfrak{A}} B\varphi\} \vdash T(\alpha)$. Because the proof must be finite, there must be a finite subset $\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_n\} \subseteq \{\varphi\colon \models_{\mathfrak{A}} B\varphi\}$, such that $PA; \gamma_1, ..., \gamma_n \vdash T(\alpha)$. By repeatedly using the deduction theorem, we deduce that $PA \vdash \gamma_1 \to (\gamma_2 \to \cdots(\gamma_n \to T(\alpha)))$. Then, by the first of the Hilbert Bernay's conditions (necessitation), $P\,A \vdash B[\gamma_1 \to (\gamma_2 \to \cdots(\gamma_n \to T(\alpha)))]$. In particular, $\models_{\mathfrak{A}} B[\gamma_1 \to (\gamma_2 \to \cdots(\gamma_n \to T(\alpha)))]$, because $\mathfrak{A} \in \mathrm{Mod}(\mathrm{PA})$.
      Note next that by the second of the Hilbert Bernay's conditions (K axiom) if we have $\models_{\mathfrak{A}} B\psi_1$ and $\models_{\mathfrak{A}} B(\psi_1 \to \psi_2)$, then we have $\models_{\mathfrak{A}} B\psi_2$. Because $\models_{\mathfrak{A}} B[\gamma_1 \to (\gamma_2 \to \cdots(\gamma_n \to T(\alpha)))]$ and $\models_{\mathfrak{A}} B\gamma_j$ for all j, we have $\models_{\mathfrak{A}} B\,T(\alpha)$ as desired.

4 Extension to Public Announcement Logic

We extend the language $\mathcal{L}$ to a language $\mathcal{L}'$ whose formulas are of the form $\psi = \bot | P_\varphi | \neg\psi_1 | \psi_1 \wedge \psi_2 | \Box\psi_1 | [\psi_1]\psi_2$, where φ is a formula of first order logic and ψ_1, ψ_2 are formulas of $\mathcal{L}'$. Here, the $[\alpha]\beta$ operator is taken to have the usual semantics [van Ditmarsch, p. 74] that is:

$\mathcal{M}, \mathfrak{A} \models [\alpha]\beta$ iff $\mathcal{M}, \mathfrak{A} \models \alpha$ implies $\mathcal{M}|_\alpha, \mathfrak{A} \models \beta$.

Here, if $\mathcal{M} = \langle W, R, V \rangle$, then $\mathcal{M}|_\alpha = \langle W', R', V' \rangle$ where $W' = \{\mathfrak{A} \in W\colon \mathcal{M}, \mathfrak{A} \models \alpha\}$, $V' = \{\langle \mathfrak{A}, p \rangle\colon \langle \mathfrak{A}, p \rangle \in V, \mathfrak{A} \in W'\}$ and $R' = \{\langle \mathfrak{A}, \mathfrak{B} \rangle\colon \langle \mathfrak{A}, \mathfrak{B} \rangle \in R, \mathfrak{A} \in W', \mathfrak{B} \in W'\}$.

The main result of this section is that formulas in $\mathcal{L}'$ can be translated to equivalent formulas in $\mathcal{L}$. (In the sense that $\psi_1 \cong_\mathcal{M} \psi_2$ is defined as $\mathcal{M}, \mathfrak{A} \models \psi_1$ iff $\mathcal{M}, \mathfrak{A} \models \psi_2$ for any $\mathfrak{A} \in |\mathcal{M}|$.)

Note first that if α does not contain brackets then $\mathcal{M}, \mathfrak{A} \models \alpha$ iff $\models_{\mathfrak{A}} T(\alpha)$ iff $\mathcal{M}, \mathfrak{A} \models P_{T(\alpha)}$. It follows that $\mathcal{M}|_{P_{T(\alpha)}} = \mathcal{M}|_\alpha$ and therefore that $\mathcal{M}, \mathfrak{A} \models [\alpha]\beta$ iff $\mathcal{M}, \mathfrak{A} \models \left[P_{T(\alpha)}\right]\beta$.

Definition: Let $S_p(\psi)$ be defined as follows: $S_p(\bot) = \bot$, $S_p(P_\varphi) = P_\varphi$, $S_v(\neg\alpha) = \neg S_p(\alpha)$, $S_p(\alpha \wedge \beta) = S_p(\alpha) \wedge S_p(\beta)$, $S_p(\Box\alpha) = \Box(p \to S_p(\alpha))$, and $S_p([\alpha]\beta) = [S_p(\alpha)]S_p(\beta)$.

Lemma: $\mathcal{M}, \mathfrak{A} \models [p]\psi$ iff $\mathcal{M}, \mathfrak{A} \models p \to S_p(\psi)$ where p is any proposition.

Proof: If $\mathcal{M}, \mathfrak{A} \not\models p$ then $\mathcal{M}, \mathfrak{A} \models [p]\psi$ and $\mathcal{M}, \mathfrak{A} \models p \to S_p(\psi)$, as desired. Now suppose that $\mathcal{M}, \mathfrak{A} \models p$, in which case we clearly have $\mathcal{M}, \mathfrak{A} \models p \to S_p(\psi)$ iff $\mathcal{M}, \mathfrak{A} \models S_p(\psi)$. Next I will prove by induction on ψ that for any submodel $\mathcal{N} \subseteq \mathcal{M}$ and any world $\mathfrak{A} \in \mathcal{N}$ such that $\mathcal{N}, \mathfrak{A} \models p$, we have $\mathcal{N}|_p, \mathfrak{A} \models \psi$ iff $\mathcal{N}, \mathfrak{A} \models S_p(\psi)$:

1. The case $\psi = \bot$. Then trivially $\mathcal{N}|_p, \mathfrak{A} \models \bot$ iff $\mathcal{N}, \mathfrak{A} \models \bot$ iff $\mathcal{N}, \mathfrak{A} \models S_p(\bot)$.
2. Suppose $\psi = P_\varphi$. Then trivially $\mathcal{N}|_p, \mathfrak{A} \models P_\varphi$ iff $\mathcal{N}, \mathfrak{A} \models P_\varphi$ iff $\mathcal{N}, \mathfrak{A} \models S_p(P_\varphi)$.

3. Suppose $\psi = \neg\alpha$ and by hypothesis that $\mathcal{N}|_p, \mathfrak{A} \vDash \alpha$ iff $\mathcal{N}, \mathfrak{A} \vDash S_p(\alpha)$. Thus $\mathcal{N}|_p, \mathfrak{A} \vDash \neg\alpha$ iff $\mathcal{N}|_p, \mathfrak{A} \nvDash \alpha$ iff $\mathcal{N}, \mathfrak{A} \nvDash S_p(\alpha)$ iff $\mathcal{N}, \mathfrak{A} \vDash \neg S_p(\alpha)$ iff $\mathcal{N}, \mathfrak{A} \vDash S_p(\neg\alpha)$, as desired.
4. Suppose $\psi = \alpha \wedge \beta$ and by hypothesis that $\mathcal{N}|_p$, $\mathfrak{A} \vDash \alpha$ iff $\mathcal{N}$, $\mathfrak{A} \vDash S_p(\alpha)$ and that $\mathcal{N}|_p$, $\mathfrak{A} \vDash \beta$ iff $\mathcal{N}, \mathfrak{A} \vDash S_p(\beta)$. Thus $\mathcal{N}|_p, \mathfrak{A} \vDash \alpha \wedge \beta$ iff $\mathcal{N}|_p, \mathfrak{A} \vDash \alpha$ and $\mathcal{N}|_p, \mathfrak{A} \vDash \beta$ iff $\mathcal{N}, \mathfrak{A} \vDash S_p(\alpha)$ and $\mathcal{N}, \mathfrak{A} \vDash S_p(\beta)$ iff $\mathcal{N}, \mathfrak{A} \vDash S_p(\alpha) \wedge S_p(\beta)$ iff $\mathcal{N}, \mathfrak{A} \vDash S_p(\alpha \wedge \beta)$.
5. Suppose $\psi = \alpha \to \beta$ and by hypothesis that $\mathcal{N}|_p$, $\mathfrak{A} \vDash \alpha$ iff $\mathcal{N}$, $\mathfrak{A} \vDash S_p(\alpha)$ and that $\mathcal{N}|_p$, $\mathfrak{A} \vDash \beta$ iff $\mathcal{N}$, $\mathfrak{A} \vDash S_p(\beta)$. Thus $\mathcal{N}|_p$, $\mathfrak{A} \vDash \alpha \to \beta$ iff $\mathcal{N}|_p$, $\mathfrak{A} \vDash \alpha$ implies $\mathcal{N}|_p$, $\mathfrak{A} \vDash \beta$ iff $\mathcal{N}$, $\mathfrak{A} \vDash S_p(\alpha)$ implies $\mathcal{N}, \mathfrak{A} \vDash S_p(\beta)$ iff $\mathcal{N}, \mathfrak{A} \vDash S_p(\alpha) \to S_p(\beta)$ iff $\mathcal{N}, \mathfrak{A} \vDash S_p(\alpha \to \beta)$.
6. Suppose $\psi = \Box\alpha$ and by hypothesis that $\mathcal{N}|_p$, $\mathfrak{B} \vDash \alpha$ iff $\mathcal{N}$, $\mathfrak{B} \vDash S_p(\alpha)$ for any $\mathfrak{B} \in |\mathcal{N}|$, where $V(\mathfrak{B}, p)$ (is true.) Thus:
   - $\mathcal{N}|_p, \mathfrak{A} \vDash \Box\alpha$ iff
   - $\mathcal{N}|_p, \mathfrak{B} \vDash \alpha$ for all $\mathfrak{B}$ such that $V(\mathfrak{B}, p)$ and $\langle \mathfrak{A}, \mathfrak{B} \rangle \in R$ iff
   - $\mathcal{N}, \mathfrak{B} \vDash S_p(\alpha)$ for all $\mathfrak{B}$ such that $V(\mathfrak{B}, p)$ and $\langle \mathfrak{A}, \mathfrak{B} \rangle \in R$ iff
   - $\mathcal{N}, \mathfrak{B} \vDash p \to S_p(\alpha)$ for all $\mathfrak{B}$ such that $\langle \mathfrak{A}, \mathfrak{B} \rangle \in R$ iff
   - $\mathcal{N}, \mathfrak{A} \vDash \Box(p \to S_p(\alpha))$ iff
   - $\mathcal{N}, \mathfrak{A} \vDash S_p(\Box\alpha)$
7. Finally suppose $\psi = [\alpha]\beta$ and by hypotehsis that $\mathcal{N}|_p, \mathfrak{B} \vDash \alpha$ iff $\mathcal{N}, \mathfrak{B} \vDash S_p(\alpha)$ for any $\mathfrak{B} \in |\mathcal{N}|$, where $V(\mathfrak{B}, p)$ (is true.) and that for any submodel $\mathcal{N}' \subseteq \mathcal{N}$, we have $\mathcal{N}'|_p, \mathfrak{B} \vDash \alpha$ iff $\mathcal{N}', \mathfrak{B} \vDash S_p(\alpha)$ for any $\mathfrak{B} \in |\mathcal{N}'|$, where $V(\mathfrak{B}, p)$. Thus:
   - $\mathcal{N}|_p, \mathfrak{A} \vDash [\alpha]\beta$ iff
   - $\mathcal{N}|_p, \mathfrak{A} \vDash \alpha$ implies $\mathcal{N}|_p|_\alpha, \mathfrak{A} \vDash \beta$
   - Note: for any $\mathfrak{B} \in \mathcal{N}|_p$, $\mathcal{N}|_p$, $\mathfrak{B} \vDash \alpha$ iff $\mathcal{N}$, $\mathfrak{B} \vDash S_p(\alpha)$. Thus $\mathfrak{B} \in \mathcal{N}|_p|_\alpha$ iff $\mathfrak{B} \in \mathcal{N}|_{S_p(\alpha)}$ and $V(\mathfrak{B}, p)$ iff $\mathfrak{B} \in \mathcal{N}|_{S_p(\alpha)}|_p$, and so the previous bullet holds iff:
   - $\mathcal{N}|_p, \mathfrak{A} \vDash \alpha$ implies $\mathcal{N}|_{S_p(\alpha)}|_p, \mathfrak{A} \vDash \beta$, iff
   - $\mathcal{N}, \mathfrak{A} \vDash S_p(\alpha)$ implies $\mathcal{N}|_{S_p(\alpha)}|_p, \mathfrak{A} \vDash \beta$, iff
   - $\mathcal{N}, \mathfrak{A} \vDash S_p(\alpha)$ implies $\mathcal{N}|_{S_p(\alpha)}, \mathfrak{A} \vDash S_p(\beta)$, iff
   - $\mathcal{N}, \mathfrak{A} \vDash [S_p(\alpha)]S_p(\beta)$

And now that the lemma has been proven, it follows straightfowardly that for any $\mathfrak{A} \in |\mathcal{M}|$ we have $\mathcal{M}, \mathfrak{A} \vDash [\alpha]\beta$ iff $\mathcal{M}, \mathfrak{A} \vDash P_{T(\alpha)} \to S_{P_{T(\alpha)}}(\beta)$ (provided α contains no brackets.)

This, however provides us with a way to remove all the brackets from any formula. If $\psi_1 \cong_{\mathcal{M}} \psi_2$ then $\neg\psi_1 \cong_{\mathcal{M}} \neg\psi_2$ and $\alpha \wedge \psi_1 \cong_{\mathcal{M}} \alpha \wedge \psi_2$ and $\psi_1 \wedge \alpha \cong_{\mathcal{M}} \psi_2 \wedge \alpha$ and $\alpha \to \psi_1 \cong_{\mathcal{M}} \alpha \to \psi_2$ and $\psi_1 \to \alpha \cong_{\mathcal{M}} \psi_2 \to \alpha$ and $\Box\psi_1 \cong_{\mathcal{M}} \Box\psi_2$ and $[\psi_1]\alpha \cong_{\mathcal{M}} [\psi_2]\alpha$. This implies that as long as a sub-formula is not on the right hand side of an announement operator, it can be replaced with an equivalent sub-formula not containing one less pair of brackets.

However, every formula containing announcement operators must contain at least one such operator $[\alpha]\beta$ which is not on the right side of any other announcement operators and such that α contains no announcement operators. Thus the reduction can be applied, and since the reduction always eliminates one pair of square brackets, all of the announcement operators can be eliminated. Another way to think about this is that we can extend the mapping T to include a case $T([\alpha]\beta) = P_{T(\alpha)} \to T\big(S_{P_{T(\alpha)}}(\beta)\big)$.

Inituitively, the upshot is that an expression like $[\alpha]\beta$ is true whenever α is false, and when α is true, the $[\alpha]$ reinterprets all the $\Box$'s within β to mean "provable assuming $T(\alpha)$" or "provable in the new formal system that has $T(\alpha)$ as an additional axiom" rather than just "provable."

5 $\mathcal{M}_B$ satisfies Gödel Löb logic

Theorem 1: $\mathcal{M}_B$ has a central world at $\mathbb{N}$ iff B is equivalent to $\Box_{\mathrm{PA}}$.

Proof: $\mathcal{M}_B$ has a root at $\mathbb{N}$ iff $\langle \mathbb{N}, \mathfrak{B}\rangle \in R$ for every $\mathfrak{B} \in W$. Now the set $\{\mathfrak{B}\colon \langle \mathbb{N}, \mathfrak{B}\rangle \in R\} = \mathrm{Mod}(\mathrm{PA} \cup \{\varphi\colon \vDash_{\mathfrak{A}} B\varphi\})$, so $\mathcal{M}_B$ has a central world $\mathfrak{A}$ iff $\mathrm{Mod}(\mathrm{PA}) = \mathrm{Mod}(\mathrm{PA} \cup \{\varphi\colon \vDash_{\mathfrak{A}} B\varphi\})$. Now if B is equivalent to $\Box_{\mathrm{PA}}$, then $\{\varphi\colon \vDash_{\mathbb{N}} B\varphi\} = \mathrm{Th}(\mathrm{PA})$, then $\mathrm{Mod}(\mathrm{PA}) = \mathrm{Mod}(\mathrm{PA} \cup \{\varphi\colon \vDash_{\mathbb{N}} B\varphi\})$ and $\mathbb{N}$ is central. If on the other hand we B not equivalent to $\Box_{\mathrm{PA}}$ we can first rule out the case where $\{\varphi\colon \vDash_{\mathfrak{A}} B\varphi\} \subset \mathrm{Th}(\mathrm{PA})$ by the first property of B. From this we conclude that $\{\varphi\colon \vDash_{\mathfrak{A}} B\varphi\} \not\subseteq \mathrm{Th}(\mathrm{PA})$, that is, that there's some φ such that $\vDash_{\mathfrak{A}} B\varphi$ but $\varphi \notin \mathrm{Th}(\mathrm{PA})$, however, this would imply that $\mathrm{PA}; \neg\varphi$ is consistent, in which case there's some $\mathfrak{B} \in W$ where $\langle \mathbb{N}, \mathfrak{B}\rangle \notin R$, and hence $\mathbb{N}$ is not central.

Theorem 2: $\mathcal{M}_B$ is transitive.

Proof: If $\langle \mathfrak{A}, \mathfrak{B}\rangle \in R$ and $\vDash_{\mathfrak{A}} B\varphi$, then $\vDash_{\mathfrak{A}} B\,B\varphi$ and so $\mathcal{M}_B, \mathfrak{A} \vDash \Box P_{B\varphi}$, from which it follows that $\mathcal{M}_B, \mathfrak{B} \vDash P_{B\varphi}$ and therefore $\vDash_{\mathfrak{B}} B\varphi$. Thus $\{\varphi\colon \vDash_{\mathfrak{A}} B\varphi\} \subseteq \{\varphi\colon \vDash_{\mathfrak{B}} B\varphi\}$, and therefore $\{\mathfrak{C}\colon \langle \mathfrak{A}, \mathfrak{C}\rangle \in R\}] = \mathrm{Mod}(\mathrm{PA} \cup \{\varphi\colon \vDash_{\mathfrak{A}} B\varphi\}) \supseteq \mathrm{Mod}(\mathrm{PA} \cup \{\varphi\colon \vDash_{\mathfrak{B}} B\varphi\}) = \{\mathfrak{C}\colon \langle \mathfrak{B}, \mathfrak{C}\rangle \in R\}$.

Theorem 3: $\mathcal{M}_B, \mathfrak{A} \vDash \Box(\Box\psi \to \psi) \to \Box\psi$ for all worlds $\mathfrak{A}$ and arbitrary modal formulae ψ.

Proof: B is Löbian so all three of Löb's conditions hold. Furthermore, the diagonalization lemma applies to any formula B of first order arithmetic, so we can construct γ such that $\mathrm{PA} \vdash \gamma \leftrightarrow [B\gamma \to \varphi]$. These are all the ingredients required for Löb's theorem to hold, and it follows that $\mathrm{PA} \vdash B(B\varphi \to \varphi) \to B\varphi$. If we choose $\varphi = T(\psi)$ then it follows that $\mathcal{M}_B, \mathfrak{A} \vDash \Box(\Box\psi \to \Box) \to \Box\psi$ for all worlds $\mathfrak{A}$ and arbitrary modal formulae ψ.

Theorem 4: If $\mathcal{M}_B$ is non-terminal at some world $\mathfrak{A}$ then $\mathcal{M}_B$ has no negative-introspection.

Proof: By theorem 3 $\mathcal{M}_B, \mathfrak{A} \vDash \Box(\Box\bot \to \bot) \to \Box\bot$. Because $\mathfrak{A}$ is non-terminal, the consequent is false, and therefore $\mathcal{M}_B, \mathfrak{A} \vDash \neg\Box(\neg\Box\bot)$. However, because $\mathcal{M}_B, \mathfrak{A} \vDash \neg\Box\bot$ and $\mathcal{M}_B, \mathfrak{A} \nvDash \Box(\neg\Box\bot)$, negative introspection does not hold at $\mathfrak{A}$

Theorem 5: $\mathcal{M}_B$ does not satisfy the D-axiom (i.e. $\mathcal{M}_B$ does not have seriality.)

Case 1: Suppose world $\mathfrak{A}$ is terminal. Then $\mathcal{M}_B, \mathfrak{A} \vDash \Box\bot$, violating the D-axiom.

Case 2: Suppose world $\mathfrak{A}$ is non-terminal Then, by the reasoning in the previous theorem $\mathcal{M}_B, \mathfrak{A} \vDash \neg\Box\neg(\Box\bot)$. However, the statement $\mathcal{M}_B, \mathfrak{A} \vDash \neg\Box\neg(\Box\bot)$, implies that there's some world $\mathfrak{B}$ where $\langle \mathfrak{A}, \mathfrak{B}\rangle \in R$ where $\mathcal{M}_B, \mathfrak{B} \vDash \Box\bot$ again violating the D-axiom.

Corrolary 6: $\mathcal{M}_B$ is a model of Gödel Löb logic.

6 The Rank Function on $\mathcal{M}_B$

If we consider the pointed submodels of $\mathcal{M}$ (which include some node $\mathfrak{A}$ and all of its decendants) which such submodels are isomorphic to each other (in the sense of graph structure alone), how many isomorphism classes are there? We know that if B is sound and satisfies the Hilbert Bernays conditions, there are at least two classes because $\langle \mathbb{N}, \mathbb{N}\rangle \in R$ but $\vDash_{\mathfrak{A}} \Box_{\mathrm{PA}}\bot$ for some $\mathfrak{A}$, and thus $\langle \mathfrak{A}, \mathfrak{A}\rangle \notin R$. We can, in fact prove that there are infinitely many, because there are infinitely many ranks.

Lemma: The set $\{\mathtt{rk}(\mathfrak{B})\colon \mathfrak{B} \in \mathrm{WF}(\mathcal{M}_B)\}$ is an initial segment of the ordinals; that is, if $\alpha < \beta$ and there exists a world of rank β then there is also a well-founded world of rank α.

Proof: Let $Q = \{\mathtt{rk}(\mathfrak{B})\colon \mathfrak{B} \in \mathrm{WF}(\mathcal{M}_B), \mathtt{rk}(\mathfrak{B}) \geq \alpha\}$ Because Q is a set of ordinals, which is non-empty because there is some well-founded world of rank $\beta > \alpha$, $\min Q$ is well defined. Suppose for contradiction that $\min Q > \alpha$. Then there exists a world $\mathfrak{B}$ with $\mathtt{rk}(\mathfrak{B}) = \min Q$. However, by definition of $\mathtt{rk}$, there must exist a world $\mathfrak{C}$ where $\langle \mathfrak{B}, \mathfrak{C}\rangle \in R$ such that $\alpha \leq \mathtt{rk}(\mathfrak{C}) < \mathtt{rk}(\mathfrak{B})$, otherwise $\mathtt{rk}(\mathfrak{B}) \leq \alpha$. Since $\mathfrak{B} \in \mathrm{WF}(\mathcal{M}_B)$, $\mathfrak{C} \in \mathrm{WF}(\mathcal{M}_B)$, and thus $\mathtt{rk}(\mathfrak{C}) \in Q$, contradicting $\mathtt{rk}(\mathfrak{B}) = \min Q$.

Theorem 1: If B is sound and Löbian, then, for any $n \in \omega$, there exists a world with rank n.

Proof: Observe first If B is sound, and $\vDash_{\mathbb{N}} \neg\varphi$ then $\vDash_{\mathbb{N}} \neg B\varphi$. This is just the contrapositive of the definition of soundness.

- **Lemma 1a:** (Löb's theorem converse) If B is sound and satisfies the Hilbert Bernays conditions, and $\vDash_{\mathbb{N}} \neg\varphi$ then there is a world which is deluded about φ.

Proof: Because B satisfies the Hilbert Bernays conditions, it follows that $\vDash_{\mathbb{N}} B(B\varphi \to \varphi) \to B\varphi$. Because B is sound and $\vDash_{\mathbb{N}} \neg\varphi$ it follows that $\vDash_{\mathbb{N}} \neg B\varphi$, and therefore, by contraposition $\vDash_{\mathbb{N}} \neg B(B\varphi \to \varphi)$. By the correspondence theorem $\mathcal{M}, \mathbb{N} \vDash \neg\Box(_\varphi \to P_\varphi)$ that is $\mathcal{M}, \mathbb{N} \vDash \Diamond(_\varphi \wedge \neg P_\varphi)$. Therefore, $\mathcal{M}, \mathfrak{A} \vDash_\varphi \wedge \neg P_\varphi$ for some $\mathfrak{A}$ and hence $\vDash_{\mathfrak{A}} B\varphi \wedge \neg\varphi$. Thus world $\mathfrak{A}$ is deluded about φ.

Now using our observation, we see that $\vDash_{\mathbb{N}} \neg\bot$, and therefore, by induction $\vDash_{\mathbb{N}} \neg B^n\bot$, where $B^n\alpha$ is shorthand for $\overbrace{B\,B\cdots B}^{n}\,\alpha$. By our lemma it follows that, for arbitrary n there must exist a world $\mathfrak{A}$ such that $\vDash_{\mathfrak{A}} \neg B^n\bot$ and $\vDash_{\mathfrak{A}} B^{n+1}\bot$, or, using the correspondence theorem $\mathcal{M}, \mathfrak{A} \vDash \neg\Box^n\bot$, and $\mathcal{M}, \mathfrak{A} \vDash \Box^{n+1}\bot$. The first of these can be rewritten as $\mathcal{M}, \mathfrak{A} \vDash \Diamond^n\top$.

- **Lemma 1b:** (Upper Bound) If $\mathcal{M}, \mathfrak{A} \vDash \Box^n\bot$, then $\mathfrak{A} \in \mathrm{WF}(\mathcal{M})$ and $\mathrm{rk}(\mathfrak{A}) < n$.
  Proof: We proceed by induction. For our base case, lemma is vacuously true when $n = 0$ because $\mathcal{M}, \mathfrak{A} \vDash \bot$ never happens. Now, for the inductive step, suppose the lemma holds for n, and we know that $\mathcal{M}, \mathfrak{A} \vDash \Box^{n+1}\bot$. It follows that if $\langle \mathfrak{A}, \mathfrak{B} \rangle \in R$, then $\mathcal{M}$, $\mathfrak{B} \vDash \Box^n\bot$, and therefore, by inductive hypothesis $\mathrm{rk}(\mathfrak{B}) < n$ and $\mathfrak{B}$ is well-founded. However, because all of its decendants are well-founded, $\mathfrak{A}$ is also well-founded. By the definition of rank, $\mathrm{rk}(\mathfrak{A}) < n + 1$.
- **Lemma 1c:** (Lower Bound) If $\mathcal{M}, \mathfrak{A} \vDash \Diamond^n\top$ then $\mathrm{rk}(\mathfrak{A}) \geqslant n$ provided $\mathfrak{A} \in \mathrm{WF}(\mathcal{M})$
  Proof: Again, we proceed by induction. The base case is trivially true because if rk is well defined at $\mathfrak{A}$ then $\mathrm{rk}(\mathfrak{A}) \geqslant 0$ because every ordinal is at least zero. Now suppose the lemma holds for n, and we know that $\mathcal{M}, \mathfrak{A} \vDash \Diamond^{n+1}\top$ and that rk is well-defined at $\mathfrak{A}$. It follows that $\mathfrak{A}$ has a descendant $\mathfrak{B}$ where $\mathcal{M}$, $\mathfrak{B} \vDash \Diamond^n\top$. $\mathfrak{B}$ must also have a well-defined rank because it is a descendant of $\mathfrak{A}$, and therefore by inductive hypothesis $\mathrm{rk}(\mathfrak{B}) \geqslant n$. This forces $\mathrm{rk}(\mathfrak{A}) \geqslant n + 1$ by the definition of rk.

Combining lemmas 1b and 1c, we see that $\mathrm{rk}(\mathfrak{A}) < n + 1$ and is well defined and $\mathrm{rk}(\mathfrak{A}) \geqslant n$. Thus $\mathrm{rk}(\mathfrak{A}) = n$. Thus, we have proven that worlds of rank n exist for arbitrary $n \in \omega$.

Theorem 2: There are no worlds of rank greater than or equal to ω (that is, all ranks are integers.)

Proof: Suppose for contradiction that there exists a world of rank at least ω. Then there must exist a world $\mathfrak{A}$ with rank ω. The descendants of $\mathfrak{A}$ are precisely $\mathrm{Mod}(\{\varphi \colon \vDash_{\mathfrak{A}} \varphi\})$. Let $\Gamma_0 = \{\varphi \colon \vDash_{\mathfrak{A}} \varphi\}$. Let $\Gamma_{n+1} = \Gamma_n; \neg B^n\bot$, and $\Gamma_\omega = \bigcup_{n\in\omega} \Gamma_n$. Pick an arbitrary $n \in \mathbb{N}$. Because $\mathfrak{A}$ has rank ω, it must have a descendant $\mathfrak{B}$ with rank greater than n. We know that $\vDash_{\mathfrak{B}} \varphi$ for all $\varphi \in \Gamma_0$ and aditionally that $\vDash_{\mathfrak{B}} \neg B^k\bot$ for $k \leqslant n$. Thus $\mathfrak{B} \in \mathrm{Mod}(\Gamma_n)$ from which we can conclude that Γ_n is consistent. By compactness it follows that Γ_ω is consistent. Thus there exists a $\mathfrak{C} \in \mathrm{Mod}(\Gamma_\omega) \subseteq \mathrm{Mod}(\Gamma_0)$. However, by lemma 1c, $\mathrm{rk}(\mathfrak{C}) \geqslant n$ for arbitrary n, if it is defined. If $\mathrm{rk}(\mathfrak{C})$ is undefined because $\mathfrak{C}$ is not well founded then $\mathfrak{A}$ is not well-founded either, contradicting our assumption that $\mathrm{rk}(\mathfrak{A}) = \omega$. If $\mathrm{rk}(\mathfrak{C})$ is defined then $\mathrm{rk}(\mathfrak{C}) \geqslant \omega$, and therefore, either $\mathrm{rk}(\mathfrak{A})$ is undefined or $\mathrm{rk}(\mathfrak{A}) \geqslant \omega + 1$, which is a contradiction.

7 Bibliography

- Artemov, S. N. *On Modal Logics Axiomatizing Provability*. Math USSR Izvestiya, Vol. 27 (1986), No. 3.
- Boolos, George. *The logic of provability.* Cambridge Univ Pr, 1996.
- Beklemishev, Lev. *Reflection principles and provability algebras in formal Arithmetic.* (D.Sc. Dissertation) Moscow, 1998. Downloaded from:
  `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.5435&rep=rep1&type=pdf`
- Gödel, Kurt. "An interpretation of intuitionistic propositional calculus." *Kurt Gödel, Collected Works, Volume 1*. Ed. Trans. Solomon Feferman, John W. Dawson Jr., Steven C. Kleene, Gregory H. Moore, Robert M. Solovay, Jean van Heijenoort. New York: Oxford University Press, 1986.

- van Ditmarsch, Hans, Wiebe van der Hoek, and Bareld Kooi. *Dynamic Epistemic Logic*. Springer, 2008.
- Solovay, Robert M. *Provability Interpretations of Modal Logic.* Israel Journal of Mathematics, Vol 25, 1976.