Chapter 7

On the Theory of Self-Reference

By self-reference we basically mean the possibility of talking inside a theory T about T itself or related theories. Here we can give merely a glimpse into this recently much advanced area of research; see e.g. [Bu]. We will prove Gödel's second incompleteness theorem, Löb's theorem, and many other results related to self-reference, while further results are discussed only briefly and elucidated by means of applications. All this is of great interest both for epistemology and the foundations of mathematics.

The mountain we first have to climb is the proof of the derivability conditions for PA and related theories in **7.1**, and the derivable Σ_1-completeness in **7.2**. But anyone contented with leafing through these sections can begin straight away in **7.3**; from then on we will just be reaping the fruits of our labor. However, one would forgo a real adventure in doing so, namely the fusion of logic and number theory in the analysis of PA. For a comprehensive understanding of self-reference, the material of **7.1** and **7.2** (partly prepared in Chapter **6**) should be studied anyway.

Gödel himself tried to interpret the notion "provable" using a modal operator in the framework of the modal system S4. This attempt reflects some of his own results, though not adequately. Only after 1970, when modal logic was sufficiently advanced, could such a program be successfully carried out. A suitable instrument turned out to be the modal logic denoted by G (or GL). The Kripke semantics for G introduced in **7.4** is an excellent tool for confirming or refuting self-referential statements. Solovay's completeness theorem and the completeness theorem of Kripke semantics for G in **7.5** are fortunately of the kind that allows application without knowing the completeness proof itself, which in both cases are not quite easy and use several technical tricks.

W. Rautenberg, *A Concise Introduction to Mathematical Logic*, Universitext, DOI 10.1007/978-1-4419-1221-3_7,

There are several extensions of G, for example, the bimodal logic GD in **7.6**. This logic is related to Hilbert's famous ω-rule. A weakening of it can expressed by the modal operator $\boxed{1}$ of GD. A comprehensive survey can be found in [Bu, Chapter VII]; see also [Vi2]. In **7.7** we discuss some questions regarding self-reference in axiomatic set theory.

7.1 The Derivability Conditions

Put somewhat simply, Gödel's second incompleteness theorem states that $\vdash_T \mathtt{Con}_T$ cannot hold for a sufficiently strong and consistent axiomatizable theory T. Here $\mathtt{Con}_T$ is a sentence reflecting the metatheoretic statement of consistency of T *inside* T, more precisely, inside the (first-order) language $\mathcal{L}$ of T. In a popular formulation: *If T is consistent, then this consistency is unprovable in T.* As was outlined by Gödel and will be verified in this chapter, the italicized sentence is not only true but also formalizable in $\mathcal{L}$ and even provable in the framework of T.

The easiest way to obtain Gödel's theorem is first to prove the *derivability conditions* stated below. Their formulation supposes the arithmetizability of T, which includes the distinguishing of a sequence $\underline{0}, \underline{1}, \ldots$ of ground terms; see page 250. Let $\mathtt{bew}_T(y, x)$ be a formula that represents the recursive predicate bew_T in T as in **6.4**. For $\mathtt{bwb}_T(x) = \exists y\, \mathtt{bew}_T(y, x)$ we write $\Box(x)$, and $\Box\alpha$ is to mean $\mathtt{bwb}_T \frac{\ulcorner \alpha \urcorner}{x}$. We may read $\Box\alpha$ as "box α" or more suggestively "α is provable in T," because $\Box\alpha$ reflects the metatheoretic property $\vdash_T \alpha$ *in* T. If $\Box$ refers to some theory $T' \neq T$ then $\Box$ has to be indexed correspondingly. For instance, $\Box_{\mathsf{ZFC}}\alpha$ for $\alpha \in \mathcal{L}_\in$ can easily expressed also in $\mathcal{L}_{ar}$. Note that $\Box\alpha$ is always a sentence, even if α contains free variables.

Further, set $\Diamond\alpha := \neg\Box\neg\alpha$ for $\alpha \in \mathcal{L}$. If α is a sentence, $\Diamond\alpha$ may be read as α *is compatible with* T, because it formalizes '$\nvdash_T \neg\alpha$', which is, as we know, equivalent to the consistency of $T + \alpha$. First of all, we define $\mathtt{Con}_T$ in a natural way by

$$\mathtt{Con}_T := \neg\Box\bot \ \big(= \neg\, \mathtt{bwb}_T(\ulcorner \bot \urcorner)\big),$$

where $\bot$ is a contradiction, $0 \neq 0$, for instance. We shall see in a moment that $\mathtt{Con}_T$ is independent modulo T of the choice of $\bot$. The mentioned derivability conditions then read as follows:

$D1$: $\vdash_T \alpha \;\Rightarrow\; \vdash_T \Box\alpha$,
$D2$: $\Box\alpha \wedge \Box(\alpha \to \beta) \vdash_T \Box\beta$,
$D3$: $\vdash_T \Box\alpha \to \Box\Box\alpha$.

Here α, β run through all sentences of $\mathcal{L}$. These conditions are due to Löb, but they were considered in a slightly different setting already in [HB]. Sometimes $D2$ is written in the equivalent form $\Box(\alpha \to \beta) \vdash_T \Box\alpha \to \Box\beta$, and $D3$ as $\Box\alpha \vdash_T \Box\Box\alpha$.

A consequence of $D1$ and $D2$ is $D0$: $\alpha \vdash_T \beta \;\Rightarrow\; \Box\alpha \vdash_T \Box\beta$. This results from the following chain of implications:

$$\alpha \vdash_T \beta \;\Rightarrow\; \vdash_T \alpha \to \beta \;\Rightarrow\; \vdash_T \Box(\alpha \to \beta) \;\Rightarrow\; \vdash_T \Box\alpha \to \Box\beta \;\Rightarrow\; \Box\alpha \vdash_T \Box\beta.$$

From $D0$ it clearly follows that $\alpha \equiv_T \beta \;\Rightarrow\; \Box\alpha \equiv_T \Box\beta$. In particular, the choice of $\bot$ in $\mathtt{Con}_T$ is arbitrary as long as $\bot \equiv_T 0 \neq 0$.

Remark 1. Any operator $\partial\colon \mathcal{L} \to \mathcal{L}$ satisfying the conditions $d1$: $\vdash_T \alpha \Rightarrow \vdash_T \partial\alpha$ and $d2$: $\partial(\alpha \to \beta) \vdash_T \partial\alpha \to \partial\beta$ thus satisfies also $d0$: $\alpha \vdash_T \beta \Rightarrow \partial\alpha \vdash_T \partial\beta$, and hence $d00$: $\alpha \equiv_T \beta \Rightarrow \partial\alpha \equiv_T \partial\beta$, for all $\alpha, \beta \in \mathcal{L}$. It likewise satisfies $d\wedge$: $\partial(\alpha \wedge \beta) \equiv_T \partial\alpha \wedge \partial\beta$, for $\alpha \wedge \beta \vdash_T \alpha, \beta$, hence $\partial(\alpha \wedge \beta) \vdash_T \partial\alpha, \partial\beta \vdash_T \partial\alpha \wedge \partial\beta$ in view of $d0$. The converse direction $\partial\alpha \wedge \partial\beta \vdash_T \partial(\alpha \wedge \beta)$ readily follows from $\alpha \vdash_T \beta \to \alpha \wedge \beta$ by first applying $d0$ and then $d2$.

Whereas $D2$ and $D3$ represent sentence schemata in T, condition $D1$ is of metatheoretic nature and follows obviously from the representability of bew_T in T. Thus, $D1$ holds even for weak theories such as $T = \mathsf{Q}$. On the other hand, the converse of $D1$,

$D1^*$: $\vdash_T \Box\alpha \;\Rightarrow\; \vdash_T \alpha$, for all $\alpha \in \mathcal{L}^0$,

may fail. Fortunately, it holds for all ω-consistent axiomatic extensions $T \supseteq \mathsf{Q}$ such as $T = \mathsf{PA}$. Indeed, $\nvdash_T \alpha$ implies $\vdash_T \neg\, \mathtt{bew}_T(\underline{n}, \ulcorner\alpha\urcorner)$ for all n (Corollary 6.4.3). Hence, $\nvdash_T \exists y\, \mathtt{bew}_T(y, \ulcorner\alpha\urcorner)$ in view of the ω-consistency of T, that is, $\nvdash_T \Box\alpha$.

Unlike $D1$, the properties $D2$ and $D3$ are not so easily obtained. The theory T must be able not only to *speak* about provability in T (perhaps via arithmetization), but also to *prove* basic properties about provability. $D3$ is nothing else than *condition* $D1$ *formalized within* T, while $D2$ formalizes (7) from page 230, the closure under MP in arithmetical terms. Let us first realize that $D2$ holds, provided it has been shown that

$D2^*$: $\mathtt{bew}_T(u, x) \wedge \mathtt{bew}_T(v, x \dot{\to} y) \vdash_T \mathtt{bew}_T(u * v * \langle y \rangle, y)$,

where the p.r. functions $\tilde{\to}$, $*$, and $y \mapsto \langle y \rangle$ appearing in $D2^*$ must either be present or definable in T. Generally speaking, $f \in \mathbf{F}_n$ is called *definable* in an arithmetizable theory $T \subseteq \mathcal{L}$ (with respect to the sequence of terms $(\underline{n})_{n\in\mathbb{N}}$ in T) if there is a formula $\delta(\vec{x}, y) \in \mathcal{L}$ such that

(1) (a) $\vdash_T \delta(\vec{\underline{a}}, \underline{f\vec{a}})$ for all $\vec{a}$, (b) $\vdash_T \forall \vec{x}\, \exists! y \delta(\vec{x}, y)$.

Clearly, f is then also represented by $\delta(\vec{x}, y)$. For $T = \mathsf{PA}$ and related theories, (1) means that f is explicitly definable in T in the sense of **2.6** and may be introduced in T (using a corresponding symbol). From now on we will no longer distinguish between T and its definitorial extensions and apply $\vdash_T y = f\vec{x} \leftrightarrow \delta(\vec{x}, y)$ without comment. This and (1) easily imply $\vdash_T \underline{f\vec{a}} = f\vec{\underline{a}}$, e.g. $\vdash_T \underline{a \tilde{\to} b} = \underline{a} \tilde{\to} \underline{b}$. With $\ulcorner\alpha\urcorner$, $\ulcorner\beta\urcorner$ for x, y, we thus obtain from $D2^*$ in view of $\ulcorner \alpha \to \beta \urcorner = \underline{\dot{\alpha} \tilde{\to} \dot{\beta}} = \underline{\dot{\alpha}} \tilde{\to} \underline{\dot{\beta}} = \ulcorner\alpha\urcorner \tilde{\to} \ulcorner\beta\urcorner$,

$$\mathtt{bew}_T(u, \ulcorner\alpha\urcorner) \wedge \mathtt{bew}_T(v, \ulcorner \alpha \to \beta \urcorner) \vdash_T \mathtt{bew}_T(u * v * \langle \ulcorner\beta\urcorner \rangle, \ulcorner\beta\urcorner).$$

Particularization yields $D2$. But the real work, the definability of the functions appearing in $D2^*$ in theories like $T = \mathsf{PA}$, still lies ahead.

In order to better keep track of things, we restrict our considerations to the theories ZFC and PA, which are of central interest in nearly all foundational questions. ZFC is only briefly discussed. Here the proofs of $D2$ and $D3$ (with $\Box = \Box_{\mathsf{ZFC}}$) are much easier than in PA and need only a few lines as follows: $D2^*$ and hence $D2$ are clear, because the naive proof of $D2^*$ above with $bew_T = bew_{\mathsf{ZFC}}$ can easily be formalized *inside* ZFC. This includes the definability of all functions occurring in $D2^*$, for we *did* define them; for instance, the operation $*$ on page 224 may be defined by setting $a * b = \emptyset$ if $a \notin \omega$ or $b \notin \omega$. We arithmetize $\mathcal{L}_\in$ according to the pattern in **6.2**, encoding formulas with Gödel numbers,[1] so that $\mathcal{L}_\in$-formulas are encoded within ZFC by certain ω-terms, defined in **3.4**. Formulas from $\mathcal{L}_{ar}$ are identified with their ω-relativized in $\mathcal{L}_\in$, called the *arithmetical formulas* of $\mathcal{L}_\in$. Moreover, the arithmetical predicate bew_{ZFC} is certainly representable in ZFC by Theorem 6.4.2, since this theorem can be viewed, just like every theorem in this book, as a theorem *within* ZFC. Thus, the naive proof of $D1$ based on this theorem (up to Corollary 6.4.3) can as a whole be carried out in ZFC, and so $D3$ is proved.

[1] This is not actually necessary, since in ZFC one can talk directly about finite sequences and hence about $\mathcal{L}_\in$-formulas (Remark 2 in **6.6**), but we do so in order to maintain coherence with the exposition in **6.2**.

Roughly speaking, $D2$ and $D3$ hold for ZFC because ordinary mathematics, in particular the material in Chapter **6**, is formalizable in ZFC. In all of the above, no special set-theoretic constructions such as transfinite recursion are needed. Only relatively simple combinatorial facts are required. Hence there is some hope that the proofs of $D2$ and $D3$ can also be carried out in sufficiently strong arithmetical theories like PA. This is indeed so. The proof of $D3$ for PA will need the most effort and will be completed only in **7.2**. Our first goal will be to show that the p.r. functions occurring in $D2^*$, and in fact *all* p.r. functions, are explicitly definable in PA.[2] They turn out to be definable even in a sense stronger than required by (1) from the previous page.

Definition. An n-ary recursive function f is called *provably recursive* or Σ_1*-definable in* PA if there is a Σ_1-formula $\delta_f(\vec{x}, y)$ in $\mathcal{L}_{ar}$ such that

(2) (a) $\vdash_{\mathsf{PA}} \delta_f(\vec{a}, \underline{f\vec{a}})$ for all $\vec{a} \in \mathbb{N}^n$; (b) $\vdash_{\mathsf{PA}} \forall \vec{x}\, \exists! y \delta_f(\vec{x}, y)$.

Since PA is Σ_1-complete, 2(a) is equivalent to $\mathcal{N} \vDash \delta_f(\vec{a}, \underline{f\vec{a}})$ for all $\vec{a}$, which is often more easily verified than 2(a) and could replace 2(a). We will show that *all* p.r. functions are Σ_1-definable in PA, which strengthens their explicit definability in PA. Thereafter we may treat all occurring p.r. functions in PA as if they had been available in the language right from the outset. Essentially this fundamental fact allows a treatment of elementary number theory and combinatorics within the boundaries of PA and hence is particularly interesting for a critical foundation of mathematics.

If $\delta_f(\vec{x}, y)$ in (2) is Δ_0 then f is called Δ_0-definable. An example is the $\boldsymbol{\beta}$-function (Exercise 1), which from now on may be supposed to be present in PA. Basic for the Σ_1-definability of all p.r. functions is $\boldsymbol{\beta}$'s main property, Lemma 6.4.1, of which we need, of course, some provable version in PA. Since Euclid's lemma and the Chinese remainder theorem are involved here, these should be derived first. Clearly, the basic arithmetical laws applied in their proofs in **6.3** should be at our disposal, including those on the order relation and on $a - b$ for $a \geqslant b$, all provable in N.

The proof of Euclid's lemma is straightforward, Exercise 2. As for the Chinese remainder theorem, we avoid the quantification over finite

[2] In [Gö2], Gödel presented a list of 45 definable p.r. functions; the last was χ_{bew}. Following [WR], Gödel considered a higher-order arithmetical theory. That Gödel's theorems also hold in first-order arithmetic was probably first noticed in [HB].

sequences for the time being, by stating the theorem as a scheme. Let c, d denote unary provably recursive functions, which may depend on further parameters. Each such c determines for given n the sequence $\mathsf{c}_0, \ldots, \mathsf{c}_n$, with $\mathsf{c}_\nu = \mathsf{c}(\nu)$ for $\nu \leqslant n$. For suggestive reasons from now on also letters such as $n, \nu, \ldots$ may denote variables in $\mathcal{L}_{ar}$. With the Δ_0-definable relation $\perp$ of coprimeness, the Chinese remainder theorem can provisionally be stated as follows: for arbitrary c, d as arranged above, we get

$$(3)\quad \vdash_{\mathsf{PA}} \forall n[(\forall i, j{\leqslant}n)(\mathsf{c}_i{<}\mathsf{d}_i \wedge (i{\neq}j \to \mathsf{d}_i \perp \mathsf{d}_j)) \to \exists a(\forall \nu{\leqslant}n)\, \mathrm{rem}(a, \mathsf{d}_\nu) = \mathsf{c}_\nu].$$

To convert the original proof of the remainder theorem to one for (3) we require, for given provably recursive d, the term $\mathrm{lcm}\{\mathsf{d}_\nu | \nu{\leqslant}n\}$, the least common multiple of $\mathsf{d}_0, \ldots, \mathsf{d}_n$. **Claim**: $f : n \mapsto \mathrm{lcm}\{\mathsf{d}_\nu | \nu{\leqslant}n\}$ is defined in PA by the Σ_1-formula

$$\delta_f(x, y) := (\forall \nu{\leqslant}x)\mathsf{d}_\nu \,|\, y \wedge (\forall z{<}y)(\exists \nu{\leqslant}x)\, \mathsf{d}_\nu \not| \, z.$$

More precisely, $\delta_f(x, y)$ *describes* a Σ_1-formula in $\mathcal{L}_{ar}$ that is even Δ_0, provided d is Δ_0-definable. Clearly $\mathcal{N} \vDash \delta(\underline{n}, \underline{\mathrm{lcm}\{\mathsf{d}_\nu | \nu{\leqslant}n\}})$ for all n. Thus, 2(a) holds. With the minimum schema (Exercise 4 in **3.3**) applied to $\beta(x, y) := (\forall \nu{\leqslant}x)\mathsf{d}_\nu \,|\, y$, we obtain $\vdash_{\mathsf{PA}} \exists! y \delta_f(x, y)$, provided it has been shown that $\vdash_{\mathsf{PA}} \exists y \beta(x, y)$ ('$\mathsf{c}_0, \ldots, \mathsf{c}_x$ have a common multiple'), which is easily derived by induction on x; see Example 1 in **2.5**. This proves the claim. After having derived Euclid's lemma in PA (Exercise 2) we confirm (3) by following the proof of the remainder theorem in **6.2**, and, writing $\boldsymbol{\beta} s t$ for $\boldsymbol{\beta}(s, t)$, a suitable version of Lemma 6.4.1 as follows:

$$(4)\quad \vdash_{\mathsf{PA}} \forall n \exists u(\forall \nu{\leqslant}n)\, \mathsf{c}_\nu = \boldsymbol{\beta} u \nu, \text{ for any given provably recursive } \mathsf{c}.$$

Theorem 1.1. *Each p.r. function f is provably recursive. Moreover, the recursion equations for f are provable in* PA *whenever $f = \boldsymbol{Op}(g, h)$.*

Proof. For the initial functions and $+, \cdot$ the formulas $\boldsymbol{v}_0 = 0$, $\boldsymbol{v}_1 = \mathsf{S}\boldsymbol{v}_0$, $\boldsymbol{v}_n = \boldsymbol{v}_\nu$ along with $\boldsymbol{v}_2 = \boldsymbol{v}_0 + \boldsymbol{v}_1$ and $\boldsymbol{v}_2 = \boldsymbol{v}_0 \cdot \boldsymbol{v}_1$ are obviously defining Σ_1-formulas. For the composition $f = h[g_1, \ldots, g_m]$, let $\delta_f(\vec{x}, y)$ be the formula $y = h(g_1\vec{x}, \ldots, g_m\vec{x})$. In this case (2) is clear, because we might think of $h, g_1, \ldots, g_m$ as being already introduced in PA, so that $\delta_f(\vec{x}, y)$ belongs to the expanded language. Only the construction of δ_f for the case $f = \boldsymbol{Op}(g, h)$ requires some skill. We may assume that besides $\boldsymbol{\beta}$ also g, h have already been introduced in the language. Consider

$$(5)\ \delta_f(\vec{x},y,z) := \exists u[\underbrace{\boldsymbol{\beta}u0 = g\vec{x} \wedge (\forall v{<}y)\boldsymbol{\beta}u\mathsf{S}v = h(\vec{x},v,\boldsymbol{\beta}uv)\ \wedge\ \boldsymbol{\beta}uy = z}_{\gamma(u,\vec{x},y,z)}].$$

δ_f is similar to $\delta_{\exp}$ from Remark 1 in **6.4**. It is Σ_1, because $\boldsymbol{\beta}$, g, h are Σ_1-definable. Lemma 6.4.1 applied with $c_i = f(\vec{a},i)$ for $i \leqslant b$ shows that $\mathcal{N} \vDash \delta_f(\underline{\vec{a}},\underline{b},\underline{f\vec{a}})$, equivalently 2(a). Uniqueness in 2(b), that is,

$$\delta_f(\vec{x},y,z) \wedge \delta_f(\vec{x},y,z') \vdash_{\mathsf{PA}} z = z',$$

derives easily from $\gamma(u,\vec{x},y,z) \wedge \gamma(u',\vec{x},y,z') \vdash_{\mathsf{PA}} z = z'$, which clearly follows from $\gamma(u,\vec{x},y,z) \wedge \gamma(u',\vec{x},y,z') \vdash_{\mathsf{PA}} (\forall v{\leqslant}y)\boldsymbol{\beta}uv = \boldsymbol{\beta}u'v$. This is easily shown by induction on y. Also, $\vdash_{\mathsf{PA}} \exists z\delta_f(\vec{x},y,z)$ will be shown inductively on y. We get $\vdash_{\mathsf{PA}} \exists u\boldsymbol{\beta}u0 = g\vec{x}$ (hence $\vdash_{\mathsf{PA}} \exists z\delta_f(\vec{x},0,z)$) from (4), choosing c therein such that $\mathsf{c}_0 = g\vec{x}$ and $\mathsf{c}_\nu = 0$ for $\nu \neq 0$. c is provably recursive, for the term $g\vec{x}$ is Σ_1-definable. The inductive step will be verified informally, that is, we shall prove

$$(*)\qquad \exists z\delta_f(\vec{x},y,z) \vdash_{\mathsf{PA}} \exists z'\delta_f(\vec{x},\mathsf{S}y,z').$$

Suppose $\gamma(u,\vec{x},y,z)$. Consider the provably recursive $\mathsf{c}\colon \nu \mapsto \mathsf{c}_\nu$ defined by $\mathsf{c}_\nu = \boldsymbol{\beta}u\nu$ for $\nu \leqslant \mathsf{S}y$ and $\mathsf{c}_{\mathsf{S}y} = h(\vec{x},y,\boldsymbol{\beta}uy)$. Here $u,\vec{x},y$ are parameters in the defining Σ_1-formula for c. So by (4) (taking $\mathsf{S}y$ for n) there is some u' with $\boldsymbol{\beta}u'\nu = \mathsf{c}_\nu = \boldsymbol{\beta}u\nu$ for all $\nu \leqslant y$ and $\boldsymbol{\beta}u'\mathsf{S}y = \mathsf{c}_{\mathsf{S}y} = h(\vec{x},y,\boldsymbol{\beta}uy)$. With this u' and $z' = \boldsymbol{\beta}u'\mathsf{S}y$ we obtain $\gamma(u',\vec{x},\mathsf{S}y,z')$, and so $\exists z'\delta_f(\vec{x},\mathsf{S}y,z')$. This confirms $(*)$ and hence 2(b). Thus, f is provably recursive and may now be introduced in PA. We finally sketch a proof of the recursion equations for f in PA, which also in PA may be written as usual, i.e.,

$$(\mathrm{A})\ \vdash_{\mathsf{PA}} f(\vec{x},0) = g\vec{x}, \quad (\mathrm{B})\ \vdash_{\mathsf{PA}} f(\vec{x},\mathsf{S}y) = h(\vec{x},y,f(\vec{x},y)).$$

(A) holds because $\vdash_{\mathsf{PA}} \delta_f(\vec{x},0,f(\vec{x},0)) \equiv_{\mathsf{PA}} \exists u(\boldsymbol{\beta}u0 = g\vec{x} \wedge \boldsymbol{\beta}u0 = f(\vec{x},0))$ and clearly $\exists u(\boldsymbol{\beta}u0 = g\vec{x} \wedge \boldsymbol{\beta}u0 = f(\vec{x},0)) \vdash f(\vec{x},0) = g\vec{x}$. (B) follows by $<$-induction on y applied to $\alpha = \alpha(\vec{x},y) := f(\vec{x},\mathsf{S}y) = h(\vec{x},y,f(\vec{x},y))$. Assume that $(\forall v{<}y)\alpha\frac{v}{y}$. Choosing u in (5) such that $\gamma(u,\vec{x},\mathsf{S}y,f(\vec{x},\mathsf{S}y))$, we readily obtain $(\forall v{\leqslant}y)f(\vec{x},v) = \boldsymbol{\beta}uv$, so that

$$f(\vec{x},\mathsf{S}y) = \boldsymbol{\beta}u\mathsf{S}y = h(\vec{x},y,\boldsymbol{\beta}uy) = h(\vec{x},y,f(\vec{x},y)).$$

This confirms $\forall y((\forall v{<}y)\alpha\frac{v}{y} \to \alpha)$, hence $\vdash_{\mathsf{PA}} \forall y\alpha$ by $<$-induction. ❏

We thus have achieved our first goal. Next observe that the properties of $*, \ell, \ldots$ from the remark on page 230 along with the basic property (5) stated there are also readily proved *within* PA. This is a little extra program that includes the proof of unique prime factorization, see Exercise 4.

Thus, $D2^*$ and hence $D2$ are indeed provable for $T = \mathsf{PA}$. In particular, the property (6) from page 230 carries over to PA, so that

(6) $\quad \Box(x \tilde{\rightarrow} y) \vdash_{\mathsf{PA}} \Box(x) \rightarrow \Box(y)$.

We mention that $\Box$ in (3) may even denote the formula $\mathtt{bwb}_T$ for any axiomatizable (and arithmetizable) theory T. $D3$ will be proved in the next section in a somewhat broader context.

Remark 2. The formalized equations of Exercise 3 in **6.4** are now also provable in PA. For instance, item (b) reads $\vdash_{\mathsf{PA}} \mathrm{sb}_{\vec{x}}(\ulcorner\varphi\urcorner, \vec{x}) = \mathrm{sb}_{\vec{x}'}(\ulcorner\varphi\urcorner, \vec{x}\,')$ for $\varphi = \varphi(\vec{x})$, where $\vec{x}\,'(\subseteq \vec{x})$ enumerates the free variables of φ. As regards (c), consider first a special case. Let φ be $\mathsf{S}x = y$. Then $\mathrm{sb}_{xy}(\dot{\varphi}, x, \mathsf{S}x) = \mathrm{sb}_x((\varphi\,\frac{\mathsf{S}x}{y})^{\cdot}, x)$, formalized $\mathrm{sb}_{xy}(\ulcorner\varphi\urcorner, x, y)\,\frac{\mathsf{S}x}{y} = \mathrm{sb}_x(\ulcorner\varphi\,\frac{\mathsf{S}x}{y}\urcorner, x)$. For the proof of this equation in PA, just $\vdash_{\mathsf{PA}} \mathrm{cf}\,\mathsf{S}x = \tilde{\mathsf{S}}\,\mathrm{cf}\,x$ is required, which holds by Theorem 1.1. Whoever wants to write down a detailed proof should follow the example on page 249.

Exercises

1. Prove in PA the Δ_0-definability of the remainder function rem, the pairing function, and the β-function; see **6.4**. In particular, rem is defined by $\delta_{\mathrm{rem}}(a, b, r) := (\exists q{\leqslant}a)(a = b \cdot q + r \wedge r < b) \vee b = r = 0$. The laws of arithmetic as given by N (page 235) may be used.

2. Prove *in* PA (a) $(\forall a, b{>}0)\exists x\exists y(a \perp b \rightarrow ax + \underline{1} = by)$, that is, Euclid's lemma. (b) $(\forall a{>}1)\exists p(\mathsf{prim}\,p \wedge p\,|\,a)$ ('each number $\geqslant 2$ has a prime divisor'), (c) $\vdash_{\mathsf{PA}} (\forall a, b{>}0)\forall p(\mathsf{prim}\,p \wedge p\,|\,ab \rightarrow p\,|\,a \vee p\,|\,b)$.

3. Show that $\vdash_{\mathsf{PA}} \mathsf{prim}\,p \wedge p\,|\,\mathrm{lcm}\{\mathsf{d}_\nu\,|\,\nu{\leqslant}n\} \rightarrow (\exists i{\leqslant}n)p\,|\,\mathsf{d}_i$, required for carrying out the proof of the Chinese remainder theorem in PA.

4. One of several possibilities of formalizing the prime factorization in PA is $(\forall n{\geqslant}\underline{2})(\exists m{\geqslant}\underline{2})n = \prod_{i\leqslant\ell m} p_i^{(m)_i}$, where m serves as a variable for the sequence of prime exponents.[3] Prove this in PA, as well as its uniqueness, which is essentially based on Exercise 2.

5. Let $T' = T + \alpha$ and T satisfy $D1$–$D3$. Show that
   (a) $\vdash_T \Box_{T'}\varphi \leftrightarrow \Box_T(\alpha \rightarrow \varphi)$ (the formalized deduction theorem),
   (b) $D1$–$D3$ hold also for T'.

[3] An equivalent formalization of the prime factorization in PA using the β-function is $(\forall k{\geqslant}\underline{2})\exists u\exists n(k = \prod_{i\leqslant n} p_i^{\beta ui} \wedge \beta un{\neq}0)$.

7.2 The Provable Σ_1-Completeness

$D3$ is a special case of the *provable* Σ_1-completeness. This is essentially the statement $\vdash_{\mathsf{PA}} \alpha \to \Box\alpha$ for Σ_1-sentences α. The proof demands still additional preparation, and even good textbooks do not carry out all proof steps. All steps described in this section and not handled in detail can easily be completed in full by the sufficiently assiduous reader. Life could be made easier through the mutual interpretability of PA and $\mathsf{ZFC}_{\mathrm{fin}}$ mentioned in **6.6**. Let $\Box = \Box(x)$ denote the formula $\mathtt{bwb}_{\mathsf{PA}}(x)$ till the end of this section. We first introduce an additional notation. Let $\varphi = \varphi(\vec{x})$.

Definition. $\Box[\varphi] := \Box(\mathrm{sb}_{\vec{x}}(\ulcorner\varphi\urcorner, \vec{x}))$ $(= \mathtt{bwb}_{\mathsf{PA}}\,\frac{\mathrm{sb}_{\vec{x}}(\ulcorner\varphi\urcorner,\vec{x})}{x})$.

By Remark 2 in **7.1**, $\vdash_{\mathsf{PA}} \mathrm{sb}_{\vec{x}}(\ulcorner\varphi\urcorner, \vec{x}) = \mathrm{sb}_{\vec{x}\,\prime}(\ulcorner\varphi\urcorner, \vec{x}\,')$, where $\vec{x}\,'$ enumerates $\mathit{free}\,\varphi$. Hence, we may assume w.l.o.g. that $\mathit{free}\,\Box[\varphi] = \mathit{free}\,\varphi$. Moreover, for $\alpha \in \mathcal{L}^0_{ar}$ we have $\vdash_{\mathsf{PA}} \mathrm{sb}_{\vec{x}}(\ulcorner\alpha\urcorner, \vec{x}) = \mathrm{sb}_{\emptyset}(\ulcorner\alpha\urcorner) = \ulcorner\alpha\urcorner$, hence $\Box[\alpha]$ and $\Box\alpha$ may be identified. '$\vdash_{\mathsf{PA}} \varphi(\underline{\vec{a}})$ for all $\vec{a} \in \mathbb{N}^n$' is reflected in PA by '$\vdash_{\mathsf{PA}} \forall\vec{x}\Box[\varphi]$'. The latter thus reflects in PA the existence of a *collection of proofs* which, due to the ω-incompleteness of PA, may be less than $\vdash_{\mathsf{PA}} \Box\forall\vec{x}\varphi$, or what amounts to the same, $\vdash_{\mathsf{PA}} \Box\varphi$.

Example. Let $\varphi = \varphi(x, y)$ be $\mathtt{S}x = y$. We prove $\varphi \vdash_{\mathsf{PA}} \Box[\varphi]$, or equivalently, $\vdash_{\mathsf{PA}} \Box[\varphi]\,\frac{\mathtt{S}x}{y}$, where w.l.o.g. x, y do not occur bound in $\Box(x)$. In order to prove $\vdash_{\mathsf{PA}} \Box[\varphi]\,\frac{\mathtt{S}x}{y}$ observe that in view of Remark 2 in **7.1**,

$$\Box[\varphi]\,\tfrac{\mathtt{S}x}{y} = \Box(\mathrm{sb}_{xy}(\ulcorner\varphi\urcorner, x, \mathtt{S}y)) \equiv_{\mathsf{PA}} \Box(\mathrm{sb}_x(\ulcorner\varphi\,\tfrac{\mathtt{S}x}{y}\urcorner, x)) = \Box[\alpha(x)]$$

with $\alpha(x) := \mathtt{S}x = \mathtt{S}x$. Thus, it suffices to verify $\vdash_{\mathsf{PA}} \Box[\alpha(x)]$ (equivalently, $\vdash_{\mathsf{PA}} \forall x\Box[\alpha(x)]$). This reflects in PA 'for arbitrary n, $\vdash_{\mathsf{PA}} \mathtt{S}\underline{n} = \mathtt{S}\underline{n}$'. We verify $\vdash_{\mathsf{PA}} \Box[\alpha(x)]$ in detail. Consider the p.r. function $\tilde{\alpha} : n \mapsto \mathrm{sb}_x(\dot{\alpha}, n)$ (the Gödel number of $\alpha(\underline{n})$). By axiom $\Lambda 9$, $\langle\tilde{\alpha}(n)\rangle$ is for each n a trivial arithmetized proof of length 1. Stated within PA, $\vdash_{\mathsf{PA}} \mathtt{bew}_{\mathsf{PA}}(\langle\tilde{\alpha}(x)\rangle, \tilde{\alpha}(x))$. This clearly yields $\vdash_{\mathsf{PA}} \exists y\,\mathtt{bew}_{\mathsf{PA}}(y, \tilde{\alpha}(x)) = \Box(\tilde{\alpha}(x)) = \Box[\alpha]$.

Next we prove some modifications $D1$, $D2$ for $\alpha = \alpha(\vec{x})$ and $\beta = \beta(\vec{x})$:

(1) (a) $\vdash_{\mathsf{PA}} \alpha \;\Rightarrow\; \vdash_{\mathsf{PA}} \Box[\alpha]$; (b) $\Box[\alpha \to \beta] \vdash_{\mathsf{PA}} \Box[\alpha] \to \Box[\beta]$.

To see (a) let $\vdash_{\mathsf{PA}} \alpha$, hence also $\vdash_{\mathsf{PA}} \forall\vec{x}\alpha$ and so $\vdash_{\mathsf{PA}} \Box\forall\vec{x}\alpha$. Just as in the above example, a proof for $\forall\vec{x}\alpha$ provides one for $\alpha_{\vec{x}}(\underline{\vec{a}})$ in a p.r. way, or stated *within* PA: $\Box\forall\vec{x} \vdash_{\mathsf{PA}} \Box(\mathrm{sb}_{\vec{x}}(\ulcorner\alpha\urcorner, \vec{x}))$ $(= \Box[\alpha]$; thus, $\vdash_{\mathsf{PA}} \Box[\alpha])$.

(b) follows from (6) in **7.1** with $\mathrm{sb}_{\vec{x}}(\ulcorner\alpha\urcorner, \vec{x}), \mathrm{sb}_{\vec{x}}(\ulcorner\beta\urcorner, \vec{x})$ for x, y, observing that $\vdash_{\mathsf{PA}} \mathrm{sb}_{\vec{x}}(\ulcorner\alpha \to \beta\urcorner, \vec{x}) = \mathrm{sb}_{\vec{x}}(\ulcorner\alpha\urcorner, \vec{x}) \tilde{\to} \mathrm{sb}_{\vec{x}}(\ulcorner\beta\urcorner, \vec{x})$, see Exercise 3 in **6.4**. (c) of this exercise yields for all not necessarily distinct x, y

(2) $\quad \Box[\alpha]\frac{t}{x} \equiv_{\mathsf{PA}} \Box[\alpha\frac{t}{x}] \quad (t \in \{0, y, \mathsf{S}y\}$ and $y \notin bnd\,\alpha)$.

Now, $D3$ is only a special case of the *provable* Σ_1-*completeness* of PA, stated not only for sentences, but for arbitrary formulas as follows:

(3) $\quad \varphi \vdash_{\mathsf{PA}} \Box[\varphi]$ (equivalently, $\vdash_{\mathsf{PA}} \varphi \to \Box[\varphi]$), for all Σ_1-formulas φ.

Indeed, choose in (3) for φ the Σ_1-sentence $\Box\alpha$ for any $\alpha \in \mathcal{L}^0_{ar}$. Then $\Box\alpha \vdash_{\mathsf{PA}} \Box[\Box\alpha] \equiv \Box\Box\alpha$, and $D3$ is proved. We obtain (3) from Theorem 2.1 below, since by (1), (2), and since w.l.o.g. $free\,\alpha = free\,\Box[\alpha]$, the operator $\partial\colon \alpha \mapsto \Box[\alpha]$ satisfies the conditions of the theorem.

Theorem 2.1. *Let* $\partial\colon \mathcal{L}_{ar} \to \mathcal{L}_{ar}$ *be any operator with* $free\,\partial\alpha \subseteq free\,\alpha$ *satisfying*

$d1$: $\quad \vdash_{\mathsf{PA}} \alpha \Rightarrow \vdash_{\mathsf{PA}} \partial\alpha$,

$d2$: $\quad \partial(\alpha \to \beta) \vdash_{\mathsf{PA}} \partial\alpha \to \partial\beta$,

ds: $\quad \partial\alpha\frac{t}{x} \equiv_{\mathsf{PA}} \partial(\alpha\frac{t}{x}) \quad (t \in \{0, y, \mathsf{S}y\},\ y \notin bnd\,\alpha)$.

Then $\vdash_{\mathsf{PA}} \varphi \to \partial\varphi$ *holds for all* Σ_1-*formulas* $\varphi \in \mathcal{L}_{ar}$.

Proof. ∂ satisfies also $d0$, $d00$, and $d\wedge$ (see Remark 1 in **7.1**). Hence, by Theorem 6.7.2 and $d00$ we need to carry out the proof only for special Σ_1-formulas. First let φ be $\mathsf{S}x = y$. Clearly, $\vdash_{\mathsf{PA}} \varphi \to \partial\varphi$ is equivalent to $\vdash_{\mathsf{PA}} \partial\varphi\frac{\mathsf{S}x}{y}$, and this to $\vdash_{\mathsf{PA}} \partial\,\mathsf{S}x = \mathsf{S}x$ by ds, which is obvious from $d1$. Now let φ be $x + y = z$. We shall prove $\vdash_{\mathsf{PA}} \forall yz(\varphi \to \partial\varphi)$ by induction on x. Observing that $y = z \vdash_{\mathsf{PA}} \partial\, y = z$ (equivalently $\vdash_{\mathsf{PA}} \partial z = z$), we obtain $\varphi\frac{0}{x} \vdash_{\mathsf{PA}} y = z \vdash_{\mathsf{PA}} \partial\, y = z \equiv_{\mathsf{PA}} \partial(\varphi\frac{0}{x}) \equiv_{\mathsf{PA}} \partial\varphi\frac{0}{x}$. Thus, $\vdash_{\mathsf{PA}} \forall yz(\varphi \to \partial\varphi)\frac{0}{x}$. Now $\varphi\frac{\mathsf{S}y}{y} \equiv_{\mathsf{PA}} \varphi\frac{\mathsf{S}x}{x}$; hence $\partial\varphi\frac{\mathsf{S}y}{y} \equiv_{\mathsf{PA}} \partial\varphi\frac{\mathsf{S}x}{x}$, by $d00$, ds. The induction step $\forall yz(\varphi \to \partial\varphi) \vdash_{\mathsf{PA}} \forall yz(\varphi \to \partial\varphi)\frac{\mathsf{S}x}{x}$ follows then from

$$\forall yz(\varphi \to \partial\varphi) \vdash \varphi\tfrac{\mathsf{S}y}{y} \to \partial\varphi\tfrac{\mathsf{S}y}{y} \vdash_{\mathsf{PA}} \varphi\tfrac{\mathsf{S}x}{x} \to \partial\varphi\tfrac{\mathsf{S}x}{x} = (\varphi \to \partial\varphi)\tfrac{\mathsf{S}x}{x}.$$

The formula $x \cdot y = z$ is left to the reader, who should observe $d\wedge$, $d2$, the induction steps for $\wedge$ and $\exists$, and $\mathsf{S}x \cdot y = z \equiv_{\mathsf{PA}} \exists u(x \cdot y = u \wedge u + y = z)$.

We now treat the logical connectives. The induction steps for $\wedge, \vee, \exists$ are simple. Indeed, from $d\wedge$ we obtain

$$\alpha \wedge \beta \vdash \alpha, \beta \vdash_{\mathsf{PA}} \partial\alpha \wedge \partial\beta \vdash_{\mathsf{PA}} \partial(\alpha \wedge \beta).$$

For $\vee$ note that $\alpha \vdash_{\mathsf{PA}} \partial\alpha \vdash_{\mathsf{PA}} \partial(\alpha \vee \beta)$, and similarly for β. Further, since $\varphi \vdash \exists x\varphi$ we get $\varphi \vdash_{\mathsf{PA}} \partial\varphi \vdash_{\mathsf{PA}} \partial\exists x\varphi$ by $d0$, and from $x \notin free\,\partial\exists x\varphi$

follows $\exists x\varphi \vdash_{\mathsf{PA}} \partial\exists x\varphi$. The prime-term substitution step (t is prime in $\frac{t}{x}$) also runs smoothly: $\varphi \vdash_{\mathsf{PA}} \partial\varphi$ yields $\varphi\,\frac{t}{x} \vdash_{\mathsf{PA}} \partial\varphi\,\frac{t}{x} \vdash_{\mathsf{PA}} \partial(\varphi\frac{t}{x})$ by ds.

It remains to verify the step for bounded quantification. Suppose that $\alpha \vdash_{\mathsf{PA}} \partial\alpha$ and $y \notin var\,\alpha$. We prove $\varphi := (\forall x{<}y)\,\alpha \vdash_{\mathsf{PA}} \partial\varphi$ by induction on y. The initial step is obvious: $\vdash_{\mathsf{PA}} \varphi\,\frac{0}{y}$, and therefore

$$\vdash_{\mathsf{PA}} \partial(\varphi\tfrac{0}{y}) \vdash_{\mathsf{PA}} \partial\varphi\,\tfrac{0}{y} \vdash_{\mathsf{PA}} \varphi\,\tfrac{0}{y} \to \partial\varphi\,\tfrac{0}{y}\,.$$

Clearly, $\varphi\,\frac{\mathsf{S}y}{y} \equiv_{\mathsf{PA}} \varphi \wedge \alpha\,\frac{y}{x}$. Hence $\alpha\,\frac{y}{x} \vdash_{\mathsf{PA}} \partial\alpha\,\frac{y}{x} \vdash_{\mathsf{PA}} \partial(\alpha\frac{y}{x})$ because of $\alpha \vdash_{\mathsf{PA}} \partial\alpha$. That leads to

$$\varphi\,\tfrac{\mathsf{S}y}{y} \wedge (\varphi \to \partial\varphi) \vdash_{\mathsf{PA}} \varphi \wedge \alpha\,\tfrac{y}{x} \wedge (\varphi \to \partial\varphi) \vdash_{\mathsf{PA}} \partial\varphi \wedge \partial(\alpha\tfrac{y}{x})$$
$$\vdash_{\mathsf{PA}} \partial(\varphi \wedge \alpha\tfrac{y}{x}) \vdash_{\mathsf{PA}} \partial(\varphi\tfrac{\mathsf{S}y}{y}).$$

Thus, $\varphi \to \partial\varphi \vdash_{\mathsf{PA}} \varphi\frac{\mathsf{S}y}{y} \to \partial(\varphi\frac{\mathsf{S}y}{y})$, which is obviously equivalent to the inductive step. ❑

Remark 3. $D1$–$D3$ are also provable for much weaker theories than PA, e.g., for the so-called *elementary arithmetic* $\mathsf{EA} = I\Delta_0 + \forall xy\exists z\delta_{exp}(x,y,z)$. Here $I\Delta_0$ is defined in Remark 1 in **6.3** and δ_{exp} is a defining Δ_0-formula for exp, see also [FS]. Also Theorem 1.1 can essentially be strengthened and has many variants. For instance, the provably recursive functions of $I\Sigma_1$ (like PA but IS restricted to Σ_1-formulas) are precisely the p.r. ones, [Tak]. The same provably recursive functions has EA augmented by the Π_2-induction schema without parameters, [Be4]. It is noteworthy that the provable recursive functions of EA itself are precisely the elementary ones, [Si]. For more material on the metatheory of PA and related theories see [Bar, Part D], and in particular [HP].

7.3 The Theorems of Gödel and Löb

We are now in a position to harvest the yields of our efforts. As long as not stated otherwise, let T denote any arithmetizable axiomatic theory in $\mathcal{L}$, that satisfies the derivability conditions $D1$–$D3$ of **7.1** along with the fixed point lemma of **6.5**. We direct attention straight away to the uniqueness statement of Lemma 3.1(b) below. According to this claim, up to equivalence in T at most $\Box\alpha \to \alpha$ can be the fixed point of the formula $\Box(x) \to \alpha$. The proof of Theorem 3.2 will show that $\neg\Box(x)$ too has only one fixed point modulo T. Beneath all this lies, as we shall see from Corollary 5.6, a completely general result.

Lemma 3.1. *Let T be as arranged above, and let $\alpha, \gamma \in \mathcal{L}^0$ be such that $\gamma \equiv_T \Box\gamma \to \alpha$. Then* (a) $\Box\gamma \equiv_T \Box\alpha$ *and* (b) $\gamma \equiv_T \Box\alpha \to \alpha$.

Proof. The supposition yields $\Box\gamma \vdash_T \Box(\Box\gamma \to \alpha) \vdash_T \Box\Box\gamma \to \Box\alpha$, by $D0$ and $D2$. Now by $D3$, we clearly obtain $\Box\gamma \vdash_T \Box\Box\gamma$, hence $\Box\gamma \vdash_T \Box\alpha$. Since $\alpha \vdash_T \Box\gamma \to \alpha \equiv_T \gamma$ and so $\alpha \vdash_T \gamma$, it follows that $\Box\alpha \vdash_T \Box\gamma$ by $D0$. Together with the already verified $\Box\gamma \vdash_T \Box\alpha$ we get (a). Using (a) we may replace $\Box\gamma$ with $\Box\alpha$ in $\gamma \equiv_T \Box\gamma \to \alpha$, which results in (b). ❏

Theorem 3.2 (Second incompleteness theorem). PA *satisfies alongside the fixed point lemma also $D1$–$D3$. Every theory T with these properties satisfies the conditions*

(1) $\nvdash_T \mathtt{Con}_T$ *provided T is consistent,* (2) $\vdash_T \mathtt{Con}_T \to \neg\Box\,\mathtt{Con}_T$.

Proof. $D1$–$D3$ were proved for PA in **7.1**. (1) follows from (2). Assume $\vdash_T \mathtt{Con}_T$. Then $\vdash_T \Box\,\mathtt{Con}_T$ by $D1$, as well as $\vdash_T \neg\Box\,\mathtt{Con}_T$ by (2). Thus, T is inconsistent. To verify (2), let γ be a fixed point of $\neg\Box(x)$, i.e.,

(∗) $\gamma \equiv_T \neg\Box\gamma$ $(\equiv \Box\gamma \to \bot)$.

By Lemma 3.1(b) with $\alpha = \bot$, we obtain $\gamma \equiv_T \Box\bot \to \bot \equiv \neg\Box\bot = \mathtt{Con}_T$. Replacing γ in (∗) with $\mathtt{Con}_T$ gives $\mathtt{Con}_T \equiv_T \neg\Box\,\mathtt{Con}_T$. Half of this is the claim (2). ❏

Thus, by (1), no sufficiently strong consistent theory can prove its own consistency. In particular, $\nvdash_{\mathsf{PA}} \mathtt{Con}_{\mathsf{PA}}$ as long as PA is consistent which is assumed throughout this book and is a minimal assumption for a far-reaching metamathematics. The above proof shows that $\mathtt{Con}_T$ is the only fixed point of $\neg\,\mathtt{bwb}_T$ modulo T. Actually, it shows a bit more, namely

(3) $\mathtt{Con}_T \equiv_T \neg\Box\,\mathtt{Con}_T$.

This strengthens (2), but only by a little: $\neg\Box\,\mathtt{Con}_T \vdash_T \mathtt{Con}_T$ is just a special case of

(4) $\neg\Box\alpha \vdash_T \mathtt{Con}_T$ (equivalently, $\neg\,\mathtt{Con}_T \vdash_T \Box\alpha$), for every $\alpha \in \mathcal{L}$.

This follows from $\bot \vdash_T \alpha$, since $\neg\,\mathtt{Con}_T \equiv \Box\bot \vdash_T \Box\alpha$ by $D0$. (4) reflects in T 'If T is inconsistent then every formula is provable'. From (1) and (3) we get in particular $\nvdash_{\mathsf{PA}} \neg\Box_{\mathsf{PA}}\,\mathtt{Con}_{\mathsf{PA}}$, although '$\mathtt{Con}_{\mathsf{PA}}$ is unprovable in PA' is true according to (1) (again we tacitly use the consistence of PA). $\neg\Box_{\mathsf{PA}}\,\mathtt{Con}_{\mathsf{PA}}$ reflects '$\mathtt{Con}_{\mathsf{PA}}$ is unprovable in PA'; hence $\nvdash_{\mathsf{PA}} \neg\Box_{\mathsf{PA}}\,\mathtt{Con}_{\mathsf{PA}}$ is just another formulation of the second incompleteness theorem.

The above claims hold independently of the "truth content" of the sentences provable in T. Namely, a consequence of the second incompleteness theorem is the existence of consistent theories $T \supseteq \mathsf{PA}$ in which along with claims true in $\mathcal{N}$ also false ones are provable, i.e., in which truth and untruth live in peaceful coexistence with each other. Such "dream theories" are highly rich in content, for all of them include ordinary number theory. An example is $\mathsf{PA}^\perp := \mathsf{PA} + \neg\,\mathtt{Con}_{\mathsf{PA}}$. This theory is consistent because *the consistency of* $\mathsf{PA}^\perp$ *is equivalent to the unprovability of* $\mathtt{Con}_{\mathsf{PA}}$ *in* PA. The italicized sentence is even provable in PA, as (5) below will show. By the formalized deduction theorem (Exercise 5 in **7.1**), $\Box_{T+\alpha}\perp \equiv_T \Box(\alpha \to \perp) \equiv \Box\neg\alpha$; hence $\neg\Box_{T+\alpha}\perp \equiv_T \neg\Box\neg\alpha$ $(\equiv \Diamond\alpha)$, and consequently,

(5) $\mathtt{Con}_{T+\alpha} \equiv_T \neg\Box\neg\alpha$ (in particular, $\mathtt{Con}_{\mathsf{PA}^\perp} \equiv_{\mathsf{PA}} \neg\Box_{\mathsf{PA}}\,\mathtt{Con}_{\mathsf{PA}}$).

The special cases under (5) and (3) for $T = \mathsf{PA}$ now clearly yield

(6) $\mathtt{Con}_{\mathsf{PA}} \equiv_{\mathsf{PA}} \mathtt{Con}_{\mathsf{PA}^\perp}$ (hence also $\mathtt{Con}_{\mathsf{PA}} \equiv_{\mathsf{PA}^\perp} \mathtt{Con}_{\mathsf{PA}^\perp}$).

Put together, $\mathsf{PA}^\perp$ contains ordinary number theory as known to us, but also proves the indubitably false sentence $\mathtt{bwb}_{\mathsf{PA}}(\ulcorner 0 \neq 0\urcorner)$. Moreover, because of $\vdash_{\mathsf{PA}^\perp} \neg\,\mathtt{Con}_{\mathsf{PA}}$ and hence $\vdash_{\mathsf{PA}^\perp} \neg\,\mathtt{Con}_{\mathsf{PA}^\perp}$ by (6), $\mathsf{PA}^\perp$ proves (the reflection of) its own inconsistency, although along with PA also $\mathsf{PA}^\perp$ is consistent. It claims to have a mysterious proof of $\perp$. Thus, consistency of T can have a different meaning within T and seen from outside, just as the meanings of *countable* diverge, depending on whether one is situated in ZFC or is looking at it from outside. One may even say that $\mathsf{PA}^\perp$ is lying to us with the claim $\neg\,\mathtt{Con}_{\mathsf{PA}^\perp}$.

We learn from the preceding that the extension $T + \mathtt{Con}_T$ of a consistent theory T need not be consistent. $T = \mathsf{PA}^\perp$ is a concrete example, and in fact only one of arbitrarily many others. More on the meaning of $\neg\,\mathtt{Con}_T$ will be said in Theorem 3.4.

We now discuss what is, along with (3), the most famous example of a self-referential sentence. Clearly, a fixed point α of $\Box(x)$ claims just its own provability, that is, $\alpha \equiv_T \Box\alpha$. A trivial example is $\alpha = \top$, because $\vdash_T \Box\top \to \top$, and since $\vdash_T \top$, clearly $\vdash_T \Box\top$, so that $\top \equiv_T \Box\top$. What is surprising here is that $\top$ turns out to be the only fixed point of $\Box(x)$ modulo T. By $D4^\circ$ below, $\vdash_T \Box\alpha \to \alpha$ implies $\vdash_T \alpha$ and so $\alpha \equiv_T \top$ (which confirms the uniqueness), although one might perhaps expect $\vdash_T \Box\alpha \to \alpha$ for all $\alpha \in \mathcal{L}^0$ because $\Box\alpha \to \alpha$ is intuitively true.

Theorem 3.3 (Löb's theorem). *Take T to satisfy $D1$–$D3$ and the fixed point lemma. Then T has the properties*

$$D4\colon \vdash_T \Box(\Box\alpha\to\alpha)\to\Box\alpha,\quad D4^\circ\colon \vdash_T \Box\alpha\to\alpha \;\Rightarrow\; \vdash_T \alpha\quad(\alpha\in\mathcal{L}^0).$$

Proof. Let γ be a fixed point of $\Box(x)\to\alpha$, i.e., $\gamma\equiv_T\Box\gamma\to\alpha$. Then $\gamma\equiv_T\Box\alpha\to\alpha$ by Lemma 3.1(b). This and $D0$ imply $\Box\gamma\equiv_T\Box(\Box\alpha\to\alpha)$. Lemma 3.1(a) states $\Box\gamma\equiv_T\Box\alpha$, hence $\Box\alpha\equiv_T\Box(\Box\alpha\to\alpha)$. Half of this is $D4$. Now suppose $\vdash_T\Box\alpha\to\alpha$. Then by $D1$, $\vdash_T\Box(\Box\alpha\to\alpha)$. Using $D4$ results in $\vdash_T\Box\alpha$, and $\vdash_T\Box\alpha\to\alpha$ yields $\vdash_T\alpha$, thus proving $D4^\circ$. ❏

$D4$ reflects just $D4^\circ$ in T. One application of Löb's theorem is an extremely easy proof of $\nvdash_{\mathsf{PA}}\mathtt{Con}_{\mathsf{PA}}$. Indeed, $\vdash_{\mathsf{PA}}\mathtt{Con}_{\mathsf{PA}}$ ($\equiv\Box\bot\to\bot$) implies $\vdash_{\mathsf{PA}}\bot$ by $D4^\circ$. That's all. Similarly, $D4$ implies (2) for $\alpha=\bot$ by contraposition. Thus, Löb's theorem is stronger than Gödel's second incompleteness theorem, which is not obvious at first glance.

Unlike $\mathsf{PA}^\bot$, $\mathsf{PA}+\mathtt{Con}_{\mathsf{PA}}$ conforms to truth (in $\mathcal{N}$). Unfortunately it is not quite clear what $\mathtt{Con}_{\mathsf{PA}}$ means in number-theoretic terms. This is clear, however, for an arithmetical statement discovered by Paris and Harrington (see [Bar]) that implies $\mathtt{Con}_{\mathsf{PA}}$; this statement is provable in ZFC but not in PA. Since then, many such sentences have been found, mostly of a combinatorial nature. A popular example is

Goodstein's theorem. *Every Goodstein sequence ends in* 0.

A *Goodstein sequence* is a number sequence $(a_n)_{n\in\mathbb{N}}$, with arbitrary a_0 given in advance, such that a_{n+1} is obtained from a_n as follows: Let $b_n=n+2$, so that $b_0=2$, $b_1=3$, etc. Expand a_n in b-adic base for $b:=b_n$, so that for suitable k,

$$(*)\quad a_n=\textstyle\sum_{i\leqslant k}b^{k-i}c_i,\ \text{with } 0\leqslant c_i<b.$$

Also the powers $k-i$ are represented in b-adic form, so too the powers of powers, and so on. Now replace b everywhere with $b+1$ ($=b_{n+1}$) and subtract 1 from the output. The result is a_{n+1}. The table below gives an example beginning with $a_0=11$; already a_6 has the value 134 217 727.

$a_0=11=2^{2+1}+2+1$	$2\rightsquigarrow 3$	$3^{3+1}+3+1=85$
$a_1=84=3^{3+1}+3$	$3\rightsquigarrow 4$	$4^{4+1}+4=1028$
$a_2=1027=4^{4+1}+3$	$4\rightsquigarrow 5$	$5^{5+1}+3=15\,628$
$a_3=15\,627=5^{5+1}+2$	$5\rightsquigarrow 6$	$6^{6+1}+2=279\,938$
$a_4=279\,937=6^{6+1}+1$	$6\rightsquigarrow 7$	$7^{7+1}+1=5\,764\,802$

As one sees from this example, a_n initially increases enormously, and it is hardly believable that the sequence ever starts to decrease and ends in 0. But the proof of the theorem is not particularly difficult; one estimates a_n from above by the ordinal number λ_n, which, crudely put, results from a_n on replacing the basis b in $(*)$ by ω. With some ordinal arithmetic it can readily be shown that $\lambda_{n+1} < \lambda_n$ as long as $\lambda_n \neq 0$. Since there is no properly decreasing infinite sequence of ordinal numbers (these are well-ordered), the sequence $(a_n)_{n\in\mathbb{N}}$ must eventually end in 0. For more detailed information see for instance [HP].

Many metatheoretic properties can be expressed using the provability operator $\Box$ in T, often using sentence schemata. The following ones turn out to be equivalent and facilitate a better understanding of the meaning of $\neg\,\mathtt{Con}_T$ within T. None of these properties hold for a consistent T from the outside (Theorem 6.5.1′), but all of them are provable in $T = \mathsf{PA}^{\perp}$.

(i)	$\neg\,\mathtt{Con}_T$:	$\Box\perp$	(provable inconsistency),
(ii)	SyComp :	$\Box\alpha \vee \Box\neg\alpha$	(syntactic completeness),
(iii)	SeComp :	$\alpha \to \Box\alpha$	(semantic completeness),
(iv)	ω-Comp :	$\forall x\Box[\varphi(x)] \to \Box\forall x\varphi(x)$	(ω-completeness).

Theorem 3.4. *The properties* (i)–(iv) *are all equivalent in a theory T satisfying the properties named at the beginning of this section.*

Proof. By (4) (i)⇒(ii),(iii),(iv) are clear. (ii)⇒(i): By Rosser's theorem formulated in T (see **7.5**), $\mathtt{Con}_T \vdash_T \neg\Box\alpha \wedge \neg\Box\neg\alpha$ for some α. Thus, $\Box\alpha \vee \Box\neg\alpha \vdash_T \neg\,\mathtt{Con}_T$. (iii)⇒(i): For $\alpha := \mathtt{Con}_T$, SeComp and (2) yield $\alpha \vdash_T \Box\alpha, \neg\Box\alpha$ and so $\vdash_T \neg\alpha$. (iv)⇒(i): By (3) in **7.2**, we obtain $\neg\,\mathtt{bew}_T(x, \ulcorner\perp\urcorner) \vdash_T \Box[\neg\,\mathtt{bew}_T(x, \ulcorner\perp\urcorner)]$, for $\neg\,\mathtt{bew}_T(x, \ulcorner\perp\urcorner)$ is Σ_1. Hence,

$$\mathtt{Con}_T = \forall x\neg\,\mathtt{bew}_T(x, \ulcorner\perp\urcorner) \vdash_T \forall x\Box[\neg\,\mathtt{bew}_T(x, \ulcorner\perp\urcorner)].$$

ω-Comp and (2) yield $\mathtt{Con}_T \vdash_T \Box\forall x\neg\,\mathtt{bew}(x, \ulcorner\perp\urcorner) = \Box\,\mathtt{Con}_T \vdash_T \neg\,\mathtt{Con}_T$. Therefore, $\vdash_T \neg\,\mathtt{Con}_T$. ❑

Remark. $\mathtt{Con}_T$ is also equivalent in T to other properties, for example to the schema $\Box\alpha \to \alpha$ for Π_1-formulas α (the *local* Π_1-reflection principle) as well as the *uniform* Π_1-reflection principle $\forall x\Box[\alpha(x)] \to \forall x\alpha(x)$ for Π_1-formulas α. Both the theorems of Paris–Harrington and of Goodstein are equivalent in PA to the uniform Σ_1-reflection, or equivalently, to the consistency of PA plus all true Π_1-sentences; see e.g. [Bar, D8].

Define inductively $T^0 = T$ and $T^{n+1} = T^n + \mathtt{Con}_{T^n}$. This *n-times-iterated consistency extension* T^n can be written as $T^n = T + \neg\Box^n\bot$ with $\Box = \mathtt{bwb}_T$, $\Box^0\alpha = \alpha$ and $\Box^{n+1}\alpha = \Box\Box^n\alpha$ (Exercise 3). Thus, the consistency of T^n can be expressed by an iterated consistency statement on T. Let $T^\omega := \bigcup_{n\in\omega} T^n$. Since $T^n \subseteq T^{n+1}$ and $T^n = T + \neg\Box^n\bot$ (hence $T^\omega = T \cup \{\neg\Box^n\bot \mid n \in \omega\}$), the following three items are equivalent:

(i) T^ω is consistent, (ii) T^n is consistent for all n, (iii) $\nvdash_T \Box^n\bot$ for all n.

Like $\mathsf{PA}^1 = \mathsf{PA} + \mathtt{Con}_{\mathsf{PA}}$, also PA^ω conforms to truth looking at PA from outside. When considered more closely, this means only that PA^ω is relatively consistent with respect to ZFC. In other terms, $\vdash_{\mathsf{ZFC}} \mathtt{Con}_{\mathsf{PA}^\omega}$. The argument (to be formalized in ZFC) runs as follows: $\vdash_{\mathsf{PA}^\omega} \bot$ implies $\vdash_{\mathsf{PA}^n} \bot$ for some n, as was noticed above, hence $\vdash_{\mathsf{PA}} \Box^n\bot$. But this is impossible, as is seen by a repeated application of $D1^*$ (p. 271) on PA.

Exercises

1. Prove $D4^\circ$ for T by applying Theorem 3.2 to $T' = T + \neg\alpha$.
2. Show by means of Löb's theorem that $\mathtt{Con}_{\mathsf{PA}} \to \neg\Box\neg\,\mathtt{Con}_{\mathsf{PA}}$ is unprovable in PA, although this formula is true if seen from outside.
3. Let T^n recursively be defined as in the text above. Prove that $T^n = T + \neg\Box^n\bot$ and $\mathtt{Con}_{T^n} \equiv_T \neg\Box^{n+1}\bot$, where $\Box$ is $\mathtt{bwb}_T$.
4. Show that $\vdash_{\mathsf{ZFC}} \Box_{\mathsf{PA}}\alpha \to \alpha$ for all arithmetical sentences α from $\mathcal{L}_\in$ (the $\mathcal{L}_\in$-sentences relativized to ω).

7.4 The Provability Logic G

In **7.3** first-order logic was hardly required. It comes then as no surprise that many of the results there can be obtained propositionally, more precisely, in a certain modal propositional calculus. This calculus contains alongside $\wedge, \neg$ the falsum symbol $\bot$, and a further unary connective $\Box$ to be interpreted as the proof operator in $\mathcal{L}_{ar}$, denoted by $\Box$ as well. First we define a propositional language $\mathcal{F}_\Box$, whose formulas are denoted by H, G, F: (a) the variables $p_1, p_2, \ldots$ from PV (page 4) and $\bot$ belong to $\mathcal{F}_\Box$; (b) if H, G belong to $\mathcal{F}_\Box$ then so too $(H \wedge G)$, $\neg H$, and $\Box H$.

No other strings belong to $\mathfrak{F}_\Box$ in this context. $H \vee G$, $H \to G$, and $H \leftrightarrow G$ are defined as in **1.4**, $\top := \neg\bot$. Further, set $\Diamond H := \neg\Box\neg H$ and define recursively $\Box^0 H = H$, $\Box^{n+1}H = \Box\Box^n H$. Let G be the set of those formulas in $\mathfrak{F}_\Box$ derivable using substitution in $\mathfrak{F}_\Box$, modus ponens MP, and the rule MN: $H/\Box H$ from the tautologies of two-valued propositional logic, augmented by the axioms (called also the G-axioms)

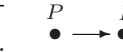
$$\Box(p \to q) \to \Box p \to \Box q, \quad \Box p \to \Box\Box p,[4] \quad \Box(\Box p \to p) \to \Box p.$$

For $H \in \mathsf{G}$ we mostly write $\vdash_\mathsf{G} H$ (read "H is derivable in G"). Rule MN corresponds to $D1$. The first G-axiom reflects $D2$, the middle $D3$, and the last (called *Löb's formula*) $D4$, hence the name *provability logic*. The connection between G and PA is described in **7.5**. Here we are concerned with the modal logic G and its *Kripke semantics*. For simplicity, we restrict ourselves to finite *Kripke frames*, which are just finite directed graphs. We can do so, since all modal logics considered here have the finite model property. We begin without further ado with the following

Definition. A G-*frame* or *Kripke frame for* G is a finite poset $(g, <)$. A *valuation* is a mapping w that assigns to every variable p a subset wp of g. The relation $P \Vdash H$, dependent on w, between points $P \in g$ and formulas $H \in \mathfrak{F}_\Box$ (read "P accepts H") is defined inductively by

$$P \Vdash p \text{ iff } P \in wp, \qquad P \nVdash \bot, \qquad P \Vdash H \wedge G \text{ iff } P \Vdash H \;\&\; P \Vdash G,$$
$$P \Vdash \neg H \text{ iff } P \nVdash H, \quad P \Vdash \Box H \text{ iff } P' \Vdash H \text{ for all } P' > P.$$

These conditions easily imply $P \Vdash \Diamond H$ iff $P' \Vdash H$ for some $P' > P$, and $P \Vdash H \to G$ iff $P \Vdash H \Rightarrow P \Vdash G$. If $P \Vdash H$ for all w and all $P \in g$, we write $g \vDash H$ and say H *holds in* g. If $g \vDash H$ for all G-frames g, we write $\vDash_\mathsf{G} H$ and say H *is* G-*valid*. The G-frame on the right, consisting of two points P, P' with $P < P'$, shows that $\nvDash_\mathsf{G} p \to \Box p$. Indeed, let $wp = \{P\}$. Then $P \Vdash p$, but $P \nVdash \Box p$ because $P' \nVdash p$. Note also $\nvDash_\mathsf{G} \Box p \to p$, for $P' \nVdash p$ but $P' \Vdash \Box p$ because there is no $P'' > P'$.



We may tacitly assume that G-frames are *initial* (have a smallest point), for $g \vDash H$ is verified pointwise. We write $H \equiv_\mathsf{G} H'$ for $\vDash_\mathsf{G} H \leftrightarrow H'$. It is readily seen that $\equiv_\mathsf{G}$ is a congruence in $\mathfrak{F}_\Box$ that extends the usual logical equivalence conservatively. For instance, $\neg\Box H \equiv_\mathsf{G} \neg\Box\neg\neg H \equiv_\mathsf{G} \Diamond\neg H$. Many more equivalences are presented in the following examples. These will later be translated into statements about self-reference.

[4] This axiom is dispensable; it is provable from the remaining, see e.g. [Boo] or [Ra1].

Examples. (a) Let g be an arbitrary G-frame. Although always $P \nVdash \bot$, we have $P \Vdash \Box\bot$, provided P is maximal in g, that is, no $Q > P$ exists. Likewise, $\Box\neg\Box\bot$ is accepted precisely at the maximal points of g. Thus, $\Box\bot \equiv_{\mathsf{G}} \Box\neg\Box\bot$, or equivalently, $\neg\Box\bot \equiv_{\mathsf{G}} \Diamond\Box\bot$ $(= \neg\Box\neg\Box\bot)$. This reflects in G the second incompleteness theorem, as will be seen in **7.5**.

(b) Let $\{P_0, \ldots, P_n\}$ be the ordered G-frame with $P_n < \cdots < P_0$. Clearly, $P_0 \Vdash \Box^m\bot$ for each $m > 0$. Induction on n shows that $P_n \Vdash \Box^m\bot$ for all $m > n$, but $P_n \nVdash \Box^n\bot$, and therefore $P_n \nVdash \Box^{n+1}\bot \to \Box^n\bot$. Hence, $\nvDash_{\mathsf{G}} \Box^{n+1}\bot \to \Box^n\bot$, and a fortiori $\nvDash_{\mathsf{G}} \Box^n\bot$ and $\nvDash_{\mathsf{G}} \neg\Box^{n+1}\bot$, for all n.

(c) $\vDash_{\mathsf{G}} \Box(\Box p \to p) \to \Box p$. For take an arbitrary g and $P \in g$. If $P \nVdash \Box p$ then there is, since g is finite, some $Q > P$ with $Q \Vdash \neg p$ and $Q' \Vdash p$ for all $Q' > Q$. Thus $Q \Vdash \Box p$; hence $Q \nVdash \Box p \to p$ and so $P \nVdash \Box(\Box p \to p)$. Consequently, $P \Vdash \Box(\Box p \to p) \to \Box p$, which proves our claim. Note also that $\vDash_{\mathsf{G}} \Box p \to \Box\Box p$. Only the transitivity of $<$ is relevant for the proof.

(d) $\vDash_{\mathsf{G}} \neg\Box^{n+1}\bot \to \Diamond R_n$, where $R_n := \bigwedge_{i=1}^{n}(\Box p_i \to p_i)$. For let $P \in g$, $P \Vdash \neg\Box^{n+1}\bot$. Then there must be a chain $P = P_0 < P_1 < \cdots < P_{n+1}$ in g. Now, it is a nice separate exercise to verify that each conjunct of R_n fails to be accepted by at most one of the $n+1$ points $P_1, \ldots, P_{n+1}$. Thus, at least one of these accepts all conjuncts. In other words, $P_i \Vdash R_n$ for some $i > 0$; hence $P \Vdash \Diamond R_n$. This nontrivial example will essentially be employed in the proof of Theorem 7.1.

By induction on $\vdash_{\mathsf{G}} H$ one easily proves $\vdash_{\mathsf{G}} H \;\Rightarrow\; \vDash_{\mathsf{G}} H$ (soundness of Kripke semantics for $\vdash_{\mathsf{G}}$). Example (c) is a part of the initial step. The induction steps over the rules are easy. For instance, $g \vDash H$ clearly implies $g \vDash \Box H$. The converse, $\vDash_{\mathsf{G}} H \Rightarrow \vdash_{\mathsf{G}} H$, holds as well. Thus, $\vdash_{\mathsf{G}} H$ can be confirmed by proving $\vDash_{\mathsf{G}} H$, and vice versa. This is the content of

Theorem 4.1 (Completeness of Kripke semantics for G). *For each formula H from $\mathfrak{F}_\Box$ it holds that $\vdash_{\mathsf{G}} H \;\Leftrightarrow\; \vDash_{\mathsf{G}} H$.*

The nontrivial direction $\Leftarrow$ follows directly from the finite model property of G, i.e., each $H \notin \mathsf{G}$ is falsified or refuted by some finite G-frame, proved, for example, in [Boo], [Ra1], and [CZ]. For the relatively simple formulas considered here, $\vDash_{\mathsf{G}} H$ is in general more easily checked than $\vdash_{\mathsf{G}} H$.

Both the formulas provable in G and those refutable are clearly recursively enumerable, thanks to the finite model property of G. Thus, in analogy to Exercise 2 in **3.6**, we obtain

Theorem 4.2. G *is decidable.*

Remark. The finite model property, decidability, and some other properties such as interpolation can all be proved in one move, see e.g. [Ra2]. An important fragment of G is $\mathsf{G}^0 := \mathsf{G} \cap \mathfrak{F}^0_\Box$, where $\mathfrak{F}^0_\Box$ denotes the set of variable-free formulas of $\mathfrak{F}_\Box$. The formulas $\neg\Box^n\bot$ ($\equiv_{\mathsf{G}} \Diamond^n\top$) form a Boolean base in G^0. One proves this most easily by showing that G^0 is complete with respect to all (totally) ordered G-frames, including the infinite ones, and applying Theorem 5.2.3 accordingly.

Exercises

1. Let g be *any* finite Kripke frame (a graph) that satisfies the axioms of G. Show that g is necessarily a poset. Only this fact justifies the identification of G-frames with posets.
2. Prove $\vdash_{\mathsf{G}} \Box p \to \Box(\Box p \to p)$, the inverse of Löb's formula. (Only the first of the three G-axioms is needed in the proof.)

7.5 The Modal Treatment of Self-Reference

Let T be a theory as in **7.3**. A mapping $\imath$ from PV to $\mathcal{L}^0$ with $p_i^\imath = \alpha_i$ is called an *insertion*. $\imath$ can be extended to the whole of $\mathfrak{F}_\Box$ by the clauses $\bot^\imath = \bot$, $(\neg H)^\imath = \neg H^\imath$, $(H \wedge G)^\imath = H^\imath \wedge G^\imath$, and $(\Box H)^\imath = \Box H^\imath$ $(= \mathtt{bwb}_T(\ulcorner H^\imath \urcorner))$. Briefly speaking, $H^\imath$ results from $H(p_1, \ldots, p_n)$ by replacing the p_ν by the sentences α_ν from $\mathcal{L}$. For instance, if $p^\imath = \alpha$ then $(\Box p \wedge \neg\Box\bot)^\imath = \Box\alpha \wedge \neg\Box\bot$, and $(\neg\Box\bot)^\imath = \neg\Box\bot = \mathtt{Con}_T$. The following lemma shows that $\vdash_{\mathsf{G}}$ is "sound" for $\vdash_T$. Already this simple fact considerably simplifies proofs about self-referential statements.

Lemma 5.1. *For each* H *with* $\vdash_{\mathsf{G}} H$ *and each insertion* $\imath$, $\vdash_T H^\imath$.

Proof by induction on $\vdash_{\mathsf{G}} H$. If H is a propositional tautology then $H^\imath \in \mathit{Taut}_{\mathcal{L}} \subseteq T$. If H is one of the modal axioms of G, then $\vdash_T H^\imath$ by $D2$, $D3$, or $D4$. If $\vdash_{\mathsf{G}} H$ and $\sigma : \mathfrak{F}_\Box \to \mathfrak{F}_\Box$ is a substitution, then $\vdash_T H^{\sigma\imath}$, since $H^{\sigma\imath} = H^{\imath'}$ with $\imath' : p \mapsto p^{\sigma\imath}$, and $\vdash_T H^{\imath'}$ holds by the induction hypothesis. As regards the induction step over MP, consider $(F \to G)^\imath = F^\imath \to G^\imath$. Finally, if MN is applied, and $\vdash_T H^\imath$ by the induction hypothesis, then $\vdash_T \Box H^\imath = (\Box H)^\imath$, due to $D1$. ❏

Example 1. We prove (3) of Theorem 3.2 with the calculus $\vdash_{\mathsf{G}}$. By Lemma 5.1 and Theorem 4.1 it suffices to show that $\vDash_{\mathsf{G}} \neg\Box\bot \leftrightarrow \neg\Box\neg\Box\bot$.

This holds by Example (a) in **7.4**. Next example: $\vDash_{\mathsf{G}} \Box(p \leftrightarrow \Diamond p) \rightarrow \neg\Diamond p$ is easily confirmed. Thus, $\vdash_T \Box(\alpha \leftrightarrow \Diamond\alpha) \rightarrow \neg\Diamond\alpha$. This tells us (if everything is related to $T = \mathsf{PA}$) that a sentence claiming its own consistency with PA is incompatible with PA, which hardly seems plausible. Even the converse is provable in PA since $\vDash_{\mathsf{G}} \neg\Diamond p \rightarrow \Box(p \leftrightarrow \Diamond p)$.

We now explain certain facts that expand upon the reasoning of above. For PA and related theories, the converse of Lemma 5.1 holds as well. That is to say, the derivability conditions and Löb's theorem already contain everything worth knowing about self-referential formulas or schemes. This is essentially the content of Theorem 5.2. For the subtle proofs of Theorems 5.2, 5.4, and 5.5, the reader is referred to [Boo].

Theorem 5.2 (Solovay's completeness theorem). *For all* $H \in \mathfrak{F}_{\Box}$: $\vdash_{\mathsf{G}} H$ *(equivalently* $\vDash_{\mathsf{G}} H$*) if and only if* $\vdash_{\mathsf{PA}} H^{\imath}$ *for all insertions* $\imath$.

Example 2 (applications). (a) $\nvdash_{\mathsf{PA}} \Box^{n+1}\bot \rightarrow \Box^{n}\bot$ because by Example (b) in **7.4**, $\nvDash_{\mathsf{G}} \Box^{n+1}\bot \rightarrow \Box^{n}\bot$. In particular, $\nvdash_{\mathsf{PA}} \mathtt{Con}_{\mathsf{PA}}$ $(\equiv \Box\bot \rightarrow \bot)$. (b) $\nvdash_{\mathsf{PA}} \neg\Box^{n+1}\bot$, since $\nvDash_{\mathsf{G}} \neg\Box^{n+1}\bot$. (c) It is easily verified with the 2-point frame on page 285 that $\nvDash_{\mathsf{G}} \neg\Box p \rightarrow \Box\neg\Box p$, in particular $\nvDash_{\mathsf{G}} \neg\Box\bot \rightarrow \Box\neg\Box\bot$. Therefore, $\nvdash_{\mathsf{PA}} \mathtt{Con}_{\mathsf{PA}} \rightarrow \Box\,\mathtt{Con}_{\mathsf{PA}}$. (d) $\mathsf{PA}_n := \mathsf{PA} + \Box^{n}\bot$ is consistent for $n > 0$ by (b), but is ω-inconsistent. Otherwise, by $D1^*$ (page 271), $\vdash_{\mathsf{PA}_n} \Box^{n}\bot \Rightarrow \vdash_{\mathsf{PA}_n} \Box^{n-1}\bot \Rightarrow \cdots \Rightarrow \vdash_{\mathsf{PA}_n} \bot$, contradicting $\nvdash_{\mathsf{PA}_n} \bot$. Since $\vdash_{\mathsf{PA}} \Box^{n}\bot \rightarrow \Box^{n+1}\bot$ by $D3$, we get $\mathsf{PA}_n \supseteq \mathsf{PA}_{n+1}$, and since $\mathsf{PA}_n \neq \mathsf{PA}_{n+1}$ by (a), we have $\mathsf{PA}_0 \supset \mathsf{PA}_1 \supset \cdots \supset \mathsf{PA}$. Observe that PA_1 is just $\mathsf{PA}^{\bot}$.

Note also the following: Since $\nvDash_{\mathsf{G}} \Box p \rightarrow p$, there must be some $\alpha \in \mathcal{L}^0_{ar}$ such that $\nvdash_{\mathsf{PA}} \Box\alpha \rightarrow \alpha$. Indeed, choose $\alpha = \bot$. The above examples point out that Theorem 5.2 and the decidability of G are very efficient tools in deciding the provability of self-referential statements.

Many other theories have the same provability logic as PA, where in general a modal propositional logic H is the *provability logic for* T when the analogue of Theorem 5.2 holds with respect to T and H. For some theories, the provability logic may be a proper extension of G. For example, the ω-inconsistent theory PA_n from Example 2(d) has the provability logic $\mathsf{G}_n := \mathsf{G} + \Box^{n}\bot$, the smallest extension of G closed under all rules of G with the additional axiom $\Box^{n}\bot$ (Exercise 1; note that G_0 is inconsistent). By the following theorem, which will be proved in **7.7**, other extensions of G to be considered as provability logics are out of the question.

Theorem 5.3 ([Vi1]). *Let T be at least as strong as* PA. *Then*

(a) *If T^ω* (page 284) *is consistent, then* G *is the provability logic of T;*

(b) *if $\vdash_{T^\omega} \bot$ and n is minimal such that $\vdash_{T^n} \bot$, then T's provability logic is G_n.*

The formulas $H \in \mathfrak{F}_\Box$ such that $\mathcal{N} \models H^\imath$ for all insertions $\imath$ in $\mathcal{L}_{ar}$ can also be surprisingly easily characterized. All $H \in \mathsf{G}$ are obviously included; but in addition also $\Box p \to p$ belongs to this sort of formula, because $\mathcal{N} \models \Box\alpha \to \alpha$ for $\alpha \in \mathcal{L}^0_{ar}$. Indeed, if $\mathcal{N} \models \Box\alpha$ then *there is* some n that codes a proof of α in PA, hence $\mathcal{N} \models \alpha$.

Let GS ($\supseteq$ G) be the set of all formulas in $\mathfrak{F}_\Box$ that can be obtained from those in $\mathsf{G} \cup \{\Box p \to p\}$ using substitution and modus ponens only. Induction in GS readily yields $H \in \mathsf{GS} \Rightarrow \mathcal{N} \models H^\imath$ for all $\imath$. Again, the converse holds as well:

Theorem 5.4 ([So]). *$H \in$ GS if and only if $\mathcal{N} \models H^\imath$ for all insertions $\imath$.*

GS is decidable as well, because it can be shown that $H \in \mathsf{GS} \Leftrightarrow H^* \in \mathsf{G}$, where $H^* := [\bigwedge_{\Box G \in \mathrm{Sf}^\Box H}(\Box G \to G)] \to H$. Here $\mathrm{Sf}^\Box H$ is the set of subformulas of H of the form $\Box G$. Thus, Theorem 5.4 reduces the decidability of GS to that of G. Using this theorem, many questions concerning the relations between *provable* and *true* are effectively decidable. For instance,

$$H(p) := \neg\Box(\neg\Box\bot \to \neg\Box p \wedge \neg\Box\neg p) \notin \mathsf{GS}$$

is readily verified. Hence $\mathcal{N} \models \neg H(\alpha) \equiv \Box(\neg\Box\bot \to \neg\Box\alpha \wedge \neg\Box\neg\alpha)$ for some $\alpha \in \mathcal{L}^0_{ar}$ by Theorem 5.4. Translated into English: *It is provable in* PA *that the consistency of* PA *implies the independence of α for some sentence α.* This is exactly Rosser's theorem, which in this way turns out to be provable in PA. As was shown in [Be1], the box in the formulas $H \in \mathsf{GS}$ in Theorem 5.4 may denote $\mathtt{bwb}_T$ for *any* axiomatizable $T \supseteq \mathsf{PA}$, provided $T \subseteq Th\mathcal{N}$. However, if T proves false sentences (as does e.g. $\mathsf{PA}^\bot$) then GS has to be redefined in a feasible manner and is always decidable.

A variable p in H is called *modalized in H* if every occurrence of p is contained within the scope of a $\Box$, as is the case in $\neg\Box p$, $\neg\Box\neg p$, and $\Box(p \to q)$. By contrast, p is not modalized in $\Box p \to p$. Another particularly interesting theorem is

Theorem 5.5 (DeJongh–Sambin fixed point theorem). *Let p be modalized in $H(p, q_1, \ldots, q_n)$, $n \geqslant 0$. Then a formula $F = F(\vec{q})$ from $\mathfrak{F}_\Box$ can effectively be constructed such that*

(a) $F \equiv_{\mathsf{G}} H(F, \vec{q})$,

(b) $\vdash_{\mathsf{G}} \bigwedge_{i=1}^{2}[(p_i \leftrightarrow H(p_i, \vec{q})) \wedge \Box(p_i \leftrightarrow H(p_i, \vec{q}))] \rightarrow (p_1 \leftrightarrow p_2)$.

This theorem easily yields a corresponding result for theories T:

Corollary 5.6. *Let p be modalized in $H = H(p, \vec{q})$ and suppose T satisfies D1–D4. Then there is an $F = F(\vec{q}) \in \mathfrak{F}_\Box$ with $F(\vec{\alpha}) \equiv_T H(F(\vec{\alpha}), \vec{\alpha})$ for all $\vec{\alpha} = (\alpha_1, \ldots, \alpha_n)$, $\alpha_i \in \mathcal{L}^0$. For each $\vec{\alpha}$ there is only one $\beta \in \mathcal{L}^0$ modulo T such that $\beta \equiv_T H(\beta, \vec{\alpha})$.*

Proof. Choose F as in (a) of the theorem. Then $F(\vec{\alpha}) \equiv_T H(F(\vec{\alpha}), \vec{\alpha})$ by Lemma 5.1 ($\vec{q}^{\,\iota} = \vec{\alpha}$). To prove uniqueness let $\beta_i \equiv_T H(\beta_i, \vec{\alpha})$ for $i = 1, 2$. By $D1$, $\vdash_T (\beta_i \leftrightarrow H(\beta_i, \vec{\alpha})) \wedge \Box(\beta_i \leftrightarrow H(\beta_i, \vec{\alpha}))$. Inserting β_i for p_i and α_i for q_i in the formula under (b) in the theorem then yields $\vdash_T \beta_1 \leftrightarrow \beta_2$ by Lemma 5.1. ❏

Example 3. For $H = \neg\Box p$ ($n = 0$), $F = \neg\Box\bot$ is a "solution" of (a) in Theorem 5.5 because $\neg\Box\bot \equiv_{\mathsf{G}} \neg\Box(\neg\Box\bot)$. According to Corollary 5.6, $\mathtt{Con}_T$ ($= \neg\Box\bot$) is modulo T the only fixed point of $\neg\,\mathtt{bwb}_T$. This is just the claim of (3) from **7.3**.

Many special cases of the corollary represent older self-reference results from Gödel, Löb, Rogers, Jeroslow, and Kreisel, which, stated in terms of modal logic, concern fixed points of $\neg\Box p$, $\Box p$, $\neg\Box\neg p$, $\Box\neg p$, and $\Box(p \rightarrow q)$ in PA. Incidentally, one gets the fixed points of these formulas—namely $\neg\Box\bot$, $\top$, $\bot$, $\Box\bot$, and $\Box q$—according to a simple recipe. All first listed formulas are of the form $H = G\,\frac{\Box H'}{p}$, where p is not modalized in $G(p, \vec{q})$ and $H'(p, \vec{q})$ is chosen appropriately. In this case, $F = H\,\frac{G(\top, \vec{q})}{p}$ is the fixed point of H, as is seen after some calculation. For $H = \neg\Box p$ from Example 3 is $G = \neg p$. Thus, according to the recipe the fixed point is

$$F = \neg\Box p\,\tfrac{\neg\top}{p} = \neg\Box\neg\top \equiv_{\mathsf{G}} \neg\Box\bot.$$

For Kreisel's formula $\Box(p \rightarrow q)$ is $G = p$. Hence, it has the fixed point

$$F = \Box(p \rightarrow q)\,\tfrac{\top}{p} = \Box(\top \rightarrow q) \equiv_{\mathsf{G}} \Box q.$$

The recipe also works for $H = \Box p \rightarrow q$, by choosing $G = p \rightarrow q$. Hence $F = (\Box p \rightarrow q)\,\frac{\top \rightarrow q}{p} = \Box(\top \rightarrow q) \rightarrow q \equiv_{\mathsf{G}} \Box q \rightarrow q$ is the only fixed point of

H modulo T. Exactly this is the claim of Lemma 3.1(b), used in Gödel's second incompleteness theorem.

Exercises

1. Prove that the theory PA_n from Example 2(d) has the provability logic G_n.
2. Show that $\mathsf{PA}^n_\perp := \mathsf{PA}^n + \neg\,\mathtt{Con}_{\mathsf{PA}^n}$ equals $\mathsf{PA} + \Box^{n+1}\perp \wedge \neg\Box^n\perp$ and that it has the provability logic $\mathsf{G}_1 = \mathsf{G} + \Box\perp$. Here $\Box$ means $\Box_{\mathsf{PA}}$.
3. Prove that $\top$, $\perp$, and $\Box\perp$ are the fixed points of $\Box p$, $\neg\Box\neg p$, and $\Box\neg p$.
4. (Mostowski). Let $T \supseteq \mathsf{PA}$ be axiomatizable and suppose $\mathcal{N} \models T$. Show that there are two mutually independent Σ_1-sentences α, β in T, that is, $\alpha \to \beta$, $\alpha \to \neg\beta$, $\beta \to \alpha$, $\beta \to \neg\alpha$ (hence also α, β, $\neg\alpha$, and $\neg\beta$) are unprovable in T.

7.6 A Bimodal Provability Logic for PA

Hilbert remarked jokingly that the incompleteness phenomenon can be forcefully removed from the world by use of the so-called *ω-rule*

$$\rho_\omega : \ \frac{X \vdash \varphi(\underline{n}) \text{ for all } n}{X \vdash \forall x \varphi}.$$

ρ_ω has infinitely many premises. It is an easy exercise to derive with the aid of ρ_ω every sentence α valid in $\mathcal{N}$ from the axioms of PA, even from those of Q. Indeed, all sentences can (up to equivalence) be obtained from variable-free literals with $\wedge, \vee, \forall, \exists$, bypassing formulas with free variables. Due to the Σ_1-completeness of Q, all valid variable-free literals are derivable. The inductive steps for $\wedge, \vee, \exists$ are simple, applying Σ_1-completeness in the $\exists$-step once again. Only in the $\forall$-step is ρ_ω used.

Clearly, an unrestricted use of the infinitistic rule ρ_ω (in spite of its relevance for higher order arithmetic) contradicts Hilbert's own intention of giving mathematics a finitistic foundation. However, things look different if we restrict ρ_ω each time to a *single* application. In view of Remark 1 in **6.2**, we no longer distinguish between φ and $\dot{\varphi}$, so that φ itself is a number and $\ulcorner\varphi\urcorner = \underline{\varphi}$ is the corresponding Gödel term. Let us define

$$\mathit{1bwb}_{\mathsf{PA}}(\alpha) := (\exists\varphi{\in}\mathcal{L}^1_{ar})[\mathit{bwb}_{\mathsf{PA}}(\forall x\varphi \to \alpha)\ \&\ \forall n\ \mathit{bwb}_{\mathsf{PA}}(\varphi(\underline{n}))].$$

$\mathit{1bwb}_{\mathsf{PA}}$ is arithmetical, in fact it is Σ_3, for $\mathit{bwb}_{\mathsf{PA}}$ is Σ_1 and $\forall n\ \mathit{bwb}_{\mathsf{PA}}(\varphi(\underline{n}))$ is Π_1. We read $\mathit{1bwb}_{\mathsf{PA}}(\alpha)$ as "α is 1-provable." Let $\mathtt{1bwb}(x)$ be the Σ_3-formula in $\mathcal{L}_{ar}$ defining $\mathit{1bwb}_{\mathsf{PA}}$. Here let x be $\boldsymbol{v}_0$. Write $\boxed{1}\alpha$ for $\mathtt{1bwb}(\ulcorner\alpha\urcorner)$ and $\diamondsuit\!\!\!\!\!\!\!\!\!\!\boxed{1}\alpha$ for $\neg\boxed{1}\neg\alpha$. Clearly, $\Box\alpha$ for $\alpha \in \mathcal{L}^0_{ar}$ ($\Box = \Box_{\mathsf{PA}}$) can be read '$\mathsf{PA} + \neg\alpha$ is inconsistent', while $\boxed{1}\alpha$, by Lemma 6.1, formalizes '$\mathsf{PA} + \neg\alpha$ is ω-inconsistent'. Thus, $\diamondsuit\!\!\!\!\!\!\!\!\!\!\boxed{1}\top$ ($\equiv \neg\boxed{1}\bot$) means 'PA ($= \mathsf{PA} + \neg\bot$) is ω-consistent'. This explains the interest in the operator $\boxed{1}$.

If $\mathit{bwb}_{\mathsf{PA}}(\alpha)$ *then certainly* $\mathit{1bwb}_{\mathsf{PA}}(\alpha)$ (choose α for φ). The italicized statement is reflected in PA as '$\vdash_{\mathsf{PA}} \Box\alpha \to \boxed{1}\alpha$ for every $\alpha \in \mathcal{L}^0_{ar}$'. The converse fails, since $\nvdash_{\mathsf{PA}} \mathtt{Con}_{\mathsf{PA}}$, while $\mathtt{Con}_{\mathsf{PA}}$ is easily 1-provable: $\vdash_{\mathsf{PA}} \varphi(\underline{n})$ for all n, with $\varphi(x) := \neg\,\mathtt{bew}_{\mathsf{PA}}(x, \bot)$, and trivially $\vdash_{\mathsf{PA}} \forall x\varphi(x) \to \mathtt{Con}_{\mathsf{PA}}$. In what follows, some claims will not be proved in detail.

Define $\Omega := \{\varphi{\in}\mathcal{L}^1_{ar} \mid \vdash_{\mathsf{PA}} \varphi(\underline{n}) \text{ for all } n\}$. By its definition, Ω and hence also $\mathsf{PA}^\Omega := \mathsf{PA} + \Omega$ are formally Σ_3. As Theorem 6.2 will show, PA^Ω is properly Σ_3 and hence is no longer recursively axiomatizable.

Lemma 6.1. *The following properties are equivalent for* $\alpha \in \mathcal{L}^0_{ar}$:

(i) $\mathit{1bwb}_{\mathsf{PA}}(\alpha)$, (ii) $\vdash_{\mathsf{PA}^\Omega} \alpha$, (iii) $\mathsf{PA} + \neg\alpha$ *is* ω*-inconsistent.*

Proof. (i)⇒(ii) follows with a glance at the definitions (read (i) naively). (ii)⇒(iii): Let $\vdash_{\mathsf{PA}^\Omega} \alpha$. Since Ω is closed under conjunctions, there is some $\forall x\varphi(x) \in \Omega$ with $\forall x\varphi \vdash_{\mathsf{PA}} \alpha$, hence $\vdash_{\mathsf{PA}} \neg\alpha \to \exists x\neg\varphi$ and so $\vdash_{\mathsf{PA}+\neg\alpha} \exists x\neg\varphi$. Now, $\forall x\varphi \in \Omega$, therefore $\vdash_{\mathsf{PA}} \varphi(\underline{n})$ and a fortiori $\vdash_{\mathsf{PA}+\neg\alpha} \varphi(\underline{n})$, for all n. Thus, $\mathsf{PA} + \neg\alpha$ is ω-inconsistent. (iii)⇒(i): Let $\vdash_{\mathsf{PA}+\neg\alpha} \beta(\underline{n})$ for all n, but $\vdash_{\mathsf{PA}+\neg\alpha} \exists x\neg\beta$. Then $\vdash_{\mathsf{PA}} \forall x\beta \to \alpha$. With $\varphi(x) := \neg\alpha \to \beta(x)$ clearly $\vdash_{\mathsf{PA}} \varphi(\underline{n})$ for all n. Now, $\forall x\varphi \equiv \alpha \vee \forall x\beta \vdash_{\mathsf{PA}} \alpha$. Hence $\vdash_{\mathsf{PA}} \forall x\varphi \to \alpha$. Thus, altogether $\mathit{1bwb}_{\mathsf{PA}}(\alpha)$. ❏

Theorem 6.2 (the 1-provable Σ_3-completeness of PA). *All true* Σ_3*-sentences are 1-provable. Moreover, for every* β *of this kind,* $\vdash_{\mathsf{PA}} \beta \to \boxed{1}\beta$.

Proof. Let $\mathcal{N} \vDash \beta := \exists y\forall x\gamma(y, x)$ where $\gamma(y, x)$ is Σ_1. Then there is some m such that $\mathcal{N} \vDash \gamma(\underline{m}, \underline{n})$ for all n. Therefore, $\vdash_{\mathsf{PA}} \gamma(\underline{m}, \underline{n})$ for all n, because PA is Σ_1-complete. Hence, $\forall x\gamma(\underline{m}, x) \in \Omega$ and so $\vdash_{\mathsf{PA}^\Omega} \exists z\forall x\gamma$, or equivalently, $\mathit{1bwb}_{\mathsf{PA}}(\beta)$ by Lemma 6.1. Because of the provable Σ_1-completeness of PA, this argumentation is comprehensible in PA, so that also $\vdash_{\mathsf{PA}} \beta \to \boxed{1}\beta$. ❏

$D1$–$D4$ are also valid for the operator $\boxed{1}\colon \mathcal{L}^0_{ar} \to \mathcal{L}^0_{ar}$. Indeed, $D1$ holds because $\vdash_{\mathsf{PA}} \alpha \Rightarrow \vdash_{\mathsf{PA}} \Box\alpha \Rightarrow \vdash_{\mathsf{PA}} \boxed{1}\alpha$, and $D2$ formalizes (or reflects) '$\vdash_{\mathsf{PA}^\Omega} \alpha, \alpha \to \beta \Rightarrow \vdash_{\mathsf{PA}^\Omega} \beta$' in PA (observe Lemma 6.1). $D3$ follows from Theorem 6.2 with $\beta = \boxed{1}\alpha$. The proof of $D4$ in **7.3** uses, along with the fixed point lemma, only $D1$–$D3$; so $D4$ holds as well. Therefore, nearly everything said in **7.3** on $\Box$ applies also to $\boxed{1}$, including Theorem 3.2, which now reads $\nvdash_{\mathsf{PA}} \neg\boxed{1}\bot$ ($\equiv \Diamond\top$). To put it more concisely, although the consistency of PA is provable with the extended means, ω-consistency is not. Hence, this property, which is Π_3-definable according to Exercise 3 in **6.7**, cannot be Σ_3 by Theorem 6.2, and must therefore be properly Π_3. Equivalently, ω-inconsistency is properly Σ_3.

Alongside $\Box\alpha \to \boxed{1}\alpha$, there are other noteworthy interactions between $\Box$ and $\boxed{1}$, in particular $\vdash_{\mathsf{PA}} \neg\Box\alpha \to \boxed{1}\neg\Box\alpha$. This formalizes 'If $\nvdash_{\mathsf{PA}} \alpha$ then $\neg\Box\alpha$ is 1-provable'. To verify the latter notice that $\nvdash_{\mathsf{PA}} \alpha$ implies $\vdash_{\mathsf{PA}} \varphi(\underline{n})$ for all n, where $\varphi(x)$ is $\neg\,\mathtt{bew}_{\mathsf{PA}}(x, \ulcorner\alpha\urcorner)$, and since $\vdash_{\mathsf{PA}} \forall x\varphi \to \neg\Box\alpha$, we get $\vdash_{\mathsf{PA}} \boxed{1}\neg\Box\alpha$. On the other hand, $\vdash_{\mathsf{PA}} \neg\Box\alpha \to \Box\neg\Box\alpha$ fails in general; Example 2(c) in **7.5** yields a counterexample.

The language of the bimodal propositional logic GD now to be defined results from $\mathcal{F}_\Box$ by adding a further connective $\boxed{1}$ to $\mathcal{F}_\Box$, which is treated syntactically just as $\Box$. The axioms of GD are those of G stated both for $\Box$ and $\boxed{1}$, augmented by the axioms

$$\Box p \to \boxed{1}p \quad \text{and} \quad \neg\Box p \to \boxed{1}\neg\Box p.$$

The rules of GD are the same as those for G. Insertions $\imath$ to $\mathcal{L}^0_{ar}$ are defined as in **7.5**, but with the additional clause $(\boxed{1}H)^\imath = \boxed{1}H^\imath$, that is, $(\boxed{1}H)^\imath = \mathtt{1bwb}(\ulcorner H^\imath\urcorner)$. By the reasoning above, all axioms and rules of GD are sound. This proves (the easier) half of the following remarkable theorem from Dzhaparidze (1985):

Theorem 6.3. $\vdash_{\mathsf{GD}} H \Leftrightarrow \vdash_{\mathsf{PA}} H^\imath$ *for all insertions* $\imath$ *as defined above. Furthermore,* GD *is decidable.*

Thus, the modal system GD completely captures the interaction between bwb_{PA} and $1bwb_{\mathsf{PA}}$; also Theorem 5.5 carries over. However, GD no longer has an adequate Kripke semantics, which complicates the decision procedure. For further references see [Boo] or [Be3].

As an exercise, the reader should derive $\boxed{1}(\Box p \to p)$ from the axioms of GD. Thus, $\vdash_{\mathsf{PA}} \boxed{1}(\Box\alpha \to \alpha)$ for *every* $\alpha \in \mathcal{L}^0_{ar}$, while $\vdash_{\mathsf{PA}} \Box(\Box\alpha \to \alpha)$ is the

case only provided $\vdash_{\mathsf{PA}} \alpha$. In other words, the *local reflection principle* $\{\Box\alpha \to \alpha \mid \alpha \in \mathcal{L}^0_{ar}\}$ is 1-provable in PA. **Be careful:** GD expands G conservatively, so that $\nvdash_{\mathsf{GD}} \Box p \to p$.

7.7 Modal Operators in ZFC

Considerations regarding self-reference in ZFC are technically sometimes easier, but from the foundational point of view more involved because there is no superordinate theory. If ZFC is consistent, as we assume it is, then $\mathtt{Con}_{\mathsf{ZFC}}$ is a true arithmetical statement that is unprovable in ZFC. Thus, true arithmetical statements may even be unprovable in ZFC, not only in PA or similarly strong arithmetical theories. It makes sense, therefore, to consider $\mathsf{ZFC}^+ := \mathsf{ZFC} + \mathtt{Con}_{\mathsf{ZFC}}$, because after all, we want set theory to embrace as many facts about numbers and sets as possible from which interesting consequences may result.

As **7.3** shows, the consistency of ZFC alone does not guarantee that ZFC^+ is consistent. The second incompleteness theorem clearly excludes $\vdash_{\mathsf{ZFC}} \mathtt{Con}_{\mathsf{ZFC}}$ but does not preclude $\vdash_{\mathsf{ZFC}} \mathtt{Con}_{\mathsf{ZFC}} \to \mathtt{Con}_{\mathsf{ZFC}^+}$. In this case $\vdash_{\mathsf{ZFC}^+} \mathtt{Con}_{\mathsf{ZFC}^+}$, and so $\vdash_{\mathsf{ZFC}^+} \bot$ by the same theorem. On the other hand, from certain assumptions about the existence of large cardinals, the consistency of ZFC^+ readily follows. These assumptions would have to be jettisoned in case $\vdash_{\mathsf{ZFC}^+} \bot$, i.e. $\vdash_{\mathsf{ZFC}} \neg\, \mathtt{Con}_{\mathsf{ZFC}}$. Moreover, the consistency of ZFC would then not correctly be reflected in ZFC, and ZFC proves along with true arithmetical facts also false ones. This sounds strange, but there is hardly a convincing argument that this cannot be so.

Even if ZFC^+ is consistent, i.e. $\nvdash_{\mathsf{ZFC}} \neg\, \mathtt{Con}_{\mathsf{ZFC}}$, it may still be that one of the sentences from the sequence $\Box\neg\, \mathtt{Con}_{\mathsf{ZFC}}, \Box\Box\neg\, \mathtt{Con}_{\mathsf{ZFC}}, \ldots$ is provable in ZFC (where $\Box$ denotes $\Box_{\mathsf{ZFC}}$ as long as it is not redefined). The latter is excluded only if we assume that the ω-iterated consistency extension ZFC^ω is consistent, hence $\nvdash_{\mathsf{ZFC}} \Box^n\bot$, for all n (see page 284), so that by Theorem 5.3, G would be the provability logic of ZFC.

In fact, the assumption $(\forall n{\in}\mathbb{N})\ \nvdash_{\mathsf{ZFC}} \Box^n\bot$ is equivalent to G's being the provability logic of ZFC, by the general Theorem 7.1 below. Therein $Rf_T := \{\Box\alpha \to \alpha \mid \alpha \in \mathcal{L}^0\}$ denotes the already encountered reflection principle. Also Theorem 5.3 is a corollary of the theorem, simply because $(\forall n{\in}\mathbb{N})\ \nvdash_T \Box^{n+1}\bot$ is equivalent to the consistency of T^ω.

Theorem 7.1. *For a sufficiently expressive theory* T [5] *the following conditions are equivalent:*

(i) T^ω *is consistent,*

(ii) $T + Rf_T$ *is consistent,*

(iii) G *is the provability logic of* T.

Proof. (i)⇒(ii) indirect: Suppose that $T + Rf_T$ is inconsistent. Then there are formulas $\alpha_0, \dots, \alpha_n$ such that $\vdash_T \neg\varphi$, $\varphi := \bigwedge_{i=1}^{n}(\Box\alpha_i \to \alpha_i)$. Hence $\vdash_T \Box\neg\varphi \equiv_T \neg\Diamond\varphi$. Now, because $\vdash_{T^\omega} \neg\Box^{n+1}\bot$, by Example (d) in **7.4** and Lemma 5.1, we get $\vdash_{T^\omega} \Diamond R_n^i$ $(p_i^i = \alpha_i)$. Clearly, $R_n^i = \varphi$ and so $\vdash_{T^\omega} \Diamond\varphi$. Since also $\vdash_{T^\omega} \neg\Diamond\varphi$, T^ω is inconsistent. (ii)⇒(iii): The proof of Theorem 5.2 for PA, as presented in [Boo], runs nearly in the same way for T, because PA is transgressed in one place only: one uses the fact that $\mathcal{N} \vDash Rf_{\mathsf{PA}}$. However, the existence of a corresponding T-model is ensured by (ii). (iii)⇒(i): $\nvdash_{\mathsf{G}} \Box^{n+1}\bot$, hence $\nvdash_T \Box^{n+1}\bot \equiv_T \neg\,\mathtt{Con}_{T^n}$ for all n, and so T^ω is consistent. ❑

The equivalence (i)⇔(ii) is a purely proof-theoretic one. It is called *Goryachev's theorem*; see [Gor] or [Be2]. We obtained it using essentially some elementary modal logic. For $T = \mathsf{ZFC}$, perhaps a bit more interesting than (i) or (ii) is the assumption of the *ω-consistency* of ZFC, that is,

$(*)$ $\vdash_{\mathsf{ZFC}} (\exists x{\in}\omega)\varphi(x) \Rightarrow \nvdash_{\mathsf{ZFC}} \neg\varphi(\underline{n})$ for some n $\quad(\varphi(x) \in \mathcal{L}_\in)$.

This assumption implies $D1^*$, which in turn ensures $\nvdash_{\mathsf{ZFC}} \Box^{n+1}\bot$ for all n, that is, (i), and hence all other conditions in Theorem 7.1 hold for $T = \mathsf{ZFC}$. It is worthwhile to observe that the consistency of $\mathsf{ZFC} + Rf_{\mathsf{ZFC}}$ and thereby the proof of Solovay's completeness theorem for ZFC follow directly from $(*)$, without appealing to Goryachev's theorem. What is needed to see that the latter is the case is the following

Lemma. *Suppose that* ZFC *is ω-consistent. Then there exists a model* $\mathcal{V} \vDash \mathsf{ZFC}$ *such that* $\mathcal{V} \vDash Rf_{\mathsf{ZFC}}$.

Proof. Let $\Omega := \{(\forall x \in \omega)\alpha \mid \alpha{=}\alpha(x) \in \mathcal{L}_\in, \vdash_{\mathsf{ZFC}} \alpha(\underline{n}) \text{ for all } n\}$. Then $\mathsf{ZFC}{+}\Omega$ is consistent. Indeed, otherwise $\vdash_{\mathsf{ZFC}} \neg(\forall x \in \omega)\alpha \equiv (\exists x \in \omega)\neg\alpha$ for some $(\forall x \in \omega)\alpha \in \Omega$ (since Ω is closed under conjunction), in contradiction to $(*)$. Any $\mathcal{V} \vDash \mathsf{ZFC} + \Omega$ satisfies the reflection principle Rf_{ZFC}, for if

[5] By such a T we mean that the proof steps of Solovay's Theorem 5.2 not transgressing PA can be carried out in T. This does not yet imply the provability of the theorem itself. Which steps are transgressing PA is described in the following proof.

$\mathcal{V} \not\models \alpha$ then $\not\vdash_{\mathsf{ZFC}} \alpha$ and therefore $\vdash_{\mathsf{ZFC}} \neg\, \mathtt{bew}_{\mathsf{ZFC}}(\underline{n}, \ulcorner\alpha\urcorner)$ for all n. Hence $(\forall y \in \omega) \neg\, \mathtt{bew}_{\mathsf{ZFC}}(y, \ulcorner\alpha\urcorner) \in \Omega$, which clearly implies $\mathcal{V} \not\models \Box\alpha$. ❏

Now we interpret the modal operator $\Box$ no longer as *provable in* ZFC, which is equivalent to *valid in all* ZFC*-models*, but rather as *valid in particular classes of* ZFC*-models*. For undefined notions used in the sequel we refer to [Ku]. A 'model' is to mean throughout a ZFC-model.

Particularly interesting are *transitive* models, i.e. models $\mathcal{V} = (V, \in^{\mathcal{V}})$, where the set V is *transitive*. This is to mean $a \in b \in V \Rightarrow a \in V$. In these models, $\in^{\mathcal{V}}$ coincides with the ordinary $\in$-relation restricted to V (a set in our metatheory that itself is ZFC). We write V for $\mathcal{V}$. Let ρa denote the *ordinal rank* of the set a, i.e., the smallest ordinal ρ with $a \in V_{\rho+1}$. To prove the soundness half of Theorem 7.3 we will need

Lemma 7.2. ([JK]) *Let V, W be transitive models such that $\rho V < \rho W$ and let $V \models \alpha$. Then $W \models$ 'there is a transitive model U with $U \models \alpha$'.*[6]

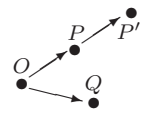
Let the modal logic Gi result from augmenting G by the axiom

(**i**) $\Box(\Box p \to \Box q) \vee \Box(\Box q \to p)$.

Gi is complete with respect to all *preference orders* g, i.e., g is a finite poset together with some function $\pi : g \to n$ $(= \{0, \ldots, n-1\})$ such that $P < Q \Leftrightarrow \pi P < \pi Q$, for all $P, Q \in g$. This implies the finite model property of Gi, which, as for G, ensures the decidability of Gi. More suitable for our aims is the characterization of preference orders g by the property

(**p**) $P < P'$ implies $P < Q$ or $Q < P'$, for all $P, P', Q \in g$,

which at once follows from the definition: Let $P < P'$, hence $\pi P < \pi P'$. If $P \not< Q$, i.e. $\pi P \not< \pi Q$, then $\pi Q \leqslant \pi P < \pi P'$, so that $Q < P'$. The proof of the converse is Exercise 1. The figure shows a poset g that is *not* a preference order (for neither $P < Q$ nor $Q < P'$). Axiom (**i**) is easily refuted in g choosing $wp = \{P'\}$, $wq = \emptyset$, and verifying that $O \Vdash \Diamond(\Box p \wedge \neg\Box q)$ and $O \Vdash \Diamond(\Box q \wedge \neg p)$ (for notice that $P \Vdash \Box p \wedge \neg\Diamond q$ and $Q \Vdash \Box q \wedge \neg p$). Hence, (**i**) does not belong to G, so that Gi is a proper extension of G. We mention that in [So] and in [Boo] a somewhat more complex axiomatization of Gi has been considered.



[6] In transitive models W the sentence in ' ' (which with some encoding can be formulated in $\mathcal{L}_\in$) is absolute, and therefore equivalent to the existence of a transitive model $U \in W$ with $U \models \alpha$.

Remark on splittings in modal logic. The completeness of Gi with respect to all preference orders follows also from the fact that Gi is the split logic arising from splitting the lattice of all extensions of G (see e.g. [Kra]) by the subdirect irreducible G-algebra belonging to the frame from the previous page.

We define insertions $\imath\colon \mathcal{F}_\Box \to \mathcal{L}^0_\in$ as in **7.5** as usual by $(\Box H)^\imath = \Box H^\imath$, where $\Box\alpha$ for the set-theoretic sentence $\alpha = H^\imath \in \mathcal{L}^0_\in$ is now to mean 'α is valid in all transitive models'. Accordingly, $\Diamond\alpha = \neg\Box\neg\alpha$ states 'α holds in at least one transitive model'.

Theorem 7.3. $\vdash_{\mathsf{Gi}} H$ *iff* $\vdash_{\mathsf{ZFC}} H^\imath$ *for all insertions* $\imath$ *as defined above.*

We prove only the direction $\Rightarrow$, that is, soundness. The converse is much more difficult, see [Boo]. As regards the axioms of Gi, since $\Box p \to \Box\Box p$ is provable from the other axioms of G (see **7.4**), it suffices to prove

(A) $\Box(\alpha \to \beta) \wedge \Box\alpha \vdash_{\mathsf{ZFC}} \Box\beta$, (B) $\Box(\Box\alpha \to \alpha) \vdash_{\mathsf{ZFC}} \Box\alpha$,

(C) $\vdash_{\mathsf{ZFC}} \Box(\Box\alpha \to \Box\beta) \vee \Box(\Box\beta \to \alpha)$, for all $\alpha, \beta \in \mathcal{L}^0_\in$.

(A) is trivial, because the sentences valid in any class of models are closed under MP. (B) is equivalent to (B′): $\Diamond\neg\alpha \vdash_{\mathsf{ZFC}} \Diamond(\Box\alpha \wedge \neg\alpha)$. Here is the proof: Suppose $\Diamond\neg\alpha$, i.e. there is a transitive model in which $\neg\alpha$ holds. Then there is also one with minimal rank, V say. We claim $V \vDash \Box\alpha$. Otherwise $V \vDash \Diamond\neg\alpha$, and hence there would be a transitive model $U \in V$ with $U \vDash \neg\alpha$ and $\rho U < \rho V$, contradicting our choice of V. Therefore, $V \vDash \Box\alpha \wedge \neg\alpha$. Thus, there is a transitive model in which $\Box\alpha \wedge \neg\alpha$ holds, which confirms (B′). Finally, (C) is verified by contraposition: suppose there are transitive models V, W and sentences α, β such that

(a) $V \vDash$ 'α holds in all transitive models and there is a transitive model in which $\neg\beta$ holds',

(b) $W \vDash$ 'β holds in all transitive models', (c) $W \vDash \neg\alpha$.

From these assumptions it follows first of all that $\rho W < \rho V$. Indeed, suppose by (a) that $U \in V$ is a transitive model for $\neg\beta$. If $\rho V \leqslant \rho W$ then $\rho U < \rho W$. Hence, by Lemma 7.2, $W \vDash$ 'there is a transitive model for $\neg\beta$', contradicting (b). Now, since $W \vDash \neg\alpha$ by (c) and because of $\rho W < \rho V$, in V holds 'there is some transitive model for $\neg\alpha$' by Lemma 7.2, in contradiction to (a). This proves (C). Soundness of the substitution rule follows as for G in **7.5**. MN is trivially sound, because if α is provable in ZFC then, of course, α is valid in all transitive models. Also MP is obvious: If α and $\alpha \to \beta$ hold in any class of models, then also β.

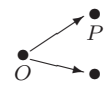
Another interesting model-theoretic interpretation of $\Box\alpha$ is 'α is valid in all V_κ'. Here κ runs through all inaccessible cardinal numbers. According to [So], the adequate modal logic for this interpretation of $\Box$ is

$$\mathsf{Gj} := \mathsf{G} + \Box(\Box p \wedge p \to q) \vee \Box(\Box q \to p).$$

More precisely, if there are infinitely many inaccessibles then we have

Theorem 7.4. $\vdash_{\mathsf{Gj}} H$ *iff* $\vdash_{\mathsf{ZFC}} H^\imath$ *for all insertions* $\imath$*, where* $\Box\alpha$ *is to mean '*α *is valid in all* V_κ*', running through all inaccessible cardinals.*

Gj is also denoted by G.3. This logic is complete with respect to all finite strict linear orders. These, of course, are also frames for Gi, so that $\mathsf{Gi} \subseteq \mathsf{Gj}$. The figure shows a Gi-frame, also called "the fork," on which the additional axiom is easily refuted at its initial point O with $wp = \{P\}$ and $wq = \emptyset$. Hence the fork is not a Gj-frame, and so $\mathsf{Gi} \subset \mathsf{Gj}$. The completeness of Gj with respect to finite orders entails the finite model property of Gj and hence its decidability.



We recommend that the reader carry out the proof of the soundness part of Theorem 7.4, without consulting the hints to the solutions (Exercise 4). It is easier than the soundness part of Theorem 7.3 proved above. All one needs to know besides Lemma 7.2 is that each V_κ is a transitive ZFC-model and that $V_\kappa \in V_\lambda$ or $V_\lambda \in V_\kappa$, for arbitrary inaccessible cardinals $\kappa \neq \lambda$. Maybe the reader can also find a new and lucid proof of the hard direction of Theorem 7.4: If $\vdash_{\mathsf{ZFC}} H^\imath$ for all $\imath$ then H holds in all finite strict linear orders, or equivalently, $\mathsf{Gj} \vdash H$.

Exercises

1. Let g be a G-frame with property (**p**), page 296. Show by induction on the length of a maximal path in g that g is a preference order.

2. Show (using Exercise 1) that axiom (**i**) for Gi holds in a G-frame g iff g is a preference order. This is an essential step in proving the completeness of Gi with respect to all preference orders.

3. This exercise is a crucial step in the completeness proof of Gj. Show that a G-frame g is a frame for Gj, i.e., $\Box(\Box p \wedge p \to q) \vee \Box(\Box q \to p))$ holds in g if and only if g is (totally) ordered.

4. Verify the soundness part of Theorem 7.4, i.e., $\vdash_{\mathsf{Gj}} \Rightarrow \vdash_{\mathsf{ZFC}} H^\imath$ for all insertions $\imath$.