INCOMPLETENESS RESULT AND PROVABILITY LOGIC

YANG LIU

ABSTRACT. This notes is an introduction to Gödel's incompleteness results, the Löb theorem, and the arithmetical completeness theorems of Solovay. It reflects the early development of provability logic.

CONTENTS

1. PRELIMINARIES

1.1. **Gödel's Incompleteness Results.** Let $\mathscr{L}_A$ be the standard first-order language of arithmetic, and PA be the first-order theory of *Peano Arithmetic* in $\mathscr{L}_A$, whose intended interpretation is over the universe of natural numbers. The crucial step in Gödel's incompleteness proofs is the *arithmetization* of PA, through the process of which every legitimate syntactic expression, say, α of PA receives a unique *Gödel number*, written $\mathrm{GN}(\alpha)$, whose name within PA is represented by the *numeral* $\underline{\mathrm{GN}(\alpha)}$, also denoted by $\ulcorner\alpha\urcorner$ (see Figure 1.1). With Gödel numbering, various syntactic properties of PA can be described by the corresponding arithmetic properties of the Gödel numbers of the syntactic structures involved. For instance, a proof in PA is a syntactic construction which can be described through Gödel numbering by the arithmetic relation $\mathrm{Pf}(y, x)$, where

$\mathrm{Pf}(y, x)$: y is the Gödel number of a proof of the wff with Gödel number x.

It can be shown that PA is sufficiently strong such that these arithmetic properties – which are number-theoretic characterizations of the syntactic properties of PA

FACULTY OF PHILOSOPHY, UNIVERSITY OF CAMBRIDGE, CAMBRIDGE CB3 9DA, UNITED KINGDOM
E-mail address: liu@yliu.net.

$$\begin{tikzcd}
\alpha \arrow[r, "\text{coded by}"] \arrow[d, dotted, "\ulcorner\cdot\urcorner"'] & \mathrm{GN}(\alpha) \arrow[dl, "\text{named in } \mathscr{L}_A \text{ by}"] \arrow[d, dotted, "\mathrm{Name}(\cdot)"] \\
\ulcorner\alpha\urcorner = \underline{\mathrm{GN}(\alpha)} \arrow[r, "\text{coded by}"'] & \mathrm{GN}(\ulcorner\alpha\urcorner)
\end{tikzcd}$$

FIGURE 1.1. Gödel numbering

– can in turn be *represented* by some wff within PA.[1] It is through this process of arithmetization and representation that PA is said to be able to "handle" its own syntax. In our example, there is a wff $\mathsf{Proof}(y, x)$ that represents the arithmetic relation $\mathrm{Pf}(y, x)$ such that, for any wff φ with Gödel number m, the following holds:

I.1 if n is a Gödel number of a proof of φ, i.e., if $(n, m) \in \mathrm{Pf}$, then

$$\mathbf{PA} \vdash \mathsf{Proof}(\underline{n}, \ulcorner\varphi\urcorner);$$

I.2 if n is not a Gödel number of a proof of φ, i.e., if $(n, m) \notin \mathrm{Pf}$, then

$$\mathbf{PA} \vdash \neg\mathsf{Proof}(\underline{n}, \ulcorner\varphi\urcorner).$$

At the heart of Gödel's proof is a diagonalization argument which relies, among other things, on the following recursive functions:

$\mathrm{Name}(x)$: the Gödel number of the expression which is the numeral of Gödel number x, i.e., for any (Gödel) number n, $\mathrm{Name}(n) = \mathrm{GN}(\underline{n})$;

$\mathrm{Sub}(x, y, z)$: the Gödel number of the result of substituting all free occurrences of the variable (whose Gödel number is x) in the expression (whose Gödel number z) with the term (whose Gödel number is y).

The *diagonal function*, $\mathrm{Diag}(x)$, is the result of substituting all free occurrences of variable ν in the expression whose Gödel number is x with the Gödel number of its own name, i.e., $\mathrm{Name}(x)$:

$$\mathrm{Diag}(x) =_{\mathrm{Df}} \mathrm{Sub}\big(\mathrm{GN}(\nu), \mathrm{Name}(x), x\big)$$

In other words, for any wff $\varphi(\nu)$ with Gödel number m (i.e., $\mathrm{GN}\big(\varphi(\nu)\big) = m\big)$, $\mathrm{Diag}(m)$ is the Gödel number of the expression which is the result of substituting all free occurrences of the variable ν in φ with the name of m in PA, i.e,

$$\mathrm{Diag}(m) = \mathrm{GN}\big(\varphi(\underline{m})\big) = \mathrm{GN}\Big[\varphi\Big(\underline{\mathrm{GN}\big(\varphi(\nu)\big)}\Big)\Big] = \mathrm{GN}\Big[\varphi(\ulcorner\varphi(\nu)\urcorner)\Big].$$

The diagonal function $\mathrm{Disg}(\cdot)$ is represented in PA by a wff $\mathsf{Diag}(x, y)$ such that

II For any wff $\varphi(\nu)$, $\mathbf{PA} \vdash \forall y\Big[\mathsf{Diag}(\ulcorner\varphi(\nu)\urcorner, y) \leftrightarrow y \approx \ulcorner\varphi(\ulcorner\varphi(\nu)\urcorner)\urcorner\Big].$

[1]It is well known that Gödel's results can be given in systems that are weaker than PA, this however does not concern us in this note.

Lemma 1.1 (Gödel-Carnap Fixed-point Theorem)**.** Let $\mathsf{Diag}(x,y)$ represents in PA the diagonal function such that (II) holds. Then, given any wff $\Phi(\nu)$, there exists a sentence η such that

$$\mathbf{PA} \vdash \eta \leftrightarrow \Phi(\ulcorner \eta \urcorner). \tag{1.1}$$

Proof. Let $\varphi(\nu) = \forall y\big[\mathsf{Diag}(\nu,y) \to \Phi(y)\big]$. Define $\eta = \varphi(\ulcorner \varphi(\nu) \urcorner)$, then

$$\eta = \forall y\big[\mathsf{Diag}(\ulcorner \varphi(\nu) \urcorner, y) \to \Phi(y)\big].$$

By (II) we have, $\mathbf{PA} \vdash \forall y\big[\mathsf{Diag}(\ulcorner \varphi(\nu) \urcorner, y) \leftrightarrow y \approx \ulcorner \eta \urcorner\big]$. Note that the following holds for any wffs $\alpha(x,y), \beta(x)$ in predicate calculus:

$$\vdash \forall y\Big[\alpha(t_1,y) \leftrightarrow y = t_2\Big] \to \Big\{\forall y\big[\alpha(t_1,y) \to \beta(y)\big] \leftrightarrow \beta(t_2)\Big\}.$$

Now, let $\alpha(x,y) = \mathsf{Diag}(x,y), \beta(x) = \Phi(\nu), t_1 = \ulcorner \varphi(\nu) \urcorner$ and $t_2 = \ulcorner \eta \urcorner$, we get

$$\mathbf{PA} \vdash \forall y\,[\mathsf{Diag}(\ulcorner \varphi(\nu) \urcorner, y) \to \Phi(y)] \leftrightarrow \Phi(\ulcorner \eta \urcorner),$$

which, by the definition of η, is (1.1). □

Definition 1.2. The wff $\mathsf{Prv}_{\mathbf{PA}}(x)$—which says "$x$ is *provable* in PA"—is defined by:

$$\mathsf{Prv}_{\mathbf{PA}}(x) =_{\mathrm{Df}} \exists y \mathsf{Proof}(y,x). \tag{1.2}$$

We write $\mathsf{Prv}(x)$ for $\mathsf{Prv}_{\mathbf{PA}}(x)$ if there is no chance of confusion.

Now, in the fixed-point Theorem, let $\Phi(x) = \neg\mathsf{Prv}(x)$, we can construct a sentence γ which asserts its own unprovability, that is,

$$\mathbf{PA} \vdash \gamma \leftrightarrow \neg\mathsf{Prv}(\ulcorner \gamma \urcorner). \tag{1.3}$$

Gödel's first incompleteness is proved by showing that, under certain consistency assumptions, neither γ nor $\neg\gamma$ in (1.3) is provable in PA. More precisely, we say a theory $\mathbf{T}$ of arithmetic is *ω-inconsistent* if there is some wff $\psi(x)$ with one free variable such that (i) $\mathbf{T} \vdash \exists x\, \psi(x)$, but (ii) $\mathbf{T} \vdash \neg\psi(\underline{n})$ for all $n \in \mathbb{N}$; $\mathbf{T}$ is said to be *ω-consistent* if it is not ω-inconsistent. It is plain that if $\mathbf{T}$ is inconsistent then it is ω-inconsistent, hence ω-consistency implies consistency (It is known that ω-consistency is strictly stronger than consistency).

Theorem 1.3 (Gödel's First Incompleteness Theorem)**.** Given (I) and (II) above, there exists a (Gödel) sentence γ in $\mathscr{L}_A$ such that

(1) if PA is consistent, then $\mathbf{PA} \nvdash \gamma$;
(2) if PA is ω-consistent, then $\mathbf{PA} \nvdash \neg\gamma$.

Proof. (1) We break the proof of the first part of the first incompleteness theorem into following four steps, which, as we shall see, will be the basis of the second incompleteness theorem.

(a) By (1.3), we have $\mathbf{PA} \vdash \gamma \to \neg\mathsf{Prv}(\ulcorner \gamma \urcorner)$.
(b) Assume, to the contrary, that $\mathbf{PA} \vdash \gamma$, we have $\mathbf{PA} \vdash \neg\mathsf{Prv}(\ulcorner \gamma \urcorner)$.
(c) Further, if $\mathbf{PA} \vdash \gamma$, then there shall be a proof of γ coded by, say, n. They by (I.1), we have $\mathbf{PA} \vdash \mathsf{Proof}(\underline{n}, \ulcorner \gamma \urcorner)$. It follows that $\mathbf{PA} \vdash \exists y \mathsf{Proof}(y, \ulcorner \gamma \urcorner)$, hence, by definition, $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner \gamma \urcorner)$.
(d) (b) and (c) show that $\mathbf{PA} \vdash \gamma$ implies $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner \gamma \urcorner) \wedge \neg\mathsf{Prv}(\ulcorner \gamma \urcorner)$, that is, $\mathbf{PA} \vdash \bot$. Thus, if PA is consistent then $\mathbf{PA} \nvdash \gamma$.

(2) Assume, to the contrary, that $\mathbf{PA} \vdash \neg\gamma$, then, by (1.3), $\mathbf{PA} \vdash \mathsf{Prv}(\gamma)$. That is, by definition, $\mathbf{PA} \vdash \exists y\, \mathsf{Proof}(y, \gamma)$. Then, there must exist a proof n of γ. For, otherwise, by (I.2), $\mathbf{PA} \vdash \neg\mathsf{Proof}(\underline{n}, \ulcorner\gamma\urcorner)$ for all n, PA is ω-inconsistent, a contradiction. But now we have PA proves both γ and $\neg\gamma$, it is hence inconsistent, again, a contradiction. Therefore, $\mathbf{PA} \nvdash \neg\gamma$. □

Observe that the first part of Gödel's first incompleteness theorem says that "if PA is consistent, then γ is not provable in PA." Suppose that *that* statement itself can be proved in PA then it shall be that PA cannot prove that "PA is consistent," i.e., its own consistency. For, otherwise, by modus ponens PA proves "γ is not provable in PA," then, by (1.3), PA proves γ, which contradicts the conclusion of the first part of Theorem 1.3. This leads to Gödel's second incompleteness theorem that *if PA is consistent then it cannot be proved in PA that PA is consistent*.

The proof of Gödel's second incompleteness theorem hence requires to *formalize within* PA the intuitive argument we just gave. The first step is to express within PA what it means to say that PA is consistent. This can be achieved by saying that some sentence is not provable in PA. To anticipate our discussion on provability logic where the atomic sentence $\perp$ which expresses falsehood is taken as a primitive symbol, we take the sentence $\mathsf{Con}_{\mathbf{PA}}$ expressing the consistency of PA to be:

$$\mathsf{Con}_{\mathbf{PA}} =_{\mathrm{Df}} \neg\mathsf{Prv}(\ulcorner\perp\urcorner). \tag{1.4}$$

The second incompleteness theorem is proved if it can be shown within PA that the first part of the first incompleteness theorem holds, that is, if it can be proved in PA

$$\mathbf{PA} \vdash \mathsf{Con}_{\mathbf{PA}} \to \neg\mathsf{Prv}(\ulcorner\gamma\urcorner). \tag{1.5}$$

It is clear that this goal is achieved if we can reconstruct *inside* PA the steps (a)-(d) taken in the proof of the first part of the first incompleteness theorem above. It turns out that this relies on the following notion of provability predicate.[2]

Definition 1.4 (Provability Predicate)**.** Prv is said to be a *provability predicate* in PA if, for any wff α, β, the following *Hilbert-Bernays-Löb* (**HBL**) conditions are satisfied:

HBL1: If $\mathbf{PA} \vdash \alpha$ then $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner\alpha\urcorner)$;
HBL2: $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner\alpha \to \beta\urcorner) \to \big[\mathsf{Prv}(\ulcorner\alpha\urcorner) \to \mathsf{Prv}(\ulcorner\beta\urcorner)\big]$;
HBL3: $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner\alpha\urcorner) \to \mathsf{Prv}\big(\ulcorner\mathsf{Prv}(\ulcorner\alpha\urcorner)\urcorner\big)$.

Theorem 1.5 (Gödel's Second Incompleteness Theorem)**.** Given (II), if Prv is a provability predicate, then, if PA is consistent then the consistency of PA is unprovable in PA.

Proof. As remarked above, it suffices to show (1.5) within PA. We organize the proof in parallel with (a)-(d) in the proof of the first part of the first incompleteness theorem. The goal is to see that (a)-(d) themselves can be given inside PA as (a')-(d').

(a') By the fixed-point theorem we have $\mathbf{PA} \vdash \gamma \to \neg\mathsf{Prv}(\ulcorner\gamma\urcorner)$. Apply HBL1, we have in PA that

$$\mathbf{PA} \vdash \mathsf{Prv}\Big(\ulcorner\gamma \to \neg\mathsf{Prv}(\ulcorner\gamma\urcorner)\urcorner\Big). \tag{1.6}$$

(b') By an instance of HBL2, (1.6) yields that

$$\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner\gamma\urcorner) \to \mathsf{Prv}\big(\ulcorner\neg\mathsf{Prv}(\ulcorner\gamma\urcorner)\urcorner\big).$$

[2]Conditions of similar kind were adopted in Hilbert and Bernays' *Grundlagen der Mathematik*. Löb (1955) however was the first who took the step of stating explicitly the provability conditions in the current form. See Boolos (1995, chapter 2) for detailed proofs that HBL1-3 indeed hold in PA.

(c′) By HBL3, we immediately get

$$\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner\gamma\urcorner) \to \mathsf{Prv}\big(\ulcorner\mathsf{Prv}(\ulcorner\gamma\urcorner)\urcorner\big). \tag{1.7}$$

(d′) Note that in PA the tautology $\mathsf{Prv}(\ulcorner\gamma\urcorner) \to (\neg\mathsf{Prv}(\ulcorner\gamma\urcorner) \to \bot)$ holds, then, by HBL1, we have

$$\mathbf{PA} \vdash \mathsf{Prv}\big(\ulcorner\mathsf{Prv}(\ulcorner\gamma\urcorner) \to (\neg\mathsf{Prv}(\ulcorner\gamma\urcorner) \to \bot)\urcorner\big). \tag{1.8}$$

By HBL2 and (1.4), we have

$$\mathbf{PA} \vdash \mathsf{Prv}\big(\ulcorner\mathsf{Prv}(\ulcorner\gamma\urcorner)\urcorner\big) \to \Big[\mathsf{Prv}\big(\ulcorner\neg\mathsf{Prv}(\ulcorner\gamma\urcorner)\urcorner\big) \to \neg\mathsf{Con}_{\mathbf{PA}}\Big].$$

Truth functionally, this yields

$$\begin{aligned} \mathbf{PA} \vdash &\Big[\mathsf{Prv}(\ulcorner\gamma\urcorner) \to \mathsf{Prv}\big(\ulcorner\mathsf{Prv}(\ulcorner\gamma\urcorner)\urcorner\big)\Big] \to \\ &\Big\{\big[\mathsf{Prv}(\ulcorner\gamma\urcorner) \to \mathsf{Prv}\big(\ulcorner\neg\mathsf{Prv}(\ulcorner\gamma\urcorner)\urcorner\big)\big] \to \big[\mathsf{Prv}(\ulcorner\gamma\urcorner) \to \neg\mathsf{Con}_{\mathbf{PA}}\big]\Big\}. \end{aligned} \tag{1.9}$$

Finally, by (1.7) and (1.8), we have $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner\gamma\urcorner) \to \neg\mathsf{Con}_{\mathbf{PA}}$, that is,

$$\mathbf{PA} \vdash \mathsf{Con}_{\mathbf{PA}} \to \neg\mathsf{Prv}(\ulcorner\gamma\urcorner).$$

This completes the proof of the second incompleteness theorem. □

1.2. **The Löb Theorem.** In the Fixed-Point Theorem, let $\Phi(x) = \mathsf{Prv}(x)$, then we get a sentence ρ which asserts its own provability, that is,

$$\mathbf{PA} \vdash \rho \leftrightarrow \mathsf{Prv}(\ulcorner\rho\urcorner). \tag{1.10}$$

Henkin (1952) raised the question whether or not this sentence ρ itself is provable or independent (in PA). This was answered by Löb (1955) in the following results, which presupposes Prv as a provability predicate satisfying HBL1-3.

Theorem 1.6 (Löb)**.** For any sentence ζ, if $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner\zeta\urcorner) \to \zeta$ then $\mathbf{PA} \vdash \zeta$.

Proof. The following proof is done in PA so we omit the prefix '$\mathbf{PA} \vdash$'. Note that, in Lemma 1.1, let $\Phi(x) = \mathsf{Prv}(x) \to \zeta$, then there exists a sentence α such that

1.	$\alpha \leftrightarrow \big(\mathsf{Prv}(\ulcorner\alpha\urcorner) \to \zeta\big)$	
2.	$\mathsf{Prv}\big(\ulcorner\alpha \to \big(\mathsf{Prv}(\ulcorner\alpha\urcorner) \to \zeta\big)\urcorner\big)$	by the '$\to$' direction of (1) and HBL1
3.	$\mathsf{Prv}(\ulcorner\alpha\urcorner) \to \mathsf{Prv}\big(\ulcorner\mathsf{Prv}(\ulcorner\alpha\urcorner) \to \zeta\urcorner\big)$	(2) and HBL2
4.	$\mathsf{Prv}(\ulcorner\alpha\urcorner) \to \big[\mathsf{Prv}\big(\ulcorner\mathsf{Prv}(\ulcorner\alpha\urcorner)\urcorner\big) \to \mathsf{Prv}(\ulcorner\zeta\urcorner)\big]$	(3) and HBL2
5.	$\mathsf{Prv}(\ulcorner\alpha\urcorner) \to \mathsf{Prv}\big(\ulcorner\mathsf{Prv}(\ulcorner\alpha\urcorner)\urcorner\big)$	HBL3
6.	$\mathsf{Prv}(\ulcorner\alpha\urcorner) \to \mathsf{Prv}(\ulcorner\zeta\urcorner)$	by (4) and (5)
7.	$\mathsf{Prv}(\ulcorner\alpha\urcorner) \to \zeta$	(6) and the assumption $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner\zeta\urcorner) \to \zeta$
8.	α	the '$\leftarrow$' direction of (1) and (7)
9.	$\mathsf{Prv}(\ulcorner\alpha\urcorner)$	HBL1
10.	ζ	by (7), (9) and MP

□

Thus, Löb answered Henkin's question in positive that the sentence that asserts its own provability is indeed provable in PA. Yet the Löb theorem reveals some surprising features of PA. As noted in Boolos (1995, p. 54-5), it seems that what the premise of Theorem 1.6 states—namely the soundness of provability in PA, that is, if ζ is provable then it is true—is something that should hold in all cases regardless whether or not ζ itself is true or false, provable or unprovable. But if we replace ζ with $\mathsf{Con}_{\mathbf{PA}}$, then, by the Löb theorem, we get $\mathbf{PA} \vdash \mathsf{Con}_{\mathbf{PA}}$, which is impossible due to Göde's second incompleteness theorem. This means the premise $\mathsf{Prv}(\zeta) \to \zeta$ does *not* hold, in general. As it is often put, PA is modest in that it asserts its own soundness only for those that can be actually proved in it.

As a matter of fact, it is now known that Löb' theorem is equivalent to the second incompleteness theorem. To see this, note that if PA is consistent, i.e., $\mathbf{PA} \nvdash \bot$, then by Löb Theorem, $\mathbf{PA} \nvdash \mathsf{Prv}(\ulcorner \bot \urcorner) \to \bot$, but this is just to say, $\mathbf{PA} \nvdash \neg\mathsf{Prv}(\ulcorner \bot \urcorner)$, since, for any sentential letter A, $\neg A \equiv A \to \bot$. This shows that the Löb theorem implies the second incompleteness theorem. Conversely, if for some ζ, we have $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner \zeta \urcorner) \to \zeta$ but $\mathbf{PA} \nvdash \zeta$, the latter implies that $\neg\zeta$ is consistent with $\mathbf{PA}$, and hence $\mathbf{PA}^* = \mathbf{PA} + \{\neg\zeta\}$ is consistent iff $\mathbf{PA}$ does not prove ζ. Put in formula, we have that $\mathsf{Con}_{\mathbf{PA}^*} = \neg\mathsf{Prv}_{\mathbf{PA}}(\ulcorner \zeta \urcorner)$. On the other hand, from $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner \zeta \urcorner) \to \zeta$ we get $\mathbf{PA} \vdash \neg\zeta \to \neg\mathsf{Prv}(\ulcorner \zeta \urcorner)$. It follows that $\mathbf{PA} + \{\neg\zeta\} \vdash_{\mathbf{PA}} \neg\mathsf{Prv}(\ulcorner \zeta \urcorner)$, and hence $\mathbf{PA}^* \vdash \mathsf{Con}_{\mathbf{PA}^*}$. This contradicts the second incompleteness theorem. This shows that Gödel's second incompleteness theorem implies the Löb theorem.

2. Provability Predicate as Modality

2.1. **Normal Systems.** Let the modal language $\mathscr{L}_M$ consist of sentences of the form

$$p ::= \top \mid \bot \mid p \mid \neg p \mid p \wedge q \mid p \vee q \mid p \to q \mid \Box p \mid \Diamond p$$

(Occasionally, we take '$\bot$', '$\to$' and '$\Box$' as primary connectives leaving other connectives to be defined in terms of the primary ones.)

Definition 2.1. A modal system $\mathbf{S}$ is said to be *normal* if the following conditions are satisfied.

(1) For any tautology A, $\mathbf{S} \vdash A$.
(2) $\mathbf{S}$ contains the following *distribution* axiom $\mathbf{K}$
 $\mathbf{K}$: $\Box(A \to B) \to (\Box A \to \Box B)$
(3) Modus Ponens

$$\frac{A \quad A \to B}{B} \tag{MP}$$

(4) Necessitation

$$\frac{A}{\Box A} \tag{Nec}$$

(5) $\mathbf{S} \vdash \Diamond A \leftrightarrow \neg\Box\neg A$.

We sometimes use '$\mathbf{K}$' ambiguously as referring also to conditions (1)-(5) of normal systems themselves. Then various normal systems can be obtained by adding to $\mathbf{K}$ one or more axioms from the following list

$$\mathbf{D}: \Box A \to \Diamond A \quad \mathbf{T}: \Box A \to A \quad \mathbf{B}: A \to \Box\Diamond A \quad \mathbf{4}: \Box A \to \Box\Box A \quad \mathbf{5}: \Diamond A \to \Box\Diamond A$$

For instance, $\mathbf{K4}$ is a system which is the result of adding to $\mathbf{K}$ axiom $\mathbf{4}$. We list some simple properties of normal systems.

Lemma 2.2 (Normality). Assume $\mathbf{S}$ is a normal system, then

(1) $\mathbf{S} \vdash \Box(A \wedge B) \leftrightarrow \Box A \wedge \Box B$
(2) $\mathbf{S} \vdash \Diamond(A \vee B) \leftrightarrow \Diamond A \vee \Diamond B$
(3) If $\mathbf{S} \vdash A \to B$ then $\mathbf{S} \vdash \Box A \to \Box B$
(4) $\mathbf{S} \vdash \Diamond A \wedge \Box B \to \Diamond(A \wedge B)$

2.2. **GL and GLS.** The provability logics to be introduced consider the following extra axiom and inference rule:

L: $\Box(\Box A \to A) \to \Box A$
Löb Rule:

$$\frac{\Box A \to A}{A} \tag{LR}$$

Clearly, the Löb rule is motivated by the Löb Theorem 1.6 where the box operator is intended to be interpreted as the provability predicate. Now, let **K4+LR** be the system obtained by adopting in the normal system **K4** the Löb rule as an extra rule of inference. That is to say, for any A, A can be inferred in **K4+LR** if $\Box A \to A$ can be proved in it. Further, it is easily seen that axiom **L** is the axiom version of the Löb rule, let **GL** be the system obtained by adding to **K** the axiom **L** (i.e. **GL** $=$ **KL**). Apparently, **K4+LR** and **GL** are normal systems, and hence satisfy properties listed in Lemma 2.2. We show that these two systems are provably equivalent.

Lemma 2.3. K4+LR $\dashv\vdash$ **GL**.

Proof. $\Rightarrow$. We show that axiom **L** holds in **K4+RL**. Note that

1. **K4** $\vdash \Box\Big[\Box(\Box A \to A) \to \Box A\Big] \to \Big[\Box\Box(\Box A \to A) \to \Box\Box A\Big]$ by axiom **K**
2. **K4** $\vdash \Box(\Box A \to A) \to \Box\Box(\Box A \to A)$ axiom **4**
3. **K4** $\vdash \Box\Big[\Box(\Box A \to A) \to \Box A\Big] \to \Big[\Box(\Box A \to A) \to \Box\Box A\Big]$ (1) and (2)
4. **K4** $\vdash \Box(\Box A \to A) \to \big(\Box\Box A \to \Box A\big)$ by axiom **K**
5. **K4** $\vdash \Box\Big[\Box(\Box A \to A) \to \Box A\Big] \to \Big[\Box(\Box A \to A) \to \Box A\Big]$ (3) and (4)
6. **K4+LR** $\vdash \Box(\Box A \to A) \to \Box A$ by the Löb rule (LR)

$\Leftarrow$. Conversely, we show that, for any A, if **K4+LR** $\vdash A$ then **GL** $\vdash A$. To this end, it suffices to show (a) that

$$\mathbf{GL} \vdash \Box A \to \Box\Box A \tag{2.1}$$

(i.e. axiom **4** holds in **GL**), and (b) that **GL** $\vdash \Box A \to A$ implies **GL** $\vdash A$. To see (2.1), note that, by the tautology $A \to \big((B \wedge C) \to (C \wedge A)\big)$, we have

1. **GL** $\vdash A \to \big[(\Box\Box A \wedge \Box A) \to (\Box A \wedge A)\big]$
2. **GL** $\vdash A \to \big[\Box(\Box A \wedge A) \to (\Box A \wedge A)\big]$ by normality
3. **GL** $\vdash \Box A \to \Box\big[\Box(\Box A \wedge A) \to (\Box A \wedge A)\big]$ by the distribution axiom **K**
4. **GL** $\vdash \Box A \to \Box(\Box A \wedge A)$ in axiom **L** let $A = \Box A \wedge A$ then by (3)
5. **GL** $\vdash \Box(\Box A \wedge A) \to \Box\Box A$ by normality
6. **GL** $\vdash \Box A \to \Box\Box A$ by (4) and (5)

Further, if **GL** $\vdash \Box A \to A$, then, by necessitation, **GL** $\vdash \Box(\Box A \to A)$, this yields **GL** $\vdash \Box A$, via axiom **L**, and hence, by the hypothesis, **GL** $\vdash A$. $\Box$

GL is commonly referred to as the *propositional provability logic* (named after Gödel and Löb), where the box operator is intended to be interpreted as the provability predicate of **PA**. Here it shall be noted that in order for this intended interpretation to hold it is necessary that the three provability conditions HBL1-3 be satisfied. But it is easily seen that, under the provability interpretation, the necessitation rule, the distribution axiom, and (2.1) (i.e., axiom **4**) correspond respectively to HBL1-3. To state these correspondences between **GL** and **PA** more precisely, let us introduce the following notions of *realization* and *translation*.

TABLE 2.1.

K4	**PA**
$\Box(A \to B) \to (\Box A \to \Box B)$	$\mathsf{Prv}(\ulcorner f(A \to B) \urcorner) \to \big(\mathsf{Prv}(\ulcorner f(A) \urcorner) \to \mathsf{Prv}(\ulcorner f(B) \urcorner)\big)$
$\Box A \to \Box\Box A$	$\mathsf{Prv}(\ulcorner f(A) \urcorner) \to \mathsf{Prv}(\ulcorner \mathsf{Prv}(\ulcorner f(A) \urcorner) \urcorner)$
if $\vdash A$ then $\vdash \Box A$	if **PA** $\vdash f(A)$ then **PA** $\vdash \mathsf{Prv}(\ulcorner f(A) \urcorner)$
from A, $A \to B$ infer B	from $f(A)$, $f(A) \to f(B)$ infer $f(B)$
from $\Box A \to A$ infer A	from $\mathsf{Prv}(\ulcorner f(A) \urcorner) \to f(A)$ infer $f(A)$

Definition 2.4. A *realization* f is a function mapping from modal sentences in **GL** to arithmetic sentences in **PA**, a *translation* of a modal sentence with respect to a given realization f is defined recursively on the complexity of the sentence as follows

(1) for any atomic sentence A, $f(\mathsf{A}) = \gamma$, where γ is some sentence in **PA**.
(2) $f(\bot) = \bot$
(3) $f(B \to C) = f(B) \to f(C)$
(4) $f(\Box B) = \mathsf{Prv}(\ulcorner f(B) \urcorner)$

where 'Prv' is the wff formula defined in (1.2).

Always provable. According to this definition, a given modal sentence may have different translations under different realizations. But the translations of logical connectives and falsity are invariant under all realizations. We call a modal sentence A *always provable* if, for any realization f, the translation of A under f is provable in **PA**, that is, **PA** $\vdash f(A)$ for all f.[3] Lemma 2.3 then gives rise to the following observation.

Theorem 2.5. For any A and any realization f, if **GL** $\vdash A$ then **PA** $\vdash f(A)$

Proof. By Lemma 2.3, **GL** $\vdash A$ iff **K4** + **LR** $\vdash A$. The latter implies that A is either an instance of an axiom of **K4** or proved from a previous step via one of the inference rules, MP, Necessitation, or RL. Given any realization f, the Table 2.1 contains a list of axioms and rules in **K4** and their corresponding translation under f in **PA**. The items in the right column hold in **PA** because of HBL2, HBL3, HBL1, MP, and the Löb theorem (Theorem 1.6). $\Box$

Always true. The theorem establishes that, for any modal sentence A, if A is provable in **GL** then A is always provable (in **PA**), that is,

$$\mathbf{GL} \vdash A \quad \Rightarrow \quad \mathbf{PA} \vdash f(A) \text{ for any } f. \tag{2.2}$$

The converse of (2.2) is shown by Solovay (1976) also to be true, i.e.,

$$\mathbf{PA} \vdash f(A) \text{ for all } f \quad \Rightarrow \quad \mathbf{GL} \vdash A. \tag{2.3}$$

That is to say, every modal sentence that is aways provable is a theorem of **GL**. We shall delay this result of Solovay's to later (Theorem 3.4 below). For the time being, note that, by Theorem 2.5, all theorems of **GL** are always provable in **PA**, and hence *true* in the standard model of **PA**. We call a modal sentence *always true* if, for every realization f, $f(A)$ is true (in the standard model of **PA**). We have that all theorems of **GL** are always true.

There is another class of model sentences that are always true under translation. Observe that, for any A and any realization f, the transition of $\Box A \to A$ under

[3] This corresponds to the notion of **P**-*valid* in Solovay (1976).

f is $\mathsf{Prv}(\ulcorner f(A)\urcorner) \to f(A)$. The latter says that if $f(A)$ is provable in **PA** then it is true, which, as a statement of soundness of **PA**, is itself a true statement. That is, if $\mathsf{Prv}(\ulcorner f(A)\urcorner)$ is true then $f(A)$ is indeed a theorem of **PA** and hence true, therefore $\mathsf{Prv}(\ulcorner f(A)\urcorner) \to f(A)$ is always true. Moreover, if both $A \to B$ and A are alway true so is B, in other words, MP preserves what is alway true. These considerations give rise to the introduction of another provability logic **GLS** (named after Gödel, Löb, and Solovay) which consists of

$$\mathbf{GLS} := \begin{cases} \text{all theorems of } \mathbf{GL} \\ \text{all sentences of the form } \Box A \to A \\ \text{MP as the sole inference rule.} \end{cases} \tag{2.4}$$

Theorem 2.6. For any modal sentence A and any realization f, if $\mathbf{GLS} \vdash A$ then $f(A)$ is true (in the standard model of **PA**).

Note that **GLS** is *not* a normal system, this is simply because condition (4) of Definition 2.1 is not satisfied (Necessitation is not a rule of inference in **GLS**). Otherwise, if from the fact that $\Box\bot \to \bot$ is an instance of $\Box A \to A$ we conclude that $\Box(\Box\bot \to \bot)$ is a theorem of **GLS** via necessitation, which, by definition, is just $\Box(\neg\Box\bot)$, whose translation in **PA** is $\mathsf{Prv}(\mathsf{Con}_{\mathbf{PA}})$. By Theorem 2.6, the latter—which says the consistency of **PA** is provable in **PA**—is true, but this contradicts the second incompleteness theorem.

Theorem 2.6 establishes that, for any modal sentence A, if A is provable in **GLS** then it is always true (in the standard model of **PA**), that is,

$$\mathbf{GLS} \vdash A \quad \Rightarrow \quad \mathbf{PA} \vDash f(A) \text{ for any } f. \tag{2.5}$$

The converse of (2.5) is also shown by Solovay (1976) to be true, i.e.,

$$\mathbf{PA} \vDash f(A) \text{ for all } f \quad \Rightarrow \quad \mathbf{GLS} \vdash A. \tag{2.6}$$

That is, if A is always true that A is a theorem of **GLS** (Theorem 3.6 below). Before we turn to Solovay's arithmetical completeness results of (2.3) and (2.6) in Section 3. Let us first give a semantic analysis of the normal system **GL** in the standard Kripke structure for modal logic.

2.3. Soundness and Completeness in Kripke Structure.

A standard (Kripke) model $\mathcal{M}$ for propositional modal logic is a structure of the form $\mathcal{M} = \langle M, R, V\rangle$, where

(1) W is a non-empty set
(2) R is a binary relation on W
(3) V is a valuation function such that, for any atomic sentence A and for any $w \in W$, $V(w, \mathsf{A}) \in \{\mathbf{T}, \mathbf{F}\}$.

Members of W are often referred to as *possible worlds* and R is called the *accessibility relation* among possible worlds, that is, $(w_1, w_2) \in R$ iff w_2 is accessible from w_1. We refer to the first two component of $\mathcal{M}$ as the *frame* $\mathcal{F}$ of $\mathcal{M}$. In this case, we say that $\mathcal{M}$ is *based on* $\mathcal{F}$.

As usual, the *truth definition* of an arbitrary modal sentence A at a world w in a given model $\mathcal{M}$ is given in terms of the *value* of A in $\mathcal{M}$ at w denoted by $\mathrm{val}_w^{\mathcal{M}}(A)$ which is defined recursively on the complexity of A as follows:

(1) $\mathrm{val}_w^{\mathcal{M}}(\bot) = V(w, \bot) = \mathbf{F}$; for any atomic sentence A other than $\bot$, $\mathrm{val}_w^{\mathcal{M}}(\mathsf{A}) = V(w, \mathsf{A})$;

TABLE 2.2. Validities in $\mathcal{F}$

Axiom		**Property of R in frame $\mathcal{F} = (W, R)$**	
D	$\Box A \to \Diamond A$	Serial	$\forall u \exists v R(u, v)$
T	$\Box A \to A$	Reflexive	$\forall u R(u, u)$
B	$A \to \Box\Diamond A$	Symmetric	$\forall u, v[R(u, v) \to R(v, u)]$
4	$\Box A \to \Box\Box A$	Transitive	$\forall u, v, w[R(u, v) \wedge R(v, w) \to R(u, w)]$
5	$\Diamond A \to \Box\Diamond A$	Euclidean	$\forall u, v, w[R(u, v) \wedge R(u, w) \to R(v, w)]$

(2) $\text{val}_w^{\mathcal{M}}(\neg B) = \begin{cases} \mathbf{T}, & \text{if } \text{val}_w^{\mathcal{M}}(B) = \mathbf{F} \\ \mathbf{F}, & \text{if } \text{val}_w^{\mathcal{M}}(B) = \mathbf{T}; \end{cases}$

(3) $\text{val}_w^{\mathcal{M}}(B \to C) = \begin{cases} \mathbf{T}, & \text{if } \text{val}_w^{\mathcal{M}}(B) = \mathbf{F} \text{ or } \text{val}_w^{\mathcal{M}}(B) = \mathbf{T} \\ \mathbf{F}, & \text{if } \text{val}_w^{\mathcal{M}}(B) = \mathbf{T} \text{ and } \text{val}_w^{\mathcal{M}}(B) = \mathbf{F}; \end{cases}$

(4) $\text{val}_w^{\mathcal{M}}(\Box B) = \begin{cases} \mathbf{T}, & \text{if } \text{val}_v^{\mathcal{M}}(B) = \mathbf{T} \text{ for all } v \text{ such that } (w, v) \in R \\ \mathbf{F}, & \text{otherwise.} \end{cases}$

We say A is *satisfied* (or is *true*) in $(\mathcal{M}, w)$ if $\text{val}_w^{\mathcal{M}}(A) = \mathbf{T}$, written

$$(\mathcal{M}, w) \vDash A \quad =_{\text{Df}} \quad \text{val}_w^{\mathcal{M}}(A) = \mathbf{T}. \tag{2.7}$$

A is said to be *valid in model* $\mathcal{M}$, denoted by $\mathcal{M} \vDash A$, if $\text{val}_w^{\mathcal{M}}(A) = \mathbf{T}$ for all $w \in \mathcal{M}$, that is, $(\mathcal{M}, w) \vDash A$ for all $w \in \mathcal{M}$. A is said to be *valid in frame* $\mathcal{F} = \langle W, R \rangle$ written $\mathcal{F} \vDash A$, if, for every valuation function V', A is valid in model $\mathcal{M}' = \langle M, R, V' \rangle$, i.e., if $\mathcal{M}' \vDash A$ for all $\mathcal{M}'$ based on $\mathcal{F}$.

As a direct consequence of the truth definition above, we have the following simple soundness property of the basic normal system **K** in any Kripke model $\mathcal{M}$ with the truth definition above.

Lemma 2.7 (Soundness of **K**)**.** For any A, if $\mathbf{K} \vdash A$ then $\mathcal{M} \vDash A$.

For other normal systems, their soundnesses depend on the structure of the underling frame. As a characteristic feature of Kripke structure, it is known that an axiom of {**D, T, B, 4, 5**} is valid in frame $\mathcal{F} = \langle M, R \rangle$ *if and only if* the accessibility relation R of $\mathcal{F}$ satisfies the corresponding property listed in Table 2.2. Occasionally, we say a model/frame has certain property when we mean its accessibility has the property. As an illustration we show the following.

Lemma 2.8. Axiom **4** is valid in $\mathcal{F}$ if and only if R is transitive.

Proof. $\Rightarrow$. Suppose $\Box A \to \Box\Box A$ is valid in $\mathcal{F}$, and that wRu and uRv. We show wRv. It suffices to show that if $(w, v) \notin R$ then $\mathcal{F} \nvDash \Box A \to \Box\Box A$. Now let $\mathcal{M} = \langle M, R, V \rangle$ be such that $W = \{w, u, v\}$, $R = \{(w, u), (u, v)\}$, and $\text{val}_w^{\mathcal{M}}(A) = \text{val}_u^{\mathcal{M}}(A) = \mathbf{T}$ but $\text{val}_v^{\mathcal{M}}(A) = \mathbf{F}$, hence $(\mathcal{M}, w) \nvDash \Box A \to \Box\Box A$.

$\Leftarrow$. By the truth definition, we show that, for any $w \in \mathcal{M}$ and any valuation V, if $(\mathcal{M}, w) \vDash \Box A$ then $(\mathcal{M}, w) \vDash \Box\Box A$. The latter requires that for any u that is accessible from w, i.e., wRv, we have $(\mathcal{M}, u) \vDash \Box A$, which further requires that for any v with uRv we have $(\mathcal{M}, v) \vDash A$. This is met by the fact that R is transitive (and hence wRv) and the assumption that $(\mathcal{M}, w) \vDash \Box A$. $\Box$

Soundness. Let us now turn to **GL**, the goal is to identify certain property of R that corresponds to axiom **L**. To this end, we highlight the following property.

Definition 2.9 (Well-foundedness)**.** A binary relation R on W is said to be *well-founded* is for any non-empty $X \subseteq W$ there is an *R-least* member w of X such that, for any $x \in X$, $R(x, w)$ does not hold. R is said to be *conversely well-founded* if, for every non-empty subset X of W, there is an *R-greatest* element w of X such that wRx for no x in X.

Theorem 2.10 (Soundness of **GL**)**.** Let $\mathcal{F} = \langle W, R\rangle$ be a frame, then the following statement are equivalent:

(1) All the theorems are valid in $\mathcal{F}$.
(2) R is transitive and conversely well-founded.

Proof. By Lemma 2.7 and the fact that **GL** is a normal system, it suffices to show that Axiom **L** is valid in $\mathcal{F}$ if and only if R is transitive and conversely well-founded.

$\Rightarrow$. Suppose that $\Box(\Box A \to A) \to \Box A$ is valid in $\mathcal{F}$. By (2.1) and Lemma 2.8, R is transitive. We show that R is also conversely well-founded. Suppose, to the contrary, there exists some non-empty $X \subseteq W$ such that for any $x \in X$ there is always some $y \in X$ for which xRy holds, we show axiom **L** is not valid in some model based on $\mathcal{F}$.

Let the valuation function V of $\mathcal{M}$ be such that $\mathrm{val}_x^{\mathcal{M}}(A) = \mathbf{F}$ for all $x \in X$. Now fix some w in X, then, by the non-well-foundedness assumption, there must be some $y \in X$ such that wRy for which $(\mathcal{M}, y) \not\models A$. It follows $(\mathcal{M}, w) \not\models \Box A$. By the arbitrariness of w, we have that $(\mathcal{M}, x) \not\models \Box A$ for all $x \in X$. It follows that $(\mathcal{M}, w) \models \Box(\Box A \to A)$. Hence we have $(\mathcal{M}, w) \not\models \Box(\Box A \to A) \to \Box A$.

$\Leftarrow$. We show that, for any $\mathcal{M}$ based on $\mathcal{F} = \langle M, R\rangle$ and for any $w \in W$, if R is transitive and conversely well-founded then $(\mathcal{M}, w) \models \Box(\Box A \to A)$ implies that $(\mathcal{M}, w) \models \Box A$.

Suppose, to the contrary, that there exists some w such that $(\mathcal{M}, w) \models \Box(\Box A \to A)$ but $(\mathcal{M}, w) \not\models \Box A$. The latter implies there must be some y such that wRy for which $(\mathcal{M}, y) \not\models A$. Let Y be the set of all such y's, i.e.,

$$Y = \{y \in W \mid wRy \;\text{ and }\; (\mathcal{M}, y) \not\models A\}. \tag{2.8}$$

Then, by converse well-foundedness, there is a R-greatest element of Y, call it y^*. Note that, for any z satisfying y^*Rz, it must be that $(\mathcal{M}, z) \models A$. For, otherwise, by transitivity, $z \in Y$ then y^* is no longer the R-greatest element of Y, a contradiction. Now, from $(\mathcal{M}, w) \models \Box(\Box A \to A)$ and wRy^*, we conclude that $(\mathcal{M}, y^*) \models \Box A \to A$. Since $(\mathcal{M}, y^*) \not\models A$, the latter implies that $(\mathcal{M}, y^*) \not\models \Box A$, but this means there is some z such that y^*Rz for which $(\mathcal{M}, z) \not\models A$, again, a contradiction. □

The above soundness result establishes that, for any modal sentence A, if A is provable in **GL** then A is true in any model that is based on a frame $\mathcal{F} = \langle W, R\rangle$ whose accessibility relation is transitive and conversely well-founded:

$$\mathbf{GL} \vdash A \quad \Rightarrow \quad \mathcal{F} \models A. \tag{2.9}$$

Completeness. Next we show that A is a theorem of **GL** if A is valid in every (finite) frame $\mathcal{F}$ in which R is transitive and conversely well-founded, that is,

$$\mathcal{F} \models A \quad \Rightarrow \quad \mathbf{GL} \vdash A, \tag{2.10}$$

hence a (weak) *completeness* theorem for **GL**. As usual, this is sought by proving the contrapositive that, for some model $\mathcal{M} = \langle M, R, V\rangle$ based on $\mathcal{F}$, if A is not provable in **GL** then A is not valid in $\mathcal{M}$. To this end, let us fix a modal sentence D that is *not* a theorem of **GL**, i.e., $\mathbf{GL} \nvdash D$, the goal is to construct a model $\mathcal{M}$ based on $\mathcal{F} = \langle M, R\rangle$, where R is transitive and conversely well-founded, under which D is *not* valid, that is,

$$\mathbf{GL} \nvdash D \quad \Rightarrow \quad \mathcal{M} \nvDash D. \tag{2.11}$$

We first define, for each modal sentence A, the *length* of A which is a number $\ell(A)$ defined recursively as follows

(1) for any atomic sentence A, $\ell(\mathsf{A}) = 1$,
(2) $\ell(\bot) = 1$,
(3) $\ell(B \to C) = \ell(B) + \ell(C) + 1$,
(4) $\ell(\Box B) = \ell(B) + 1$.

It is easy to see that, for any A, A has at most $2^{\ell(A)}$ many subsentences. Further, we say that a set X of subsentences of D is *D-consistent in* **GL** if $\mathbf{GL} \nvdash \neg \bigwedge Y$ for all $Y \subseteq X$, where $\bigwedge Y$ is the conjunction of all members of Y. We say that X is *maximal D-consistent* if, for any subsentence B of D, either $B \in X$ or $\neg B \in X$. Since there are at most $2^{\ell(D)}$ many subsentences of D, there are at most $2^{\ell(D)}$ many D-consistent set X. Next, define $\mathcal{M} = \langle W, R, V\rangle$ to be such that

W**:** the domain of $\mathcal{M}$ contains all maximal D-consistent sets, that is,

$$W := \{w \mid w \text{ is maximal } D\text{-consistent}\} \tag{2.12}$$

R**:** for any $w, v \in W$,

$$wRv \quad \text{iff} \quad \begin{cases} \text{(i) for all } \Box A & \Box A \in w \Rightarrow \Box A, A \in v \\ \text{(ii) for some } \Box B & \Box B \in v \Rightarrow \neg\Box B \in w \end{cases} \tag{2.13}$$

V**:** for each atomic modal sentence A occurs in D and for any $w \in W$,

$$V(w, \mathsf{A}) = \begin{cases} \mathbf{T} & \text{if } \mathsf{A} \in w, \\ \mathbf{F} & \text{if } \mathsf{A} \notin w. \end{cases} \tag{2.14}$$

We show that D is not valid in this constructed model. This relies on the following observations.

Lemma 2.11. Let $\mathcal{M} = \langle W, R, V\rangle$ be defined as in (2.12)-(2.14), then

(1) For every subsentence $\Box A$ of D and every $w \in W$,

$$\Box A \in w \quad \Longleftrightarrow \quad \text{for any } v, wRv \text{ implies } A \in v. \tag{2.15}$$

(2) R is transitive and conversely well-founded.

Proof. (1) The '$\Rightarrow$' follows immediately from the first clause in the definition of R in (2.13). For the '$\Leftarrow$' we show that contrapositive, that is, if $\Box A \notin w$, then there is some v satisfying wRv and $A \notin v$. To this end, let

$$X = \{\neg A, \Box A\} \cup \{B, \Box B \mid \Box B \in w\}.$$

If X is inconsistent, then, let $B_1, \ldots, B_n, \Box B_1, \ldots, \Box B_n$ be an enumeration of $\{B, \Box B \mid \Box B \in w\}$ (this is due to the fact w is finite by definition), we have

1. $\mathbf{GL} \vdash \neg(\neg A \wedge \Box A \wedge B_1 \wedge \cdots \wedge B_n \wedge \Box B_1 \wedge \cdots \wedge \Box B_n)$
2. $\mathbf{GL} \vdash (B_1 \wedge \Box B_1 \wedge \cdots \wedge B_n \wedge \Box B_n \wedge \Box A) \to A$ — by 1 and pure logic
3. $\mathbf{GL} \vdash (B_1 \wedge \Box B_1 \wedge \cdots \wedge B_n \wedge \Box B_n) \to (\Box A \to A)$ — by 2 and pure logic

4. $\mathbf{GL} \vdash \big(\Box B_1 \wedge \Box\Box B_1 \wedge \cdots \wedge \Box B_n \wedge \Box\Box B_n\big) \to \Box(\Box A \to A)$ distribution
5. $\mathbf{GL} \vdash \big(\Box B_1 \wedge \cdots \wedge \Box B_n\big) \to \Box A$ by Axiom **4** and **L**

Since $\Box A$ is a subsentence of D, the last line implies that $\Box A \in w$ given that all of $\Box B_i$'s are in w. Thus if $\Box A \notin w$ then X is not inconsistent. Now if X is consistent, it is contained in some v. Note that since $\Box A \notin w$ it must be $\neg\Box A \in w$, then by the second clause of (2.13) we get wRv from $\Box A \in X \subseteq v$. Finally, since $\neg A \in X \subseteq v$, we have that $A \notin v$, which is what we want to show.

(2) Transitivity follows immediately from the first condition (i) of (2.13). For converse well-foundedness, we make use of the fact that the worlds in W are finite. It is easily seen that in this case that R is conversely well-founded iff it is irreflexive. Now if R is not conversely well-founded then R is reflexive, that is, wRw for all $w \in W$. Then by (ii) of (2.13), for some $\Box B$, we have both $\Box B$ and $\neg\Box B$ are in w which contradicts the consistency assumption of w. Therefore, R is indeed conversely well-founded. □

Lemma 2.12. For every subsentence A of D and any $w \in W$,

$$A \in w \quad \text{iff} \quad (\mathcal{M}, w) \vDash A. \tag{2.16}$$

Proof. The proof is given by induction on the complexity of A. We show the only non-trial case where $A = \Box B$. By Lemma 2.11 (1), $\Box B \in w$ iff, for any v, wRv implies $B \in v$. By the inductive hypothesis, the latter holds iff $v \vDash B$. That is, $\Box B \in w$ iff, for any v, wRv implies $v \vDash B$. This means, by the truth definition for '$\Box$', $\Box B \in w$ iff $w \vDash \Box B$. Therefore, (2.16) holds. □

Note that, by the assumption that $\mathbf{GL} \nvdash D$, we have that $\{\neg D\}$ is consistent in **GL**, and hence is contained in some maximal D-consistent set, say w^*. Clearly, $D \notin w^*$, and hence by Lemma 2.12, $(\mathcal{M}, w^*) \nvDash D$, which is what we seek to show. This leads to the following completeness theorem.

Theorem 2.13 (Completeness of **GL**)**.** For any modal sentence A, if A is valid in every (finite) frame $\mathcal{F} = \langle W, R\rangle$ in which R is transitive and conversely well-founded then $\mathbf{GL} \vdash A$.

3. Arithmetical Completeness Theorems of Solovay

3.1. **GL is proof-complete with respect to PA.** We now return to the first arithmetical completeness result of Solovay (1976). As shown in Theorem 2.5, for any modal sentence A, if $\mathbf{GL} \vdash A$ then $\mathbf{PA} \vdash f(A)$ for all realization f. We seek to show that the converse is also true, that is, if A is always provable then A is provable in **GL** (cf. (2.3)). It is clear that this is achieved if it can be shown that, for any given D, if D is not a theorem of **GL** then there is some realization f^* under which $f^*(D)$ is not provable in **PA**, i.e.,

$$\mathbf{GL} \nvdash D \;\Rightarrow\; \mathbf{PA} \nvdash f^*(D) \tag{3.1}$$

The task is hence to construct such a realization function f^* under which the translation of D is not provable in **PA**. Note that, by Theorem 2.13, if $\mathbf{GL} \nvdash D$ then there is some finite, transitive, and conversely well-founded model $\mathcal{M} = \langle W, R, V\rangle$ such that, for some $w_0 \in W$, $(\mathcal{M}, w_0) \nvDash D$. The main step of Solovay's proof can be viewed as constructing f^* using this finite model $\mathcal{M}$.

To simplify matters, let $\mathcal{M}$ be such that $W = \{1, \ldots, n\}$ with $w_0 = 1$ and R is a transitive and conversely well-founded relation on W satisfying also that $1Ri$ for all $1 < i \leq n$. We have $(\mathcal{M}, 1) \nvDash D$. Further, for any i, denote by S_i the set of j's for which (i, j) stand in relation R:

$$S_i =_{\text{Df}} \{j \in W \mid iRj\} \text{ for } 1 \leq i \leq n, \text{ and we specify}$$
$$S_0 =_{\text{Df}} \{1, \ldots, n\}.$$

Next we seek to define a function $h : \omega \to \{0, 1, \ldots, n\} (= n + 1 = \{0\} \cup W)$ which features the following property

$$h(0) = 0 \text{ and if } h(m) = i \text{ then } h(m+1) = \begin{cases} j & \text{for some } j \in S_i \text{ (i.e. } iRj\text{)}, \\ i & \text{otherwise.} \end{cases} \tag{3.2}$$

Admittedly, whether or not $h(m + 1) = i$ or j depends on further specification. But, for time being, it is clear from (3.2) that, if well defined, h is non-decreasing, and, by converse well-foundedness of R and the fact that W is finite, h has a limit in $n + 1$. Let l denote the limit value of h, that is, $l = \lim_{m\to\infty} h(m)$. We give a formal inductive definition of h using l as follows

$$\begin{aligned} h(0) &= 0 \\ h(m) &= i \\ h(m+1) &= \begin{cases} j & \text{if, for some } j \in S_i, \textbf{PA} \vdash \mathsf{Proof}(\underline{m}, \ulcorner \underline{l} \not\approx \underline{j} \urcorner), \\ i & \text{otherwise.} \end{cases} \end{aligned} \tag{3.3}$$

That is to say, h is so defined that $h(m + 1)$ remains i unless, for some $j \in S_i$, m is the Gödel number of a proof in **PA** that j is *not* the limit of h, i.e., $j \neq l = \lim_{m\to\infty} h(m)$. Obviously, the inductive step of the recursive definition of h above refers explicitly to the limit of the function h being defined. The circularity is handled by applying a generalized fixed-point theorem, from which we get that function h can be represented in **PA** by a wff $H(x, y)$ such that if $h(a) = b$ then $\textbf{PA} \vdash \forall y[H(\underline{a}, y) \to y \approx \underline{b}]$.[4] The expression "the limit of h is i" ($0 \leq i \leq n$) can then be expressed in **PA** by the following *Solovay sentences*:

$$\chi_i := \exists z \forall x \Big[x \geq z \to \exists y \big(y \approx \underline{i} \wedge H(x, y) \big) \Big] \quad (0 \leq i \leq n). \tag{3.4}$$

The sentence says that if the limit of h is i, i.e., if $l = i$, then there is some m for which $h(m) = i$ and that, for any $m' > m$, $h(m') = i$. The following is a list of properties can be sought for Solovay sentences.

Lemma 3.1. (1) $\textbf{PA} \vdash \chi_0 \vee \chi_1 \vee \cdots \vee \chi_n$.
(2) χ_0 is true in standard model of **PA**.
(3) For all $0 \leq i \leq n$, $\textbf{PA} \nvdash \neg\chi_i$.
(4) For any $0 \leq i \leq n$ and for any $j \in S_i$, $\textbf{PA} \vdash \chi_i \to \neg\mathsf{Prv}(\ulcorner \neg\chi_j \urcorner)$.
(5) For any $1 \leq i \leq n$, if $j \notin S_i$ then $\textbf{PA} \vdash \chi_i \to \mathsf{Prv}(\ulcorner \neg\chi_j \urcorner)$.

Proof. See §4.1-4.5 in Solovay (1976, p. 296-297). □

[4]See Boolos (1995, 126ff) for detailed representation of h in **PA**.

Embedding. Now we proceed to construct a realization function f^* using model $\mathcal{M}$ and the Solovay sentences introduced above. To this end, we first extend $\mathcal{M} = \langle W, R, V\rangle$ to $\mathcal{M}' = \langle W', R', V'\rangle$ which includes world 0:

$$\begin{aligned} W' &= W \cup \{0\}, \\ R' &= R \cup \{(0,i) \mid 1 \leq i \leq n\}, \\ V'(i, \mathsf{A}) &= \begin{cases} V(1, \mathsf{A}) & \text{if } i = 0, \\ V(i, \mathsf{A}) & \text{if } 0 < i \leq n. \end{cases} \end{aligned} \tag{3.5}$$

It is easy to see that $\mathcal{M}'$ is transitive and conversely well-founded and $(\mathcal{M}', 1) \nvDash D$. We seek to *embed* $\mathcal{M}'$ into **PA** through the following process of translation.

Define a realization f^* function from sentences of **GL** to that of **PA** to be such that, for any atomic sentence A,

$$f^*(\mathsf{A}) = \bigvee_{V'(i,\mathsf{A})=\mathbf{T}} \chi_i \tag{3.6}$$

(let $f^*(\mathsf{A}) = \bot$ if no i verifies A). Translations of compound sentences with respect to f^* are defined similarly as (2)-(4) in Definition 2.4.

Lemma 3.2. Let A be any modal sentence. For any $1 \leq i \leq n$,

(1) if $(\mathcal{M}', i) \vDash A$ then $\mathbf{PA} \vdash \chi_i \to f^*(A)$;
(2) if $(\mathcal{M}', i) \nvDash A$ then $\mathbf{PA} \vdash \chi_i \to \neg f^*(A)$

Proof. The proof is by induction on complexity of A. The basic case where A is an atomic sentence follows directly from (3.6). For compound sentences we discuss only the case where A is in the form of $\Box B$.

(1) If $(\mathcal{M}', i) \vDash \Box B$ the for all $j \in S_i$ we have $(\mathcal{M}', j) \vDash B$ (recall that $S_i = \{j \mid iRj\}$ and, for all $1 \leq i, j \leq n$, iRj iff $iR'j$). By the inductive hypothesis, the latter yields,
1. $\mathbf{PA} \vdash \chi_j \to f^*(B)$, for all $j \in S_i$
2. $\mathbf{PA} \vdash \bigvee_{j \in S_i} \chi_j \to f^*(B)$ by (1)
3. $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner \bigvee_{j\in S_i} \chi_j \urcorner) \to \mathsf{Prv}(\ulcorner f^*(B) \urcorner)$ distribution
4. $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner \bigvee_{j\in S_i} \chi_j \urcorner) \to f^*(A)$ $f^*(A) = \mathsf{Prv}(\ulcorner f^*(B) \urcorner)$
5. $\mathbf{PA} \vdash \chi_i \to \mathsf{Prv}(\ulcorner \bigvee_{j\in S_i} \chi_j \urcorner)$ by Lemma 3.1(1)&(5)
6. $\mathbf{PA} \vdash \chi_i \to f^*(A)$ by (4) and (5)

(2) If $(\mathcal{M}', i) \nvDash \Box B$ then there is some $j \in S_i$ for which $(\mathcal{M}', j) \nvDash B$. By the inductive hypothesis, the latter yields,
1. $\mathbf{PA} \vdash \chi_j \to \neg f^*(B)$
2. $\mathbf{PA} \vdash f^*(B) \to \neg \chi_j$ by (1)
3. $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner f^*(B) \urcorner) \to \mathsf{Prv}(\ulcorner \neg\chi_j \urcorner)$ distribution
4. $\mathbf{PA} \vdash \neg\mathsf{Prv}(\ulcorner \neg\chi_j \urcorner) \to \neg f^*(A)$ by 3 and $f^*(A) = \mathsf{Prv}(\ulcorner f^*(B) \urcorner)$
5. $\mathbf{PA} \vdash \chi_i \to \neg f^*(A)$ by (4) and Lemma 3.1(4)

□

The following is a variant of the lemma above, which will become handy in the next section.

Lemma 3.3. Let A be any modal sentence, suppose that for any subsentence of A of the form $\Box B$, $(\mathcal{M}', 1) \vDash \Box B \to B$, then for any subsentence C of A:

(1) if $(\mathcal{M}',1) \vDash C$ then $\mathbf{PA} \vdash \chi_0 \to f^*(C)$;
(2) if $(\mathcal{M}',1) \nvDash C$ then $\mathbf{PA} \vdash \chi_0 \to \neg f^*(C)$

Proof. As usual, we show the only non-trivial case where C is in the form of $\Box D$.

(1) If $(\mathcal{M}',1) \vDash \Box D$ then $(\mathcal{M}',i) \vDash D$ for all $i \in S_1$. By the hypothesis of the lemma we have $(\mathcal{M}',1) \vDash \Box D \to D$, hence $(\mathcal{M}',1) \vDash D$. Apply Lemma 3.2 we get, $\mathbf{PA} \vdash \chi_i \to f^*(D)$ for all $1 \leq i \leq n$. Apply the inductive hypothesis, from $(\mathcal{M}',1) \vDash D$ we get $\mathbf{PA} \vdash \chi_0 \to f^*(D)$. Together, we have

$$\mathbf{PA} \vdash \big(\chi_0 \vee \chi_1 \vee \cdots \vee \chi_n\big) \to f^*(D)$$

Then, by Lemma 3.1(1), $\mathbf{PA} \vdash f^*(D)$, and hence $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner f^*(D)\urcorner)$. The latter yields $\mathbf{PA} \vdash \chi_0 \to f^*(\Box D)$.

(2) If $(\mathcal{M}',1) \nvDash \Box D$ then, for some $j \in S_1$, $(\mathcal{M}',j) \nvDash D$. Apply Lemma 3.2(2) we have,

1. $\mathbf{PA} \vdash \chi_j \to \neg f^*(D)$
2. $\mathbf{PA} \vdash f^*(D) \to \neg\chi_j$ — by (1)
3. $\mathbf{PA} \vdash \mathsf{Prv}\big(\ulcorner f^*(D)\urcorner\big) \to \mathsf{Prv}\big(\ulcorner \neg\chi_j\urcorner\big)$ — distribution
4. $\mathbf{PA} \vdash \neg\mathsf{Prv}\big(\ulcorner \neg\chi_j\urcorner\big) \to \neg f^*(\Box D)$ — by (3) and $f^*(\Box D) = \mathsf{Prv}(\ulcorner f^*(D)\urcorner)$
5. $\mathbf{PA} \vdash \chi_0 \to \neg f^*(\Box D)$ — by 4 and Lemma 3.1(4)

$\Box$

Theorem 3.4 (Arithmetic Completeness of **GL**)**.** For any modal sentence A, if, for any realization f, $\mathbf{PA} \vdash f(A)$ then $\mathbf{GL} \vdash A$.

Proof. Suppose that D is not a theorem of **GL**. Let $\mathcal{M}'$ and f^* be defined as above. We have $(\mathcal{M}',1) \nvDash D$, then, by Lemma 3.2, $\mathbf{PA} \vdash \chi_1 \to \neg f^*(D)$. Note that (3) of Lemma 3.1, χ_1 is consistent with **PA**, hence $\neg f^*(D)$ is also consistent with **PA**, from which we conclude that $\mathbf{PA} \nvdash f^*(D)$. This is what we want to show.

Alternatively, it can also be shown that

1. $\mathbf{PA} \vdash \chi_1 \to \neg f^*(D)$
2. $\mathbf{PA} \vdash f^*(D) \to \neg\chi_1$ — by (1)
3. $\mathbf{PA} \vdash \mathsf{Prv}(\ulcorner f^*(D)\urcorner) \to \mathsf{Prv}\big(\ulcorner \neg\chi_1\urcorner\big)$ — distribution
4. $\mathbf{PA} \vdash \neg\mathsf{Prv}\big(\ulcorner \neg\chi_1\urcorner\big) \to \neg\mathsf{Prv}(\ulcorner f^*(D)\urcorner)$ — by (3)
5. $\mathbf{PA} \vdash \chi_0 \to \neg\mathsf{Prv}(\ulcorner \neg\chi_1\urcorner)$ — by Lemma 3.1(3)
6. $\mathbf{PA} \vdash \chi_0 \to \neg\mathsf{Prv}(\ulcorner f^*(D)\urcorner)$

By (2) of Lemma 3.1 and the soundness of **PA**, the last line above implies that $\neg\mathsf{Prv}(\ulcorner f^*(D)\urcorner)$ is also true in the standard model of **PA**, and hence $f^*(D)$ is not provable in **PA**. $\Box$

3.2. **GLS is truth-complete with respect to PA.** As remarked in §2.2, all theorems of **GLS** are always true, that is, for any theorem A of **GLS**, the translation of A under any realization f is true in the standard model of **PA**. We show that the converse, i.e., the second arithmetical completeness result of Solovay, is also true. To this end, we first note that, given the construction of **GLS** in (2.4), theorems of **GL** are closely related to that of **GLS**. We specify this relationship by highlighting the following correspondence: given any modal sentence A, let $\Box B_1, \ldots, \Box B_n$ be all subsentences of A with principle connective $\Box$, define A^S of A to be such that

$$\Big[(\Box B_1 \to B_1) \wedge \cdots \wedge (\Box B_m \to B_m)\Big] \to A. \qquad (A^S)$$

Lemma 3.5. For any A, $\mathbf{GLS} \vdash A$ if and only if $\mathbf{GL} \vdash A^S$.

Proof. $\Leftarrow$. If $\mathbf{GL} \vdash A^S$ then $\mathbf{GLS} \vdash A^S$, but all of $\Box B_i \to B_i$ $(1 \leq i \leq m)$ are axioms of **GLS**, hence $\mathbf{GLS} \vdash A$.

$\Rightarrow$. We show $\mathbf{GLS} \vdash A$ implies $\mathbf{GL} \vdash A^S$. Suppose, to the contrary, $\mathbf{GL} \nvdash A^S$, we derive a contradiction by constructing a realization f^* under which $f^*(A)$ is false (because $\mathbf{GLS} \vdash A$ implies that A is always true).

Apply the methods towards the proof of Theorem 3.4, from $\mathbf{GL} \nvdash A^S$ we can construct a model $\mathcal{M}'$ and a realization f^* defined in (3.6) such that $(\mathcal{M}', 1) \nvDash A^s$. Truth functionally, the latter implies $(\mathcal{M}', 1) \vDash \Box B_i \to B_i$ for all $\leq i \leq m$ but $(\mathcal{M}', 1) \nvDash A$. By Lemma 3.3, we have

$$\mathbf{PA} \vdash \chi_0 \to \neg f^*(A) \tag{3.7}$$

Again, by Lemma 3.1(2) and the soundness of **PA**, (3.7) implies $\neg f^*(A)$ is true in the standard model of **PA**. Hence $f^*(A)$ is false, which is what we want to show. □

Theorem 3.6 (Arithmetical Completeness of **GLS**)**.** For any modal sentence A, if, for any realization f, $f(A)$ is true then $\mathbf{GLS} \vdash A$.

Proof. Suppose, to the contrary, that $\mathbf{GLS} \nvdash A$. Then, by Lemma 3.5, $\mathbf{GL} \nvdash A^S$ which further implies, via (3.7), that there is some realization f^* under which $f^*(A)$ is false, a contradiction. □

References

Boolos, G. (1995). *The Logic of Provability*. Cambridge University Press.

Boolos, G. and G. Sambin (1991). Provability: the emergence of a mathematical modality. *Studia Logica 50*(1), 1–23.

Gaifman, H. (2000). What Gödel's incompleteness result does and does not show. *The Journal of Philosophy 97*(8), 462–470.

Henkin, L. (1952). A problem concerning provability, problem 3. *Journal of Symbolic Logic 17*(2), 160.

Löb, M. H. (1955). Solution of a problem of Leon Henkin. *The Journal of Symbolic Logic 20*, 115–118.

Solovay, R. (1976). Provability interpretations of modal logic. *Israel Journal of Mathematics 25*(3-4), 287–304.

Verbrugge, R. L. (2014). Provability logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2014 ed.).